Data Sharing in NIH-Funded Research:
a Case Study of Data Sharing Practices and Opportunities

By
CHINONYE E. HARVEY
B.S., Michigan State University, 1998
M.P.H., George Washington University, 2001

DISSERTATION

Submitted as partial fulfillment of the requirements for the degree of Doctor of Public Health in Leadership in the School of Public Health of the University of Illinois at Chicago
Chicago, Illinois. USA
May 2019

Dissertation Committee:

Dr. Eve Pinsker, Chair, Community Health Sciences
Dr. Kristina Risley, de Beaumont Foundation
Dr. Elizabeth Gillanders, National Cancer Institute, National Institutes of Health
Dr. Michael Petros, Environmental and Occupational Health Sciences
Dr. Steven Seweryn, Epidemiology and Biostatistics

DISCLAIMER

The findings and conclusions in this report are those of the author(s) and do not necessarily represent the official position of the National Cancer Institute, National Institutes of Health.

DEDICATION

This dissertation is dedicated to my dear husband Nik and my two beautiful children, Nyla (13 years) and Nevin (9 years), without whom I would not have started on this journey let alone gotten this far. You are my world and words cannot express how deeply grateful I am to you for all the sacrifices you made, every single day for the past five years, to help me reach this milestone. Through the good and hard times, through lots of laughter and lots of tears, your love and warm hugs kept me going. This is for you. We did it – together!

To my parents, Dr. Luke and Chinwe Umeh – thank you for teaching me hard work, resilience, and faith, and for inspiring me to always aim for the highest and the best. To my brothers Dr. Ifeanyi Umeh, Chidulue Umeh, and my sister Ogechi Umeh – thank you for being the best siblings anyone could ask for, my biggest supporters, always ready to jump in. To the rest of my extended family – thank you for all of your support, prayers and encouragement throughout the years. You were always there when I needed you.

Finally, I want to honor my late grandmother, mama (Eunice) and my late uncle Bosa, who both passed within three months of each other last year, in the middle of this dissertation. It was one of the most challenging years of my life but the celebration of your lives and legacies inspired me to keep going - I know you are proud.

ACKNOWLEDGEMENTS

To Dr. Kristina Risley, my committee chair (until Fall 2018), coach and champion from day one - thank you so much for your tremendous guidance and confidence in my work. To my other committee members, I am forever indebted to all of you for guiding me on this long and rewarding journey, and staying committed to my success. Thank you Dr. Mike Petros and Dr. Steve Seweryn for your unwavering support over the years. Thank you Dr. Eve Pinsker for stepping in as my chair at the final stages of my dissertation. A special thanks to my mentor, EGRP colleague and committee member - thank you Dr. Gillanders for your tremendous support, genuine care, invaluable insights on my research topic, and for always reminding me to be kind to myself throughout this rigorous process.

A huge thanks to all my study participants who took time out of their very busy schedules to participate in my research. I am extremely grateful and fortunate to have had such a unique opportunity to learn first-hand from your experiences.

I want to especially thank my EGRP colleague, Dr. Danielle Daee, for her willingness to help with my analysis as a second coder. To Dr. Kathy Helzlsouer, my boss, thank you for your leadership, support, insights and mantra - "focus to finish." To my former EGRP boss Dr. Muin Khoury, former EGRP supervisor Dr. Britt Reid, and Stacey Vandor, thank you for being open to this idea five years ago when I presented it to you. I will always cherish your trust, support and care. To the rest of my EGRP and DCCPS colleagues and friends, thank you for sharing your expertise and experiences with me, and for being some of my biggest supporters since my tenure at NCI fourteen years ago.

It truly has been a privilege to be a part of the DrPH in Leadership program and to have had so many opportunities to learn about and embody leadership in my work. Thank you to my fellow UIC DrPH 2014 cohort classmates whose support and friendships have been invaluable. I want to also thank former students, Dr. Alina Flores, Dr. Vanessa Byams, and Dr. David Reynen whose insights and experiences were very helpful to me at different stages of my dissertation.

CEH

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES (continued)

LIST OF TABLES (continued)

LIST OF FIGURES

ABBREVIATIONS

BioLINCC	Biologic Specimen and Data Repository Information Coordinating Center
CEC	Cancer Epidemiology Cohort
DAC	Data Access Committee
dbGaP	Database of genotypes and phenotypes
DCCPS	Division of Cancer Control and Population Sciences
EGRP	Epidemiology and Genomic Research Program
ESI	Early Stage Investigator
GDS	Genomic Data Sharing
GWAS	Genome Wide Association Studies
IRB	Institutional Review Board
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NIH	National Institutes of Health
NOT	Notice
PI	Principal Investigator
RFI	Request for Information
R01	Research Project grant

SUMMARY

Data Sharing advances the pace of scientific discovery, improves public health and enhances clinical benefit. It promotes reproducibility, replication and validity of research results, generates new hypotheses or research questions not already addressed in the original study, and advances the state of research and innovation (Borgman, 2012). The National Institutes of Health (NIH) has long fostered a culture which promotes data sharing, and has developed data sharing policies that require researchers to share data from their federally-funded studies. The three major policies of relevance are the 2003 NIH Data Sharing policy; the 2007 policy on Genome-Wide Association Studies (GWAS); and the more recent NIH Genomic Data Sharing (GDS) policy. NIH has also invested significantly in developing an infrastructure for genomic data sharing.

Despite the benefits, policies and resources in place for sharing data, investigators are not motivated to share their data with others (Olson 2017). The uptake of data sharing practices among researchers is not optimal, especially for non-genomic data sharing. Researchers are reluctant to share their data in public or controlled-access data repositories for a variety of reasons. However, the field of genomics has pioneered the culture of open data access and data sharing (Kaye et al, 2009) in advancing progress and much of what we know to be critical to successful sharing already exists in the field of genomics. For example, there are systems, structures and policies in place to facilitate sharing of genomic data.

A case study design was used to explore organizational level factors influencing the sharing of data in public or controlled-access data repositories, to design organizational

approaches to achieve wider data sharing. Exploring the successes and challenges of the implementation of the NIH GDS policy, as an exemplar case, provided insights into factors that facilitate or hinder data sharing. In addition, it provided opportunities to apply that knowledge to epidemiological data sharing and data sharing practices among the research community in general. The perspectives of NIH supported investigators (both new and experienced) of major research initiatives and with some experience with data sharing, as well as NIH staff who lead the development and implementation of the GDS policy, was critical in understanding what it takes for successful data sharing to occur and opportunities for achieving this. Key public documents were also reviewed and analyzed as part of this qualitative research study.

The findings from this case study generated knowledge that resulted in a set of recommendations for changes to the existing data sharing policies and implementation processes, as well as informing the development of new polices and enhanced implementation strategies. Addressing data sharing issues around culture, policy, resources and technological infrastructure is important in improving data sharing in biomedical research. The implications of this research while it cannot be generalized across funding agencies or institutions, are transferable to similar organizations supporting data sharing activities in research settings, with the ultimate goal to advance scientific progress and improve public health outcomes.

I. BACKGROUND AND PROBLEM STATEMENT

A. Background

i. Introduction

The National Institutes of Health (NIH) has always been an advocate for sharing research data to advance medical science, accelerate discovery and innovation, and improve the health of populations. In 2003, the NIH released a data sharing policy encouraging investigators with research grants of $500,000 or more in total direct costs in any single year to provide a plan for how they will share their data, regardless of the type of data. This was a general policy that applied to all types of data generated from large research grants, including genetic data and observational data. The policy does not explicitly mandate sharing; but rather simply requires an evidence of a plan for data sharing as confirmed by NIH Program Officers. Subsequently in 2007, the NIH published a policy on Genome-Wide Association Studies (GWAS) that required researchers conducting GWAS studies to deposit both genotype and phenotype datasets in a controlled-access repository e.g., the database of Genotypes and Phenotypes (dbGaP). GWAS studies are useful in exploring genetic factors associated with diseases that can influence health.

More recently in 2015, the NIH released the Genomic Data Sharing (GDS) policy, an expansion of the GWAS policy, which requires that all genomic data from NIH-funded research be submitted in a NIH-designated data repository (dbGaP). Unlike the NIH general Data Sharing Policy of 2003 and the NIH GWAS policy, the NIH GDS policy specifically mandates broad and responsible sharing of large scale human and non-human genomic research data generated from NIH-funded studies regardless of funding levels and NIH has systems in place to monitor and check compliance. There are other examples of successful data sharing across the

institute but these will not be included in this dissertation. For example, the National Heart Lung and Blood Institute (NHLBI) supports a data and biospecimen repository (Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC))[1] for submission and access to clinical trial and observational data.

Regardless of the nuances in the language used in the policies, the expectation of these NIH data sharing policies is that all researchers would share their data, taking appropriate measures to protect patient confidentiality, and contribute towards the advancement of new scientific discoveries for public health and clinical benefit. However, the uptake of these policies among NIH-funded investigators is not optimal. There are ongoing efforts at various levels of the institute, other funding agencies, and science journals to promote the broad sharing of data from federally-funded research, but these are not without challenges. Currently, sharing is either minimal or non-existent which warrants further investigation. Specifically, researchers funded through the National Cancer Institute's (NCI) Epidemiology and Genomics Research Program (EGRP) face more challenges with sharing epidemiological data in public or controlled-access databases than with genomic data due to a variety of factors that were explored in this dissertation.

This dissertation explored organizational / institutional level factors influencing the sharing of data generated by NIH-funded research in public or controlled-access databases or data repositories in order to design organizational approaches to achieve wider data sharing. There are also individual level and interpersonal factors that facilitate or hinder the sharing of research data however, these were not the primary focus of this research because they have been studied elsewhere. The inter-relationships between the organizational, individual and interpersonal factors are not all mutually exclusive. For example, organizational / institutional

[1] https://biolincc.nhlbi.nih.gov/home/

factors such as technology will have to address individual level factors such as motivation for successful sharing of research data to occur. Understanding these factors will help shed light on some of the gaps in organizational / institutional systems designed to support and promote data sharing in biomedical research and opportunities for improving these systems.

ii. Data Sharing in Biomedical Research and the Impact on Public Health

There have been major strides made over the past several decades in biomedical research and public health however populations continue to face some of the most complex scientific and emerging public health problems of the twenty-first century. Many leaders in the biomedical research field have argued that a different approach is necessary to address these multifaceted problems. This has led to a shift from a siloed and independent approach that is ubiquitous in academic institutions, where individuals focus on their own research and limit access to their data, to a more multidisciplinary and collaborative approach that involves sharing of data, knowledge and expertise across disciplines and teams (Tenopir, 2011).

Midgley et al (2006) suggest that leaders must employ a systems approach to look at multiple factors and groups, and engage researchers from various fields and disciplines to help drive creativity, innovation and rigor in scientific research methods, processes and practices to solve these adaptive public health challenges. For example, the NIH supports many multidisciplinary research studies through a variety of mechanisms, such as the multiple Principal Investigator (PI) model. This model allows more than one principal investigator on a single grant, fostering integrated expertise from different disciplines and promoting collaborative research and team science.

Due to the complex nature of public health and the many layers of influence, collaborative approaches including the broad sharing of data among researchers is important in addressing these problems. Stokols et al (2008) describe contextual factors such as organizational, technological, sociopolitical, physical environmental, interpersonal and intrapersonal factors that influence collaborative research and inherently, the practice of data sharing among researchers. The implications of these factors have been examined in the literature but more needs to be done to fully understand the relationships between such factors and the role they play in the implementation of policies that guide research practices; from discovery through replication and reproducibility.

Two Elusive Concepts: Data and Data Sharing

The concepts of "data" and "data sharing" are elusive and mean different things to different people in different fields. Understanding what they mean will gain more appreciation for the national efforts around promoting data sharing practices within the scientific community. There is variation in beliefs regarding what data is to be shared, where to share data, how to share data and with whom data should be shared. Some may think of data as the "form" (e.g. physical or electronic) or "type" of data (e.g. aggregate or individual level data; questionnaire data or clinical data, etc.). Butlin (2011) defines data as the "building blocks of science, the basic observations around which we construct our theories." Most recently, NIH defines digital scientific data as the electronic form of data that is "commonly acceptable in the scientific community as necessary to validate research findings including dataset used to support scholarly publications" and excludes software, data collection and analytic tools, preliminary analyses, draft papers and laboratory specimens (NIH Plan, 2015). The definition also includes "data that are used to support a scientific publication as well as data from completed studies that might

never be published, including data that support or refute a hypothesis, but does not include draft or preliminary data sets (NIH Plan, 2015)."

In Wikipedia data sharing is defined as "the practice of making data use for scholarly research available to other investigators" (Wikipedia, n.d.). Another definition of data sharing is the "deposition and preservation of data primarily to provide access for data use and reuse (Tenopir, 2011)." Essentially, data sharing encompasses the release of data for use by others including posting on websites and submitting to journals or submitting in data repositories or databases for controlled or public access. According to the NIH Plan (2015), public access to data refers to the "availability of data for public use, including data that are openly available for any use or data with controlled-access to protect the privacy of research participants, intellectual property or security." Another layer of complexity around defining "data" and "data sharing" has to do with the variation regarding who owns the data and who pays for the data to be shared.

Benefits of Data Sharing

Data is essential to basic scientific and public health research and sharing data broadly is critical in advancing science and improving health outcomes. Data sharing promotes reproducibility, replication and validity of research results, generates new hypotheses or research questions not already addressed in the original study, advances the state of research and innovation (Borgman, 2012) and ensures rigor in study methods and analyses. Ultimately, data sharing plays a key role in enhancing public value and clinical benefit by allowing data and results of publicly funded research to become accessible to the public. This improves knowledge and understanding of research data and results, informs prevention strategies, risks and benefits of treatment options, and influences clinical practice and health decisions at the individual, social, economic and political levels.

A recent re-analysis of two large prostate cancer screening trials on the effectiveness of prostate cancer screening shows the benefit of data sharing in research. For several years, there has been controversy in the scientific and medical community about the effectiveness of prostate-specific antigen (PSA) screening, the gold standard since the 1980s, as a way of detecting people who have the disease so they can get treatment and prevent death. The issue is also related to whether the benefits outweigh the harms especially with possible over diagnosis and over treatment in men who are less likely to benefit from the screening.

In 2009, two prostate cancer screening trials, one in the U.S. and the other in Europe, published contradictory results on the effectiveness of prostate-specific antigen (PSA) testing in the same issue of the *New England Journal of Medicine*. The Prostate, Lung, Colorectal and Ovarian (PLCO) Trial in the United States (U.S.) suggested that screening did not reduce the risk for prostate cancer mortality i.e. the rates were the same for the control group and the intervention group. On the contrary, the European Randomized Study of Screening for Prostate Cancer (ERSPC) suggested a reduction in prostate cancer mortality by 20%" (Vickers AJ, 2017). Considering this, the U.S. Preventive Services Task Force's (USPSTF) determined that the benefits were very small given results of these two trials that averaged about a 10% reduction in mortality from prostate cancer. This was the basis for the task force's recommendation in 2012 against PSA testing.

To assess PSA screening benefits and effects on prostate cancer mortality from the PLCO and ESRPC trials which involved over 200,000 men, Tsodikov et al (2017) pooled and extracted data from the trials, re-analyzed the data and compared results from the two trials to see if there were differences in effects of screening compared to no screening. They used more rigorous approaches which identified errors in the study methods and analyses especially around

screening intensity used in each trial group. As a comparison with those who were not screened, the researchers found a 25% to 32% lower risk in prostate cancer death in men who were screened. This is a significant increase and confirms PSA testing as an effective public health and clinical intervention in the prevention and treatment of prostate cancer.

This is an example of where data sharing led to a translational and clinically actionable intervention in public health. It speaks to the importance of other researchers being able to reproduce findings using same data as the primary investigator / original data generator; the need for checks in rigor, scrutiny, accuracy, evaluation or validity of published findings to help uncover errors in published findings. If the data weren't available for others to use, it wouldn't have been confirmed that PSA screening can significantly lower risk of prostate cancer by a significant.

While the USPSTF has previously based its recommendations about PSA screening on the findings from the PLCO and ERSPC trials, it is now considering re-evaluating and updating its recommendations based on the evidence from this new study that PSA screening has a significant reduction in prostate cancer death in men. This study brings to light the missed opportunities over the past five years in interventions for men with prostate cancer as well as delays in reduction in prostate cancer mortality due to implementation of the 2012 recommendations and guidelines. Prostate cancer deaths in the U.S. have decreased by about 50%, from 38.6 to 19.1 per 100,000 men and while the factors are multifactorial with contributions from screening it is not clear what the impact of the five-year delay in intervention is relative to the U.S. prostate cancer mortality rate.

Challenges with Data Sharing

Sharing data generated from publicly funded research is an issue and not something that all researchers are fond of due to its demonstrated complexity and undue administrative and financial burden (Tenopir, 2015). More specifically, data sharing is not perceived as rewarding among researchers if there is a lack of funding and infrastructure to support sharing. For the scientific research community to appreciate the value of data sharing, the social context and implications should be emphasized (Azberger, 2004). Of concern among many scientists is the risk of their data being scooped and published in a high-impact journal, in addition to concern for patient confidentiality. Many studies have examined the benefits and challenges associated with data sharing practices and show minimal data sharing among researchers and variation across different disciplines (Tenopir, 2011).

Funding agencies such as the NIH, some science journals and other research institutions have established policies to encourage data sharing. However, not all researchers are willing to share their data in public or controlled-access data repositories. Another concern is that not all research is subject to data sharing requirements, and policies are either enforced loosely, or inconsistently across institutes. Despite the challenges faced with data sharing in scientific research, the field of genomics shows that "data sharing cuts duplication, speeds progress and increases career opportunities for researchers; also, leading to better policies and healthier people (Pisani et al, 2010)." This dissertation will focus on leveraging the experiences of data sharing in genomics at NCI to elucidate best practices for implementing data sharing in epidemiology research funded by NCI, the largest of twenty-seven institutes and centers of the NIH.

NIH's Perspective of Data Sharing

The NIH is an agency of the U.S. Department of Health and Human Services (DHHS), the largest supporter of biomedical research in the world and the nation's medical research agency. It is comprised of 27 institutes and centers (Figure 1) and invests almost $40 billion annually in medical research. Each institute and center is funded separately by the U.S. Congress and has a different mission, budget, and priorities which often focus on specific diseases or body systems. For example, NCI conducts and supports research related to the cause, diagnosis, prevention and treatment of cancer and cancer-related outcomes. The NIH through its extramural research program supports more than 300,000 researchers at over 2,500 academic research institutions in the U.S. and globally. This represents more than 80% of NIH's budget.

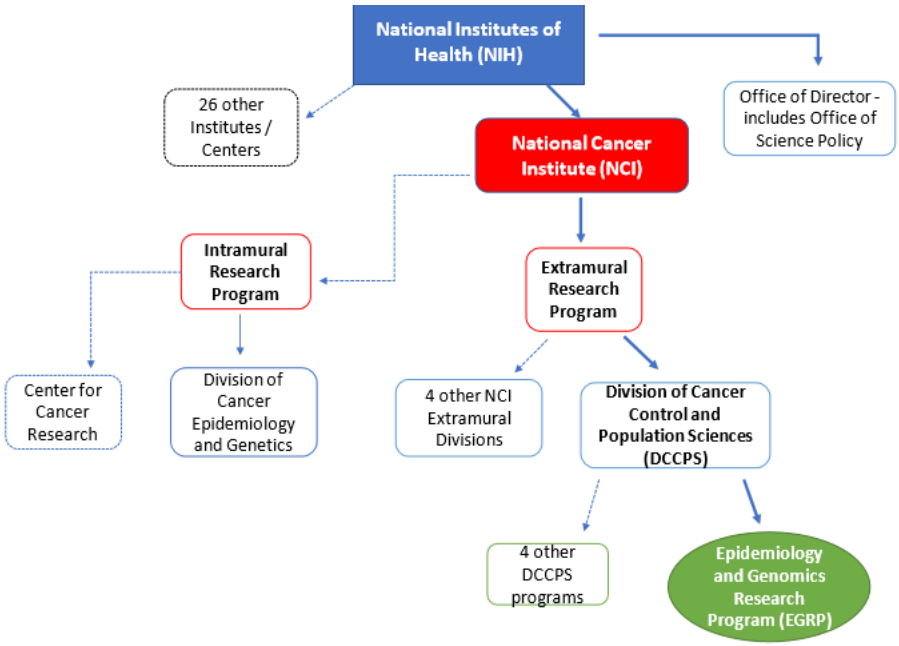

Figure 1: NIH Organizational Structure (High-Level)

The mission of the NIH is "to seek fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability.[2] " The goals of the NIH are: "1) to foster creative discoveries, innovative research strategies and their applications, 2) to develop, maintain, and renew scientific human and physical resources towards disease prevention, 3) to expand the knowledge base in medical and associated sciences to enhance the Nation's economic well-being and ensure high return on public investment in research, and 4) to exemplify and promote the highest level of scientific integrity, public accountability, and social responsibility in the conduct of science.[3]"

In order to realize these goals, NIH, under the leadership of Dr. Francis Collins (the current Director of NIH) and the directors of other NIH Institutes and Center, supports a wide range of research looking at 1) "the causes, diagnosis, prevention, and cure of human diseases; 2) the processes of human growth and development; 3) the biological effects of environmental contaminants; 4) the understanding of mental, addictive and physical disorders; and 5) in directing programs for the collection, dissemination, and exchange of information in medicine and health.[3]"

The NIH has a strong culture of supporting the sharing of final research data to achieve its mission and scientific research goals. According to the NIH Data Sharing Policy of 2003, NIH believes that "all data from NIH-funded research should be made as freely and widely available as long as doing so safeguards the privacy of participants and the confidentiality and proprietary nature of the data.[4]" While NIH believes that sharing research data "allows scientists to expedite the translation of research results into knowledge, products, and procedures to improve human health" (NIH Genomic Data Sharing, n.d.), it "recognizes that data sharing may

[2] https://www.nih.gov/about-nih/what-we-do/mission-goals
[3] https://www.nih.gov/about-nih/what-we-do/mission-goals
[4] https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

be complicated or limited, in some cases, by institutional policies, local IRB rules, as well as local, state and Federal laws and regulations, including the Privacy Rule.[3] ”

Historical Context of Data Sharing Policies at NIH

The institutionalization of data sharing in biomedical research (and at NIH) dates as far back as 1990 when the Human Genome Project (HGP) was launched. Following the launch, a large international collaborative effort to map and sequence all the genes of the human genome convened leaders in the scientific community during a summit held in Bermuda in 1996. They agreed on a set of guiding principles known as the Bermuda principles (developed during the 1996 summit in Bermuda) that would require the release of all DNA sequence data in publicly accessible databases within twenty-four hours after generation (Wikipedia, n.d.). A few years later in 2003, after a meeting in Fort Lauderdale, Florida, organized by the Wellcome Trust, the Fort Lauderdale Agreement was established. This was a public declaration of the scientific research community in favor of free and unrestricted use of genome sequencing data in biomedical research prior to data being published (Wikipedia, n.d.).

The NIH leadership at that time, with Dr. Francis Collins as the Director of the National Human Genome and Research Institute (NHGRI), along with other funding agencies, supported the adoption of this agreement to increase access and sharing of research resources to the public. The Fort Lauderdale agreement served as the foundation for promoting data sharing policies at the NIH. Building on previous data sharing policies, the NIH released a final statement on February 2003 (NOT-OD-03-032)[5] that serves as the current “general” policy supporting NIH’s commitment to sharing results and research resources generated from research that it funds and making those results and data available to the public. Figure 2 below shows a compressed

[5] https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

historical timeline of the release of data sharing policies at NIH to date, with a focus on genetic and genomic data sharing policies.

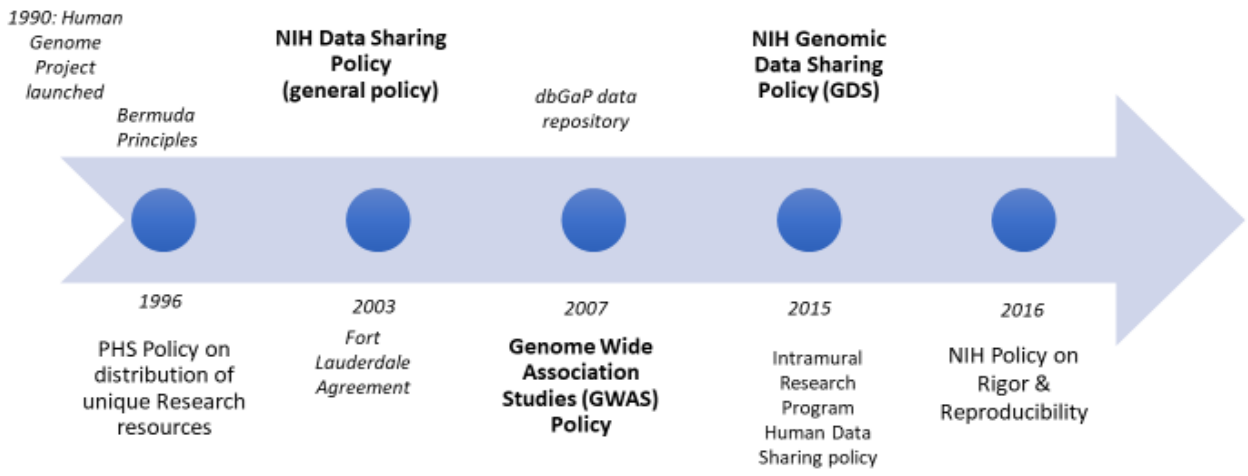

Figure 2: Historical timeline of data sharing policies and laws

More Recent U.S. National Initiatives and Policies Driving Data Sharing

Data sharing and open access have become a national conversation among policy makers, the research community, funding agencies, the private sector and the general public especially around genomic testing and privacy concerns with the sharing of data. In February 2013, the Office of Science Technology and Policy (OSTP) within the Executive Office of the President sent a memo to all federal agencies calling for increased access to results of federally funded scientific research (Holdren, 2013). Subsequently, President Obama's *All of Us* Research

Program and Vice President Biden's *Beau Biden Cancer Moonshot* initiative were launched in January 2015 and January 2016 respectively. Both initiatives call for increased data sharing among researchers, research participants and health care providers to help generate new and innovative ideas, and improve treatment and prevention strategies.

On December 13, 2016, Congress passed the twenty-first century Cures Act that will increase funding ($6.3 billion in funding) for biomedical research, with most of the money earmarked for NIH to promote the *Cancer Moonshot* and *All of Us* Research Program initiatives, as well as the BRAIN Initiative to improve our understanding of diseases such as Alzheimer's. This bill also gives the NIH Director, Dr. Francis Collins, the authority to mandate the sharing of scientific data by all NIH-funded researchers to advance rapid scientific progress in biomedical research. The NIH is currently exploring new policies that would be an updated and expanded version of the NIH 2003 general Data Sharing Policy.

This dissertation research study will shed light on institutional / organizational level factors that facilitate or hinder data sharing, as well as processes to overcoming the barriers associated with knowledge and information transfer in biomedical research conducted across regions of the world. These informed strategies may influence policy and practice of data sharing among science teams and collaborative research groups.

B. Statement of the Problem

The United States (U.S.) national initiatives such as the *All of Us* Research Program and the *Beau Biden Cancer Moonshot* initiative call for increased data sharing among researchers, research participants and health care providers to help generate new and innovative ideas, and improve treatment and prevention strategies. Most recently the 21st Century Cures Act, a bipartisan bill was signed into law on December 13, 2016 by President Obama includes several

measures that will "cut bureaucratic red tape that slows the progress of science, enhance data sharing and privacy protections for research volunteers, improve support for the next generation of biomedical researchers" (Hudson, 2017).

Data sharing policies at NIH have evolved over the years, since 1996, with the most recent being the Genomic Data Sharing (GDS) policy which became effective in January 2015. Despite these policies, data generated from NIH-funded research could be shared more widely in public or controlled-access databases, with the more recent exception concerning sharing of genomic data, and clinical trial data. There are systems, dedicated resources and clear processes in place for researchers to submit or access other genomic data in controlled-access data repositories. Although these facilitate the implementation of the NIH GDS policy, it is not relevant to the sharing of non-genomic data such as epidemiological data, which lag behind.

While the recent law mandates the sharing of all research data, in the absence of clear guidelines for implementation, an understanding of the challenges and how to address those challenges, it will be difficult to enforce sharing of data generated from NIH funded research in public or controlled-access databases or data repositories. The lack of data sharing by researchers in public or controlled-access databases is a leadership issue that has garnered national level interest. Current gaps in cancer research could benefit from leveraging existing data and knowledge to advance scientific progress to prevent and find a cure for cancer.

C. Purpose of the Study

This dissertation focused on the successes and challenges of data sharing practices and policy implementation NIH using a case study approach. Despite challenges, the NCI has had successes with implementing the GDS policy for the past couple of years which requires investigators to submit their genomic data in dbGaP. The goal of this study was to learn from

the implementation of existing data sharing policies primarily at the NCI in order to enhance sharing of epidemiological data generated from research funded by EGRP. In addition, this study provides recommendations for enhancing data sharing practices and implementation of policies among NIH-funded researchers. Understanding NCI's experiences, strategies for facilitating sharing, the challenges with researchers sharing data in public or controlled-access databases or data repositories and how they responded to those challenges will be insightful for EGRP as well as all other institutes at NIH and other research organizations dealing with similar issues.

D. Research Questions

This dissertation focuses on cancer research studies funded by NCI/EGRP with the following specific research questions.

1. How do organizational / institutional level factors facilitate or hinder the sharing of research data in public or controlled-access databases or data repositories?
    a) What are the organizational / institutional level factors that facilitate sharing of research data in public or controlled-access databases or data repositories?
    b) How do these organizational / institutional factors facilitate the sharing of research data in public or controlled-access databases or data repositories?
    c) What are the organizational / institutional level factors that hinder sharing of research data in public or controlled-access databases or data repositories?
    d) How do these organizational / institutional factors hinder the sharing of research data in public or controlled-access databases or data repositories?
2. What are opportunities for improving / enhancing the sharing of federally funded research data in public or controlled-access databases or data repositories?

3. How can what has been learned from genomic data sharing be transferred to epidemiological data sharing?
    a) What has been learned from genomic data sharing that could support epidemiological data sharing?
    b) In what ways can these lessons learned support epidemiological data sharing?

E. Leadership Implications and Relevance

i. Funding Agencies and Research Organizations

Data sharing has leadership implications in terms of NIH's leadership role in biomedical research and policy development. The NIH Office of the Director (OD) is responsible for developing and implementing data sharing policies for NIH funded research. According to the Director of OSTP (Holdren, 2013), "federal agencies investing in research and development must have clear and coordinated policies for increasing such access [to federally funded published research and digital scientific data]." This helps to maximize investment in scientific research to achieve optimal public health outcomes and clinical benefits.

Data sharing is complicated and has a down-stream and up-stream impact on many levels. It has direct impact on the researcher who generates and provides the data, the data users, patients who provide their data through participation in research, health care providers, policy makers and institutional / organizational leadership. The findings of this study will prove valuable to the NIH by illuminating current facilitators or barriers to data sharing and how they can help maximize the value of data generated in scientific research which are funded by tax payer dollars.

Understanding the underlying challenges with sharing research data funded by NIH will be useful in developing recommendations for how to enhance data sharing policies, particularly, the sharing of epidemiological data in NCI/EGRP-funded studies. The results of this study can be applied beyond EGRP, NCI and the NIH, and on a broader scale can influence the development and implementation of data sharing policies and practices in research communities, other organizations, and other federal and non-federal agencies of comparable size, structure and that fund similar types of research as NIH e.g. research organization or foundations. This study will also elucidate best practices and models for effective policies and implementation guidelines, highlight opportunities to best maximize federal investment in scientific research to advance scientific progress and improve health outcomes of all populations. It will foster economies of scale in large collaborative research studies, presenting significant cost savings on the long run.

ii. Public Health Surveillance

Data sharing has implications for public health on the local, national and global levels. In public health surveillance, studies have shown that the limited sharing of data as seen in the SARS outbreak of 2003 resulted in a delayed response to prevent the spread of the virus. Similarly, during the 2014 Ebola Virus epidemic in West Africa, the limited sharing of viral sequences made the assessment of the virus's potential for mutations more difficult (Sane & Edelstein, 2015). These examples illustrate the need for enhanced data sharing practices and how the lack of adequate systems and process for sharing information could have a negative impact in public health. It is therefore important that public health leaders address the challenges related to the rapid dissemination of data and results during outbreaks (Yozwiak et al, 2015), to reduce spread of diseases and illnesses, through efficient and effective public health

interventions, technology and programs.

In addition, cancer registries, whether population-based registries or hospital based registries, collect and publish data and statistics on new cancer cases, mortality, survival and prevalence of certain risk factors. An example is the NCI Surveillance, Epidemiology and End Results (SEER) registries and CDC's National Program of Cancer Registries (NPCR). The cancer data available through cancer registries "enable public health professionals to understand and address the cancer burden more effectively.[6] " The data can be used to assess and address cancer burden at the local, state and national levels and also valuable for public health planning and surveillance work.

iii. Data Sharing as Technical and Adaptive Leadership Issue

Large collaborative research projects encounter major problems related to the sharing of research data. These problems can be classified as both technical and adaptive and are worth addressing because they impede scientific progress, resulting in delays in the pace of discovery, translation and implementation in clinical and public health practice. This may mean poorer health outcomes and increased disparities in health among diverse populations, nationally and globally. The technical challenges of data sharing have a negative impact on biomedical research and as noted widely in the literature, include the lack of adequate technology, infrastructure support, funding, policies and guidelines. The issues with technology during electronic data transfers and communication or information exchange through different media platforms, guidance on standardized data formats, and data collection methods are considered technical challenges.

[6] https://www.cdc.gov/cancer/npcr/index.htm

Data sharing in biomedical research is also an adaptive issue that negatively impacts the research being done and the way research is being conducted. This is evident in the complex dynamics of interpersonal and intrapersonal relationships between multiple stakeholders (e.g. scientists, institutions, governments, and the public); multiple disciplines; the variation use of terminology as well as in the interpretation of policies, expectations and requirements for sharing; and the decision-making processes around implementation of policies and the implications. The institutionalization of an adequate rewards-based system to incentivize researchers to share their data willingly is an ongoing issue in the scientific community; it is complicated, requires a systems approach with multiple key stakeholders, and is influenced by a variety of factors.

The political climate's effect on partnerships between and within institutions and agencies are indicative of power influences and power dynamics. Leaders can influence decisions and develop standardized research policies at funding agencies, inform efforts to share, translate and implement research and elucidate best practices and models. For example, a diverse group of stakeholders in academia, industry, funding agencies, journal publishers and industry convened and designed a set of concise and measurable principles to enhance the data sharing and data reuse (Wilkinson et al, 2016). These principles are known as the Findability, Accessibility, Interoperability, and Re-usability (FAIR) Guiding Principles; an international framework developed to increase findability, accessibility, interoperability, and re-usability of research data.

Tenopir (2011) says that "Underlying policies and practices have great influence on encouraging or inhibiting data sharing." Therefore, this study will help inform research policies and practices on a systems level that will promote the field of biomedical and public health

research and improve health outcomes. Depending on the type of research being done, the design and method, researchers need to access data beyond what they have collected in their own studies to effectively replicate, reproduce, validate findings and address new scientific questions with increased rigor in their data collection and analyses. This will impact not just the quality of research that is being conducted, but has an influence on the overall health and well-being of the public.

II. CONCEPTUAL AND ANALYTICAL FRAMEWORK

A. Literature Review

i. Overview

"If rewards of big data are to be reaped, then researchers who produce those data must share them, and do so in such a way that the data are interpretable and reusable by others" (Borgman, 2012). Before we can understand the value of sharing data in scientific research, it is important to have a common understanding of what data sharing means. Data sharing is an elusive concept, a complex issue that involves multiple stakeholders at various levels (Figure 3).

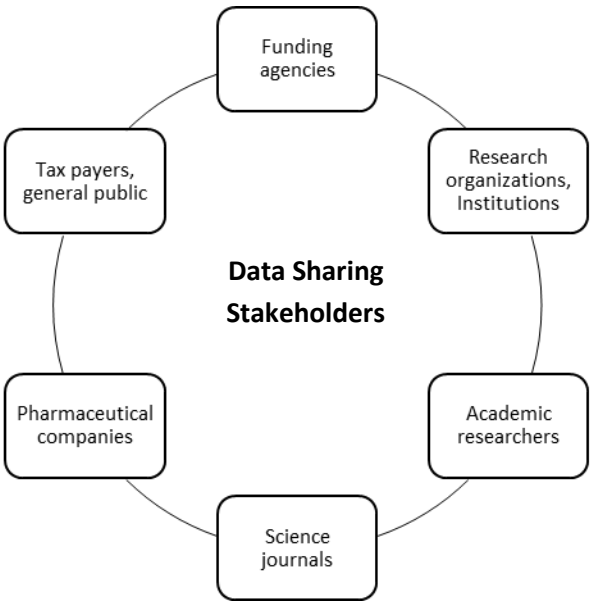

Figure 3: Multiple stakeholders involved in data sharing in scientific research

"Data sharing includes the deposition and preservation of data; however, it is primarily associated with providing access for use and reuse of data" (Tenopir, 2011). Borgman (2012) describes data sharing as the release of research data for use by others and may entail private exchange upon request to deposit datasets in a public data collection system through websites and as supplementary materials to journals. It could also be "as varied as announcing the existence of data, posting them on a website, or contributing them to a richly curated repository (Borgman, 2012)." In biodiversity research, data sharing is described as "the practice of making one's data available to others or reusing it again for subsequent analysis … includes persistent data storage, i.e. sustainable repositories for long term data storage are needed" (Enke et al, 2012).

Why Share Data?

In this era of "big data", researchers are generating an unprecedented amount of data through use of novel methods and technologies and which other researchers may reuse for new scientific discoveries and innovations (Borgman 2012). Pharmaceutical companies depend heavily on published research from academia for ideas to develop new drugs and treatment, and repeating the experiments to see if they get comparable results (Harris, 2017). The NIH, the largest funder of biomedical research is on the forefront of ensuring that there is transparency and direct access to data from federally-funded research and that the value of the data is maximized to benefit the public. NIH believes sharing data is critical and that the value of these data is maximized to benefit the public. The directive for sharing research data comes from the top-down; from the White House to the funding agencies and the scientific community.

Harris (2017) in his book referenced a March 2012 hearing at Congress where a senator challenged Dr. Francis Collins, NIH Director, to require rigor, reproducibility and replication of

studies. The reason is the known fact in the literature that shows that many published results in peer reviewed journals are not reproducible. This was the impetus for releasing the NIH policy on Rigor and Reproducibility[7] in January 2016 mandating "scientific rigor in the design and conduct of proposed research studies, as well as reproducibility of findings by multiple researchers to validate original results." According to the author (Harris, 2017), Dr. Collins supported broad sharing of genomic data during the Human Genome Project to prevent the potential of private companies hoarding and patenting human genes.

The value of data sharing has been well documented in the successes of many scientific research discoveries in biomedical research and primarily are to "(a) reproduce or verify research, (b) make results of publicly funded research available to the public, (c) enable others to ask new questions of extant data, and (d) advance the state of research and innovation (Borgman, 2012)." Sharing of research data does not only help elucidate problems associated with large datasets such as missing data, noise, etc., all of which affect the analyses and interpretation of findings, it helps with improving the quality of the data itself (Polina et al, 2012).

Data sharing promotes rigor, reproducibility and replication, thereby cutting down on scientific and methodological errors as evidenced in Tsodikov et al (2017), even before results get published. A poorly designed study cannot be reproduced or the results validated, and therefore lacks rigor. The rigor and robustness required in good scientific methods is fostered by transparency and openness in terms of raw data, statistical methods or source codes necessary to understand, develop or reproduce published research (Wikipedia, n.d.). Broad consensus in the biomedical research community is that "research data sharing is a primary ingredient for ensuring that science is more transparent and reproducible" (Vasilvesky 2017). In his book,

[7] https://grants.nihg.gov/reproducibility/index.htm

(Harris, 2017, p.145) Harris echoes the same sentiment that "increasing transparency could go a long way toward reducing reproducibility problems that plague biomedical research."

Data sharing provides researchers the opportunity to test existing hypotheses, generate novel ideas and hypotheses (Borgman, 2012; Vasilevsky et al, 2017), use data for meta-analysis and in teaching (Fox 2014; Pisani et al, 2010) and to "re-examine the data and form our own opinion of their meaning" (Butlin, 2011). According to the NIH Data Sharing Policy and Implementation Guidance,[8] data sharing allows researchers to 1) ask new scientific questions not already asked by the primary investigators, 2) explore a variety of alternative hypotheses, methods of data collection, measurement and analyses, 3) promote new research and training of new researchers, and 4) create new pooled data sets from multiple data sources.

The sharing of scientific data plays a significant role in determining the future of scientific researchers (Fox et al, 2014) and development of new drugs for treatment and cure for disease (Harris, 2017). Since pharmaceutical companies rely on findings from research to test new drugs they've developed, it is critical that errors in research methods and analysis be uncovered prior to publishing findings. This can be done through sharing of data. Harris (2017, p. 8) says that "sharing of data can accelerate progress in biomedical research by helping researchers discover errors more quickly."

From a clinical trials perspective, the Food and Drug Administration (FDA) requires scientists to register their hypothesis in advance in ClincalTrials.gov, an NIH repository, so they can "demonstrate that their studies are indeed confirmatory" (Harris, 2017, p.150). This policy forces transparency in clinical trials and ensures that drugs with negative results are published in the literature, which generally is not the case if the results were not favorable to the drug being investigated (Harris, 2017). Similarly, the Open Science Framework is a repository where

[8] https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#goals

researchers deposit not only their raw data but also the algorithms and methods used in the analysis. Tuchman (2016) says that "sharing clinical trial data should also make progress more efficient by making the most of what may be learned from each trial and by avoiding unwarranted repetition. It will help to fulfill our moral obligation to study participants, and we believe it will benefit patients, investigators, sponsors, and society."

The issue of whether to share or not to share research data, what to share, how to share and with whom to share, is one that has become an ongoing national debate among key stakeholders and must be addressed. For the purposes of this dissertation research, sharing research data in public or controlled-access repositories is the phenomenon of interest; a problem noted widely in practice and in the literature. Understanding the organizational / institutional level factors that facilitate or hinder data sharing will inform data policy and data sharing practice, and help design optimal approaches for wider data sharing in biomedical research (Borgman, 2012).

Established Frameworks for Sharing Research Data

There are two frameworks that have been established and accepted by the scientific community, and both share the same beliefs as NIH regarding the sharing of research data: 1) the Global Alliance for Genomics and Health (GA4GH)[9], and 2) the Findability, Accessibility, Interoperability, and Reusability (FAIR) data principles (Wilkinson et al, 2016). The GA4GH, a collaborative global partnership, was established in 2013 to promote broad sharing of genomics data in the scientific community. GA4GH considers data sharing critical to advancing scientific progress and improving health outcomes of populations and has developed a framework for responsible sharing of genomic and health-related data. The core elements include: "transparency; accountability; engagement; data quality and security; privacy, data protection

[9] https://genomicsandhealth.org/about-global-alliance

and confidentiality; recognition and attribution; risk-benefit analysis; sustainability; education and training; accessibility and dissemination" (Knoppers, 2014).

Similarly, the FAIR data principles (Wilkinson et al, 2016) is another framework that guides researchers in the sharing of research data. These data principles support the importance of good data sharing practices and data management within the scientific community, and help guide researchers in addressing barriers to data sharing. They were developed through a collaborative network of representatives from academic institutions, funding agencies, industry and journal publishers, with the goal to support and enhance data reuse and reproducibility of research findings.

The Field of Genomics as an Exemplar of Data Sharing in Scientific Research

"The field of genomics is regarded as the leader in the development of infrastructure, resources and policies that promote data sharing." (Kaye et al, 2009). According to Borgman (2012), "Data sharing activities appear to be concentrated in a few fields, and practices even within these fields are inconsistent." The variability of data sharing practices across scientific disciplines and fields of research pose some challenges to advancing the progress of scientific discovery. These include concerns with human subject privacy, variation in data standards, interoperability of technological infrastructure, data ownership issues and desire for attaining competitive advantage among researchers, etc. Part of the problem is the variation in forms and types of research data and the way research data is collected and interpreted by scientists (Borgman, 2012).

Most scholars agree that there are benefits of data sharing (Nelson, 2009) however this view is not unanimous (Mueller-Langer, 2014) and varies across disciplines as some disciplines are more apt to share data e.g. genomics, astronomy, economics, physics, geology, etc.

(Borgman 2012), as opposed to the biological sciences. There tends to be more affinity for and success with open access to data in a few disciplines such as geophysics, mathematics, computer science, and astronomy, atmospheric science and oceanography compared to fields such as in the sciences, wildlife ecology and many social sciences (Nelson, 2009). Although fields such as astronomy, social sciences and natural history have demonstrated benefits of data sharing (Poline et al, 2012), accessing data has not been very easy for researchers, except in the field of genetics (Borgman, 2012; Butlin, 2011).

In contrast to other biomedical disciplines, many neuroscientists have not quite embraced data sharing and are therefore unwilling to share, and ambivalent about the practice of open sharing of experimental data (Ascoli et al, 2017). The authors note that "brain science trails behind other scientific disciplines in terms of open data initiatives" and "many data sets remain unavailable to the broader research community, causing a waste of time, money and scientific opportunities (Ascoli et al, 2017)."

The NIH Brain Research through Advancing Innovative Neurotechnologies® (BRAIN) Initiative, an interdisciplinary effort among several agencies, was launched on April 2, 2013 by President Obama to accelerate discovery of novel methods and innovative technologies to improve our understanding of the human brain and advance prevention and treatment strategies for brain disorders. According to section II of the BRAIN 2025 Report,[10] one of eight high priority research areas is *Maximizing the value of the BRAIN Initiative: Core Principles*, which includes the establishment of platforms for sharing data as well as methods and software. Consistent with Ascoli et al (2017) is a statement in the report that "new data platforms would also encourage changes in the culture of neuroscience to promote increasing sharing of primary data and tools." NIH reportedly spent approximately $6 billion in funding for neuroscience

[10] https://www.braininitiative.nih.gov/2025/

research in 2015. With such a huge investment, it is only prudent that researchers make this knowledge and data available publicly.

The field of genomics sets the standard for how to advance scientific progress through data sharing (Poline et al, 2012). It has pioneered the era of open data access and the data sharing movement and leading "the development of infrastructure, resources, and policies that promote data sharing" (Kaye et al, 2009). For example, human genetic variants affecting health, response to treatment, etc. are made available through the Haplotype map (HapMap) and 1000 Genomes Project. Progress on this front is partly attributed to transparency and clear expectations. Following the Bermuda agreement in 1996, which many scientists noted as "a defining moment for genomics" (Nelson, 2009), as well as technological advances over the years, it has become evident that "genetic research is advancing faster than any other area of biomedicine (Pisani & AbouZhar, 2010)". Leaders at the meeting developed a set of agreements known as the Bermuda principles that required "sequences longer than 1,000 base pairs be made publicly available within 24 hours" (Nelson, 2009).

"Since the early days of sequencing, the whole international community has recognized the value of depositing sequence data in publicly available databases (Butlin, 2011)." Data submitted by academic institutions represent the majority (70%) of data submitted in the NIH dbGaP, followed by data submitted by "non-academic research or non-profit organizations (22%), and government research agencies and health departments (8%)" (Paltoo et al, 2014).

One of the many documented successes and benefits of genomic data sharing is seen in the scientific discovery that resulted from sharing data from GWAS studies. Since the NIH GWAS Data Sharing Policy was instituted in 2007, more than 900 secondary analyses of GWAS data has been published (Arias et al, 2014). Twenty percent of the GWAS publications focused

on cancer, twenty percent on methods development, fifteen percent on mental health disorders, and seven percent on cardiovascular disease (Paltoo et al, 2014). The secondary analysis of deposited GWAS data contributed to identification of "associations between the human leukocyte antigen and Parkinson's disease, for example, as well as other previously unknown associations" (Arias et al, 2014; Paltoo et al, 2014). In research studies that require understanding of disease progression and subsequent treatment, such as Parkinson's, data sharing can help with the discovery of clinically relevant biomarkers the disease. The success of this is going to be dependent on availability of datasets and biosamples from observational studies and interventional trials which Frasier (2016) describes as being "kept under lock and key"; a culture that must be changed to advance scientific progress.

Further description of the differences between data sharing in genomics compared to other fields such as epidemiology is listed in Table 1. This table highlights some of the key factors contributing to ease of data sharing in the field of genomics. One author describes data sharing and collaboration in epidemiology and public health research as slow-paced and lagging behind some fields such as genomics (Pisani et al, 2010). Perhaps public health research would advance more quickly if we followed a similar path and "if we squeezed more scientific and policy insights out of data that have already been collected (Pisani & AbouZhar, 2010)."

TABLE I: DATA SHARING IN GENOMICS COMPARED TO OTHER FIELDS

	Genomic Data Sharing	**Other Fields**
Repository	**Clear repository in place – dbGaP** - NIH-wide repository in place (dbGaP) since 2007 (Paltoo et al, 2014)	**No clear repository NIH wide for epidemiological data except**: - BioLINCC repository for NHLBI epidemiological and

		clinical trial studies. - dbGaP for genomic and phenotypic data
Policy	**Clear policy** - Specific to genetic and genomic data (GWAS data sharing policy, Genomic data sharing policy)	- NIH general data sharing policy of 2003 - Not specific to epidemiology data
Expectations	**Clear expectations stated in policy**	**Expectations could be ambiguous**
Guidance	**Clear guidance** - Processes, templates to facilitate sharing	**Guidance is not clear**
Data structures	**Clear data structures** - Data elements and sequence structures, with very slight variation in how reported	**Variation in data elements**
Systems / Structures for Administrative support	**Clear systems of support and resources in place** - Data Access Committee, Genomic Program Administrators	**Systems for support are not explicit**
Billing	**Clearly stated to include cost** of data sharing in grant application	**Encouraged to include cost** of data sharing in grant application

Funding agencies are making a concerted effort towards more open and broad data sharing. For example, the Open Science Prize initiative jointly sponsored by the UK Wellcome Trust, the NIH and Howard Hughes Medical Institute was created in 2017 to encourage the development of innovative and novel tools and platform for open data science, and use of data to advance discovery and public health benefit.[11] Of the 3,730 votes received representing 76 countries, the winner was the Real-Time Evolutionary Tracking for Pathogen Surveillance and Epidemiological Investigation, a project to "promote open sharing of viral genomic data and harness this data to make epidemiologically actionable inferences.[5] " This is a model that could be adopted in other fields of biomedical research to foster data sharing and collaboration.

[11] https://www.opensciencеprize.org/

Borgman (2012) says it is not totally clear what the criteria for identifying, sharing and re-using data are, however, progress is being made in better understanding why researchers are not readily sharing their data. These are indications for more research on best practices and lessons learned on data sharing from fields that do share data consistently and where data are consistently reused. This provides the rationale for the focus of this dissertation research; to learn from NIH GDS policies and practices about the factors that influence broad data sharing at the organizational / institutional level and how these can be addressed. This study provides a deeper understanding of the factors that facilitate and hinder sharing of data among researchers, and inform the development and implementation of policies, as well as systems to support data sharing.

Factors Influencing Data Sharing in Scientific Research

There are many factors critical to facilitating data sharing in scientific research which can be mapped to various levels of the socio-ecological / systems model, i.e. 1) the individual; 2) the interpersonal; 3) the organizational / institutional; and 4) the systems levels. This model is the basis of the conceptual framework developed for this dissertation research which is described more in-depth later. It serves as a guiding tool to understand the interrelationships between the different stakeholders and to help identify recommendations for targeted strategies that are appropriate for addressing data sharing challenges at and across the levels. It is hoped that these recommendations will lead to an improvement in data sharing practices among researchers.

One study described by Enke et al (2012), highlights some of the challenges researchers face with sharing data across the systems spectrum. An online survey of over 3,000 researchers in the field of biodiversity / life sciences were surveyed about their willingness to share data. The results indicate that over 80% agreed with a 'YES' response indicating willingness to deposit

data in publicly accessible databases. Some of the main individual level barriers described in this study are: fear of loss of control (53%) and huge time commitment and effort to prepare data for sharing (50%).

Other major concerns from the study include the lack of acknowledgement for sharing data (43%) which is a perceived fear that implores intervention at the organizational / institutional level; and lack of standard data formats for managing data sharing (34%). 31% of the respondents feared misinterpretation of conclusions from reuse by someone else and 27% of the respondents had privacy and legal concerns. Lack of knowledge about existing public repositories (40%) and inadequate technical infrastructures were disincentives for sharing, both of which require support from the organizations / institutions. In addition, the reluctance among researchers to deposit data in databases without funding for long-term support and maintenance highlights the relationship between individual level factors (e.g. motivation or willingness) and organizational / institutional level factors that support and promote data sharing practices among researchers.

This dissertation research complements the existing work that has been done in this area, such as the one described above, filling in gaps in the literature and practice by specifically focusing on organizational / institutional level factors that facilitate or hinder the sharing of federally-funded research data in publicly accessible or controlled-access databases or data repositories. There are a few studies that have already looked at individual / interpersonal factors that facilitate or hinder the sharing of research data, however addressing these factors require organizational / institutional level systems and supports; they are not mutually exclusive. The next section highlights some of these factors, beginning briefly with individual and interpersonal

level factors that affect data sharing and illuminate opportunities for organizational / institutional influence to increase and enhance data sharing practices in research.

ii. Individual / Interpersonal Level Factors Influencing Data Sharing

Time, Effort, Cost, and Expertise

There are relationships that exist between individual and interpersonal level factors that affect researchers' motivation, willingness and attitudes towards data sharing. These include their perception of risks and benefits associated with sharing. For example, there's the notion that sharing data can negatively impact career opportunities and sometimes these concerns outweigh the perceived benefits of data sharing that come with accelerating new discoveries and curing disease (Nelson, 2009). It takes time and effort, and money to prepare, document and format data for submission, in a logical and systematic way that is meaningful to other users.

In addition, researchers' knowledge, skill level and ability to access and submit data in data repositories could be factors in determining whether they readily share or not share data. These concerns are barriers that are echoed throughout the literature (Nelson, 2009; Borgman 2012), as well as concerns with availability and access to resources to support data management.

Data Quality, Integrity, and Expertise

Researchers are often leery about sharing data they collect from their original research. One of the concerns is the fear that users not involved in the "generation and collection of the data may not understand the choices made in defining the parameters" (Longo et al, 2016) and therefore could misinterpret or misuse the data. The fear that data will be misused or the dilemma of whether to share raw data versus cleaned data is driven by the assumption that users if given access to other researchers' data, whether raw or cleaned, will have challenges with accurately interpreting and using the data, leading to inaccurate conclusions or correlations

(Nelson, 2009). This could be due to the lack of expertise or lack of documentation of the datasets being shared but it is not entirely clear. Some will argue that this seems like a reason to share data because it provides the opportunity to train researchers, as well as identify and address any methodological or analytical issues thereby testing the validity and reproducibility of the data and results.

There are differences in data collection, analysis methods and study populations such that one cannot generalize the criteria for defining variables across studies (Longo, 2016). This could be as simple as researchers using different measures in their studies; particularly true for epidemiological or observational studies. In a recent re-analysis of data from two previous prostate cancer screening trials, the authors used same data from the two original studies (PLCO and ERSPC trials) but different methods to test validity and reproducibility of findings and disproved the original results (Tsodikov, 2017). Making data available to others promotes rigor and transparency in research but despite this, researchers are reluctant to share their data.

Trust

The nature of collaborative research is such that requires direct access, transparency, communication, shared data and resources to best maximize the value of data shared, for rapid scientific discovery and translation of research results into public health and clinical interventions. Trust is a very important factor that needs to be established among collaborative research groups or science teams because it impacts their willingness to share data and minimizes the tendency to mine data for their own personal benefit. Researchers tend to be more selective in who they are willing to share their data with, primarily giving preference to those within their research network or "immediate area of specialty" as opposed to the public (Borgman 2012). These individuals or groups of researchers who share a "community of

interest" (equivalent term used by the National Science Foundation) are said to have the right skills, knowledge, abilities and expertise to interpret the data in a way that does not compromise the [intended] meaning of the original data and subsequently, the analysis.

Establishing trust between the researchers is an important interpersonal factor that influences researchers' motivation to share data. Many patients enrolled in clinical trials consent to have their data shared so it can help others. More researchers in turn need to exert the same level of trust with colleagues when it comes to sharing data not just to advance science, but to honor the sacrifices of research participants who donate their data. However, this is not that simple. Literature shows that trust is impacted by personal (versus altruistic) goals and interests such as career advancement, funding opportunities, recognition within the scientific community, and publication in peer-reviewed journals.

Competition

To fully exploit the value of their data and ensure that they are the first to publish with their primary data as well as use it for subsequent research, scientists tend to "strategically delay" the time of submission of papers (Mueller-Langer et al, 2014). This is tied to the competitive advantage associated with the data they have worked hard to generate. At the same time, studies have shown that data sharing might increase value of a publication for its author by increasing its credibility as their work may often be replicated by others as well as facilitate new research. It may also generate positive effects for the scientific community as the data could be used for subsequent research and validation (Mueller-Langer 2014). This is a benefit to sharing data that could motivate researchers and alleviate their concern with lack of recognition and acknowledgment for their work.

iii. Organizational / Institutional Level Factors Influencing Data Sharing

This dissertation primarily focused more in-depth on organizational / institutional level factors that will inform the design of organizational / institutional recommendations for wider data sharing in the scientific community. These include: 1) culture of scientific research, 2) academic institution / organizational practice and culture including incentives / rewards, ethical and legal issues, 3) regulatory laws and policies from funding agencies and science journals, 4) technological infrastructure, and 5) support including leadership, administrative and financial.

The Culture of the Scientific Research Enterprise

"Over time, the data culture had changed to one in which research collaboration, facilitated by the Internet had led researchers to acknowledge the need to share data" (Sturges, 2015). This is also true in the field of genomics where the motivation to share stemmed from the need to increase sample size to conduct large genetic studies and the rapid increase in technological advances. However, data sharing policies challenge the current scientific cultural norms that "rely on publications of research as a means to garner recognition and professional success" (Arias et al, 2014), which hinders change. There is the tendency within the scientific community to resist change if it affects the ability to publish on original research data, given that the primary measure of success is determined by the quantity and quality of publications as well as the citation of records (Pisani et al, 2010; Stanley 1988).

This resistance or reluctance comes from the competitive culture in the scientific community which is a major barrier to sharing research data because researchers are competing for an array of resources, grant funding, publications in the high impact journals and the recognition that comes with it, opportunities for career advancements, and attainment of promotion and tenure (Harris, 2017). They must be mindful of how they choose to invest their

time and resources to achieve these personal goals, as well as how they balance that with societal and public benefits of sharing research data.

Some scientists believe that resources spent on data documentation for use by others are resources not spent on data collection, publication fees, research activities and related research needs (Borgman, 2012). Their desire to protect their scientific lead, preserve the proprietary value of their data (Pham-Kanter et al, 2014) and their ability to publish (Zinner et al, 2016) seems to outweigh their desire to have their data re-used to generate new hypotheses and answer new research questions beyond that of the initial study. The NIH policy on Rigor and Reproducibility[12] released in 2016 encourages scientific rigor in the design and conduct of proposed research studies as well as reproducibility of findings by multiple researchers to validate original results. Sharing data will improve rigor and transparency in research methods, increase reproducibility and validity of findings, and lead to knowledge generation towards improving health outcomes.

Academic Institutional Practices and Culture

Obtaining promotion and tenure at the university is a big goal for all scientists in academic research institutions because of the security it provides in their careers as it relates to grant funding, for instance. This pressure stems from the culture and practice of universities who base their decision on promotion and tenure of scientists on the number of publications in top peer reviewed journals. This is viewed as a mutual benefit for both parties because on one hand, it gives the scientists recognition and reputation among their peers in the scientific community and on the other hand, the universities get the reputation associated with the investigator's discovery which could translate to more grant funding.

[12] https://grants.nih.gov/reproducibility/index.htm

Data sharing by contrast receives almost no recognition but this is quickly taking a different turn. Changing the culture of science at the organizational / institutional level "from one where publications were viewed as the primary product for the scientific enterprise to one that also equally values data (Nelson, 2009)" is one that's being promoted by the NIH and science journals. The most recent launch by Nature Research is that of an online peer-reviewed journal, *Scientific Data*, that promotes wider sharing, data reuse and credits those who share. This effort emphasizes that scientific research is no longer solely about the publications as an end-product of research, but also about the data that is produced from the research.

Researchers are geared primarily towards publication of articles, "as they invest their time and resources into activities that can increase their reputation" (Friesike, 2015). The result of this is pressure to publish new results as much as possible and less value and no acknowledgment is placed on data sharing activities which are time consuming (Ioannidis & Khoury, 2014).

In a survey of 1,564 academic researchers, while majority agreed with the principle that sharing of research data advanced scientific progress (83%) and that researchers should make their data available to the public (76%), very few of them (13%) indicated that they had shared their own data in the past (Friesike, 2015). The major concern among these researchers was that if shared, their data will be published by other researchers which will limit the recognition and attribution to them as the data originators. A small number of respondents (12%) were less concerned about any criticism with their data or being proven wrong (Friesike, 2015) than the recognition that impacts their careers if others used their data to publish first.

Incentives and Rewards, Acknowledgment and Recognition

Despite the immense efforts of data collection by data originators and data sharing policies from funding agencies (e.g. NIH, the UK Wellcome Trust, Bill and Melinda Gates Foundation), biomedical journal policy of the International Committee of Medical Journal Editors (ICMJE), and the legislature of the 21st Century Cures Act, "there is rarely academic recognition or reward for data sharing itself (Bierer et al, 2017)." As a result, researchers are not generally sharing their data in publicly accessible databases or data repositories.

Some researchers are not motivated to share because of the perception that the cost in terms of time, funding and risks outweigh the benefits in terms of publications and career advancement. There are no clear incentives besides funding and the reward of contributing to scientific advancements, to motivate researchers to share data. Even with funding, there is the challenge of competition within the scientific community which reduces their chances of success.

The concern among researchers when it comes to data sharing is that they don't get the credit and acknowledgement from their institutions when their data is used and are therefore not motivated to share. Researchers spend a lot of time collecting their data and the reward of this effort is to be the first to publish the findings in high impact journals. While most researchers are supportive of contributing their data towards the advancement of science and for the greater good they are often conflicted because one of their major goals is advancement in their careers which is heavily dependent on the number and quality of their publications.

Interestingly, the credibility and reputation of researchers who publish are increased when they share because of opportunities for replication by others and generation of new research (Mueller-Langer, 2014). Despite this, data creators fear that others will gain competitive advantage over them by using their data for subsequent research. Friesike et al (2015) use the

term reputation economy to describe the state of academia where individual researchers place the highest value of their career on getting recognized by their peers. Therefore, researchers tend to focus more on publishing articles to increase their reputation, and less on data sharing because they don't perceive it nearly as beneficial in terms of time and resources. The authors indicate that there is little motivation to share although 76% of respondents in the survey support data sharing. A major concern to sharing data expressed by 80% of respondents is that others may publish with it, thereby minimizing chances of getting recognized for original data and work.

Proposed Ideas for Strategies and Incentives for Data Sharing

Getting researchers to share data and acknowledge individual contributions within the scientific community is "not a matter of more regulation and guidelines, but of developing norms that become an intrinsic part of a new scientific culture, in which people can trust each other because the rules and obligations are known at the outset (Kaye et al, 2009)." The current data sharing policies "either try to motivate researchers to share by invoking the common good [the "carrot"] or they force them to share through mandating data sharing policies [the "stick"]" (Friesike et al, 2015). It is not clear in the literature nor in practice if the "carrot" or the "stick" or a combination of both works and is something that needs to be addressed from an organizational perspective, to enhance data sharing among researchers.

This is important yet a challenging task because changing the scientific culture and decision-making around data sharing incentives will require joint efforts of key stakeholders who influence or are influenced directly or indirectly e.g. funding agencies, journal publishers and editors, academic and research institutions, industry, researchers, and the scientific and non-scientific research community including research participants. This is an ongoing challenge at the NIH because while the policies apply to all NIH-funded researchers, the implementation is

left to the individual institutes and vary widely. It is not clear how to effectively address these issues to enhance data sharing among researchers.

Data originators are not getting proper credit or acknowledgment for their data when reused by others. Without the acknowledgment and recognition, researchers fear their data and hard work will be scooped, losing competitive advantage if data re-users publish findings first. They also fear they will be criticized by data users which can destroy their reputation in the scientific community. The quality of datasets is likely to be compromised because of the minimal effort put forth by the "unmotivated" researcher, which in turn impacts reproducibility and reusability of the data (Friesike et al, 2015).

A way forward towards better data sharing practices is to consider what systems and strategies that could facilitate the ease of data sharing among researchers, without sacrificing quality of the data and privacy of human subject participants. There are a few suggested strategies for incentivizing or rewarding data sharing that have been described in the literature. However, it is unclear whether any of these have been institutionalized, are in consideration or whether there is evidence that they are effective in motivating researchers to share.

Bierer et al (2017) suggest modifying "faculty reporting formats and promotion criteria" by academic institutions to recognize data authorship as significant contributions, and allowing inclusion of publications that cite data, and in grant applications to further incentivize researchers to share data. Similarly, Piwowar et al (2008) emphasize the need for sharing contributions to be included in decisions around hiring, promotion and tenure; encouraging investigators to include datasets as part of their listing of accomplishments on their CVs and grant applications. This is in addition to the adoption of a data sharing citation index.

Olfson et al (2017) propose instituting a common data sharing metric or the S-index, to "measure the number and impact of peer reviewed publications in which investigators have shared their data with other research groups." The recognition of data shared would allow data to be included and considered as part of researcher's overall contribution to research. The rationale for the S-index is that "study design, data collection, and data curation are as important" as data analysis and publication in research. Funding agencies and research institutions could use this metric to "quantify each investigator's history of data sharing as part of their overall scientific contribution" (Olfson et al, 2017).

Rowhani-Farid et al (2017) conducted a systematic review of over five-hundred articles that looked at strategies and evidence-based incentives that increased data sharing. Of these, the open data badge award was the only evidence-based incentive that increased data sharing. The adoption of the badge by the journal increased the rate of sharing twenty-fold. The authors say that it's not just important to identify strategies but to test them and figure out which work. Some examples of strategies (though not tested) in the literature they reviewed include use of policy, open data campaigns, encouraging collaboration among researchers, availability and access to good data / technology systems or a combination of these.

The PQRST (Productive, Quality, Reproducible, Shareable, and Translatable) metric is designed to evaluate the number of papers that include "shareable data, materials or protocols" (Ioannidis & Khoury, 2014). This will complement "the commonly accepted metrics for academic performance (the journal citation index, the Hirsch index and even altmetrics) which are all based on research article publications and citations" (Friesike, 2015).

Challenges and Opportunities with Reward Systems

While it has been recognized that incentives such as crediting data generators are key to promoting and implementing data sharing, "there has been no systematic implementation of a standard process and method to credit original data generators" (Bierer et al, 2017). This is where funding agencies such as the NIH and academic institutions could work together and see how to "modify systems for apportioning academic credit to better align incentives for data sharing with the advancement of science and medicine" (Bierer et al, 2017). In addition to collaboration with original data generators, organizations need work with the scientific community to develop and align implementable policies, meaningful incentives, and sustainable technology "to make it both easy and rewarding for researchers to share their data" (Koers, 2016). Such collaboration and alignment will improve data sharing practices and help with overcoming some of those barriers inherent in the sharing of original data among researchers.

Identifying where organizations such as the NIH (and academic institutions) can have influence, what the role of the NIH is with creating effective reward systems for researchers that share data and identifying what those barriers are will be critical to making advances in this area. Publishers play a key role because they have policies that require researchers to share their data before they can be accepted for publishing in the journals. They are also able to use technology to "link published articles with data stored in repositories" and have scientific journals that allow publications focused on datasets and methods (Koers, 2016), e.g. *Scientific Data* online journal.

Through awards funded by the NIH Big Data to Knowledge (BD2K) program, an initiative launched in 2014 to support and facilitate the sharing and re-use of existing complex biological data ("big data") and other digital objects such as software, BD2K is developing a data

discovery index (DDI)[13] to improve data sharing in biomedical research. In the implementation of the genomics data sharing policy, publications are checked to make sure that the submitter is acknowledged. This is an example where leadership in research and funding organizations have taken a systems approach and come up with creative and innovative strategies to address this barrier to data sharing and data re-use.

These ideas will hopefully, over time, influence a shift from the current scientific culture that is heavily driven by publications, towards a new culture where the data sharing contributions are recognized and rewarded. Olfson et al (2017) says that "If embraced by funders of science and leadership in academic medicine, the S-index would provide a potentially powerful voluntary means for encouraging greater constructive collaboration among investigators." An ongoing dialogue with the scientific community on the pros and cons of formal and informal data sharing is necessary to "articulate the norms required in specific situations, and to determine a fair and equitable way to share data but also acknowledge individual contributions (Kaye et al, 2009)." The Institute is beginning to engage is these discussions at the division level.

Support - the Role of Leadership in Changing the Culture of Science

In 1997, a report on data sharing, value and recommendation for full and open access to data was published by the U.S. National Research Council to promote an international norm that allows for sharing of publicly funded research data. Despite this recommendation, the culture of sharing has not been adopted or normalized in many fields of research. Funding agencies such as the NIH, Wellcome Trust, and Medical Research Council in the UK, to name a few, have developed formal data sharing policies that require all its funded researchers to share research data broadly (Poline et al, 2012).

[13] https://datascience.nih.gov/bd2k/funded-programs/resource-indexing

Literature shows that the cultural norms in scientific research impede the sharing of data broadly and therefore requires a systems approach to address the need for a cultural shift. It takes leadership to largely influence this shift, which can be done through provision of resources, administrative support and financial support. At the NIH for example, each institute is responsible for championing efforts around institutionalizing and implementing data sharing in NIH-funded research. One example is the development and support of dbGaP, a central controlled-access data repository created by the NIH National Center for Biotechnology Information (NCBI) for researchers to archive and access genomic and phenotypic data from human studies. This is evidence for how change starts at the top; critical if we want to make that culture shift among researchers even early on in their careers.

Policies that are designed such that researchers have a perceived benefit for sharing data are key in determining whether they share or don't share their data (Borgman, 2012). To facilitate this, it is important to get buy-in of key researchers and commitment from funders (Pisani et al, 2010) and academic institutions early in the development of policies. This can play a role in influencing change at the organizational / institutional level which hopefully will trickle down to change at the individual level, and the scientific research community in general.

Ten years ago, a commentary published in *Trials* titled "Whose data set is it anyway" described how the attitudes among clinical trialists and pharmaceutical industry groups was such that they considered it standard not to share their trial data after publication. Today, there's a change in the culture and attitudes where leaders of funding organizations and science journals (e.g. the UK Wellcome Trust, the NIH, the ICMJE) are recommending sharing of raw data. Despite this, the author says that "we are yet to see clinical trial data sharing become an

unquestioned norm, where, say a researcher can readily download a data set from a trial almost as easily as they can now download the trial publication" (Vickers, 2016).

In contrast, the field of genomics has been a driver in fostering a culture of data sharing practices among researchers. This could be attributed to the leadership of the scientific community and funding agencies starting with the development of the Bermuda principles and NIH genomic data sharing, and reflecting a changing view on scientific norms (Arias et al, 2014). Changing the culture of the scientific enterprise to share data more broadly is complex and complicated for more fields such as neurosciences and epidemiology, than others. This cannot be solved by just technical issues. It requires a systems perspective and application of evidence-based strategies.

Leaders need to prioritize data sharing and according to data sharing advocates, "the power to prod researchers towards openness and consistency rests largely with those who have always had the most clout in science: the funding agencies, which can demand data sharing in return for support; the scientific societies, which can establish it as a precedent; and the journals which make sharing a condition of publication (Nelson 2009)." This is evident in the support and leadership provided by the NIH Director and others during the creation of the Bermuda Principles in 1996 and the Fort Lauderdale Agreement in 2003. The agreement to make genome sequencing data available in publicly accessible databases was one of the driving forces behind the development of NIH data sharing policies under the guiding and supportive leadership of Dr. Francis Collins, the NIH director at that time and currently.

Other Institutes such as the NCI, NHLBI, NHGRI, NIA, etc., continue to implement data sharing policies in their institutes and among their grantees. Similarly, the Associate Director of EGRP has been leading efforts in the program to evaluate data sharing practices and policies that

can be applied to EGRP-funded epidemiology studies. One of the milestones in this process is the creation of Cancer Epidemiology Data Repository (CEDR),[14] designed to interface directly with dbGaP, as a repository for researchers to deposit and access epidemiology data from their EGRP funded studies. There are ongoing discussions within the program and with NCI stakeholders on the design and implementation to facilitate sharing of data generated from observational studies.

Ethical and Legal Issues

Academic institutions are inherently bound by ethical standards when it comes to human subject research. Scientific research involving human participants is guided by ethical norms and standards as described in the Belmont Report[15] of 1979. The Belmont Report describes the ethical principles and guidelines that protect the privacy of humans participating in all types of research. Violation of ethical conduct in research or a breach of privacy can have severe consequences which create a sense reluctance among researchers to share their data widely because they assume it increases the possibility of a breach or violation. Informed consents in research are critical because it ensures transparency in the research purpose, process and outcomes for all study participants; it is hoped that the study participants will have a clear understanding of what the research is about, how their data will be used or shared, etc.

The universal elements of consent documents include the purpose statement and description of the research; a description of any forseeable risks and discomforts to the subject; a description of research benefits to the subject or others; a disclosure of procedures or treatments that may benefit the subject; a statement of the extent to which confidentiality will be maintained; an explanation of potential compensation and medical treatment for research

[14] https://epi.grants.cancer.gov/CEDR/
[15] https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/

involving more than minimal risk; description of who to contact for research or subject related injury, rights or other questions; and a statement of voluntary participation in research with no consequences for refusal. However, "many informed consent documents do not mention the possibility of data sharing and pose a major barrier to sharing data (Poline et al, 2012)." NIH policy requires that the informed consent be consistent with data sharing. Older studies do not have data sharing plans built into their informed consent forms and therefore run a risk for privacy breach or if they must, will need to re-consent individuals which could be time-consuming and deterring.

Some IRBs will grant researchers permission to share retrospective data if they were able to re-consent participants however, this could be quite a huge administrative burden on researchers and their teams, in addition to completing the paperwork for IRB submission. On the other hand, "some IRBs are not willing to approve protocols requesting open data sharing" and as such institutional leadership could influence how IRBs review protocols with data sharing requests, as well as amendments to their existing protocols as it relates to data sharing (Poline et al, 2012). "When sharing data in a collaborative research team, data use agreements between the institutions are obtained early in the research phase however, these can be difficult and time-consuming to negotiate to the point that they can inhibit or delay sharing significantly (Tenopir, 2011)." Researchers need guidelines specific to country and funding agency on how to prepare ethics applications and anonymize data in order to share freely as much as possible (Poline et al, 2012).

Regulatory Laws and Policies – Recent U.S. National Initiatives and Laws Promoting Data Sharing

The conversation around data sharing within the scientific research community has elevated to the national level. In February 2013, the Office of Science Technology and Policy (OSTP) within the Executive Office of the President sent a memo to all federal agencies calling for increased access to results of federally funded scientific research (Holdren, 2013). Subsequently, former President Obama's *All of Us* Research Program (formerly known as the Precision Medicine Initiative (PMI) Cohort Program), and former Vice President Biden's *Beau Biden Cancer Moonshot* Initiative were launched in January 2015 and January 2016 respectively. Both initiatives call for increased and enhanced data sharing among researchers, research participants and health care providers to help generate new and innovative ideas, and improve treatment and prevention strategies.

On December 13, 2016, Congress passed the twenty-first century Cures Act that will increase funding for biomedical research, with most of the money earmarked for NIH to promote the Cancer Moonshot and *All of Us* Research Program. This bill also gives the NIH Director the authority to mandate data sharing in all grant applications.

Following the launch of the *Beau Biden Cancer Moonshot* Initiative in January 2016 a Cancer Moonshot Task Force including the presidentially appointed National Cancer Advisory Board (NCAB) was formed to address the goals of the initiative which primarily is to achieve rapid progress in cancer research over five years. This entails developing strategies to improve prevention, early diagnosis and detection, and treatment of cancer. The NCI was charged with implementing this initiative in collaboration with "cancer researchers, oncologists, patient

advocates, and representatives from the private sector and government agencies.[16]" The Blue-Ribbon Panel (BRP) was formed, and consists of seven working groups who would identify innovative scientific opportunities to move the initiative forward. The BRP came up with ten recommendations of compelling and timely research opportunities for the research community and one of them is to "create a national ecosystem for sharing and analyzing cancer data so that researchers, clinicians and patients will be able to contribute data, which will facilitate efficient data analysis"[17].

Regulatory Laws and Policies - Funding Agencies (The National Institutes of Health)

The NIH has been on the fore front of data sharing in biomedical research and recognizes the value of making data widely available and accessible to the public through NIH-designated data repositories. NIH requires that all data generated from NIH funded research should be made "as freely and widely available as possible while safeguarding the privacy of participants and the confidentiality and proprietary nature of the data.[18]" This is a way to expedite research, facilitate cross-discipline collaboration among scientists and stimulate ideas for new discoveries to improve health outcomes.

The *Final NIH Statement on Sharing Research Data* was released in 2003 and is the landmark policy statement and guidance on data sharing. This data sharing policy requires a data sharing plan for "any grant $500,000 or more in total direct annual costs or an explanation of why sharing isn't possible.[18]" However, authors note that details about how, when and where to make data available "were so vague that researchers soon stopped paying attention … until someone got in trouble for not playing by the rules" (Nelson, 2009). The vagueness in the policy

[16] https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative/blue-ribbon-panel/blue-ribbon-panel-report-2016.pdf
[17] https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative/blue-ribbon-panel/blue-ribbon-panel-report-2016.pdf
[18] https://grants.nih.gov/grants/guide/notic-files/NOT-OD-03-032.html

is intentional on the part of NIH to allow for exceptions to sharing as it related to privacy concerns and issues with informed consent (Nelson, 2009). Consequently, language in this policy introduces challenges with mandating, enforcing and compliance with researchers; the policy requirements are negotiated with the NIH Program Officer. It is not fully effective in getting researchers to share their data in public or controlled-access data repositories.

The need for clear policies and guidelines is critical to developing shared expectations, in addition to being able to monitor and track compliance. Otherwise, the compliance among researchers is very low. NIH program staff enforce these policies through monitoring and tracking, however since the data sharing expectations in the NIH Genomic Data Sharing policy are more clear and explicit compared to other policies, it's observed to lead to a more consistent implementation of the policy.

Notably, NIH does not factor in data sharing as part of the merit score given to grant applications during peer review. However, there are recent efforts underway across the NIH that plan on changing this and to allow data sharing to be a factor during peer review. Just like the other sections of the grant application that are scored, the data sharing section will also be scored, which can impact overall score in the fundable range. This is important because it will influence the decision-making process for grant funding and incentivize researchers to share data given that grants with good merit scores tend to get funded, although the funding decisions vary and depend on a variety of factors unique to each institute.

There are sixty-five NIH-supported data sharing repositories that accept data generated from NIH-funded research or other studies, but there are some restrictions on data submissions from investigators involved in a specific research network.[19] Despite the availability of these data repositories, many studies show that researchers are reluctant to share their data. It will be

[19] https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

valuable to learn from the experiences of NIH-funded investigators and NIH staff on successes and challenges with using existing data repositories such as the NIH dbGaP and how NIH might improve data sharing practices. Lessons learned here can be applied to other research organizations and funding agencies.

Policies – Science Journals (Editors and Publishers)

Journal publishers are a key stakeholder to achieving the goal of increased data sharing, and many of them are requiring that authors share their data with other investigators, either by "depositing the data in a public repository or making it freely available upon request" (Savage et al, 2009) as a condition for publication. "However, many have not yet implemented data sharing policies and the requirements vary widely across journals" (Vasilevsky et al, 2017). The ICMJE was established to encourage the sharing of clinical trial data, supporting the NIH data sharing policy. Data sharing is considered an ethical obligation of all who conduct research on human subjects while protecting privacy of subjects or patients.

The findings from a study that looked at availability of research data in highly-cited journals showed that a "substantial proportion of original research papers published in high-impact journals are either not subject to any data availability policies or do not adhere to the data availability instruction in their respective journals" (Alawi A, 2011). 30% of original research papers reviewed "were not subject to any data availability policy" and "59% did not fully adhere to the data availability instructions of the journals they were published in" (Alawi A, 2011). "The other 143 papers that adhered to the data availability instructions did so by publicly depositing only the specific data type as required, making a statement of willingness to share, or sharing all the primary data. Overall, only 47 of the 500 papers (9%) deposited full primary raw

data online. None of the 149 (30%) papers not subject to data availability policies made their full primary data publicly available" (Alawi A, 2011).

This is an example of where journal editors can help foster, change and enforce their journal publication policy as it relates to sharing of research data prior to publication (Tuchman, 2016). One study confirmed that only few "biomedical journals require data sharing" and that "there's a significant association between higher impact factors and journals with a data sharing requirement" (Vasilevsky et al, 2017). The results showed that of the 318 biomedical journals analyzed, 11.9% indicated data sharing as a requirement for publication, 9.1% mentioned data sharing indirectly, and 14.8% addressed protein, proteomic, and / or genomic data sharing; and 31.8% did not mention data sharing in the journals.

While more than half of the journals that require data sharing addressed reproducibility, the authors found that specific guidance on practices that ensure data availability and data reuse were missing in most data sharing policies (Vasilevsky et al, 2017). This is true for NIH data sharing policies based on comments from the scientific community in response to a NIH Request for Information on *Strategies for NIH Data Management, Sharing and Citation*[20]. An exception to this is the NIH Genomic Data Sharing Policy which has specific guidance, procedures, systems and data repositories such as dbGaP, in place to facilitate data broad sharing of genomic data and reuse. This is particularly of importance to policy makers and funding agencies to understand how to best support the development and implementation of data sharing policies to enhance sharing and therefore the rapid progress of scientific research.

A similar study found that of the ten raw data sets requested from investigators "who had published in journals with explicit data sharing policies, only one investigator sent an original data set. The others refused to share or did not respond to the request" (Savage et al, 2009). This

[20] https://osp.od.nih.gov/wp-content/uploads/Public_Comments_Data_Managment_Sharing_Citation.pdf

implies that authors with original data do not necessarily make their data sets available to other investigators regardless of the journal policies. This impacts the advancement of scientific progress.

Literature shows that many published research findings can't be reproduced because they lack robustness and rigor in the methods and analysis, which leads to misleading results (Tsodikov, 2017). Journals can help with easing problems of reproducibility by publishing more studies that report negative results (Harris, 2017). In addition, the *Scientific Data* journal is a step in the right direction in allowing not just findings but data authors to be lead authors on publications. There is no guarantee that these will improve data sharing and recognition of the value of sharing data among the scientific community. It's also not clear what it will take to enforce such practices or essentially change the thinking or status quo as it relates to publication of findings.

Technological Infrastructure – Financial, Administrative and Technical Support

The technical challenges associated with data sharing is not just the technical aspect, for example inadequate technological resources and varying standards, but also the burden with data management and programming support which can get expensive (Poline et al, 2012). Researchers are dependent on funding agencies and academic institutions for funding for infrastructure and technical support for data management. The NIH data sharing policies encourage researchers to include the cost for data sharing in their grant application. This practice will vary across the scientific community because of the diverse ways researchers and their institutions can choose to interpret and implement this aspect of the policy. One major aspect underlying the efficiency and effectiveness with managing, storing, sharing and using data is the

availability of adequate and secure technological infrastructure in place. If this is not addressed, it poses a technical barrier to data sharing (Nelson 2009, Pisani et al, 2010, Enke et al, 2012).

There's the issue of sustainability and long-term maintenance of data repositories and funding to support not just the maintenance of the infrastructure (Nelson 2009; Poline et al, 2012; Enke et al, 2012) but also funding for data preparation, documentation, and management. Considerable expertise, effort, restructuring and proprietary software are important to consider (Borgman, 2012). Interoperability of a variety of databases and repositories is also critical in data management. Databases that are not interoperable hinder data sharing because data are housed in different systems and it makes it difficult to access data and do analyses on the data. Developing structures and systems that can facilitate data sharing, standardized processes and systems that can talk to each other is invaluable.

Other technical issues that affect infrastructure for data management and therefore hinder data sharing in publicly accessible databases have to do with 1) formatting data in a way that can be easily accessible to the scientific community (Pisani et al, 2010), and 2) formatting data in a way that can be used by a secondary user without any errors (Ascoli et al, 2017). While this is perceived as necessary, it is a big investment of time and effort on the researcher's part to format and prepare data so it can be deposited and easily accessed and interpreted by anyone. The formatting of the data and lack of standardized data formats make it challenging for data to be shared in a meaningful way.

This is less critical in the field of genomics. For example, having common data elements, data structure, and the fact that "sequencing one genome is very similar to sequencing another" (Nelson, 2009), make it very easy for sharing of genomic data. For useful genomic analyses, it's important to have "good metadata that describe sample collection procedures, clinical definition

of cases and demographic data (Kaye et al, 2009)." This should apply to analyses in other fields or disciplines, however, in fields such as environmental science, "the choice of standards is far less obvious" (Nelson, 2009).

For effective data reuse, sharing, and interpretation, it is important to understand the data type, size, requirements for storage, standards for data formats, and units of measurements, all of which impact study quality and validity (Gardner et al, 2003). Although having standards and detailed metadata make it easier to share data, they are difficult to establish, requiring a lot of effort. Without good descriptive data standards, protocols and analytic variables, secondary data users are faced with the challenges of understanding and accurately interpreting the primary data within its original context (Gardner et al, 2003).

In addition, the security of the server is important because of the sensitivity and confidentiality of the data from human research participants stored in the database. The nature of biomedical research in the 21st century is one that involves multi-site collaborations among scientists within and outside their research networks or academic institutions. Such collaborations require the sharing of data transmitted through high-speed computer networks, critical for submission and analysis (Harris, 2017). The author further states that sharing data in a secure manner, although critical for collaborative research, is a "formidable challenge". While there are many factors that affect the willingness, motivation and ability of researchers to share data, "there is a special role for technology to pave the way – both in removing barriers, as well as in delivering rewards and benefits" (Koers, 2016). This dissertation research explored these issues and determined opportunities for improving technological infrastructure to support data sharing.

B. Conceptual Framework

The socio-ecological framework and systems theory was used to inform the conceptual framework / theory of change (Figure 4) for this dissertation because it provides the opportunity for a systems perspective on the different types of factors influencing the sharing of research data in public or controlled-access data repositories.

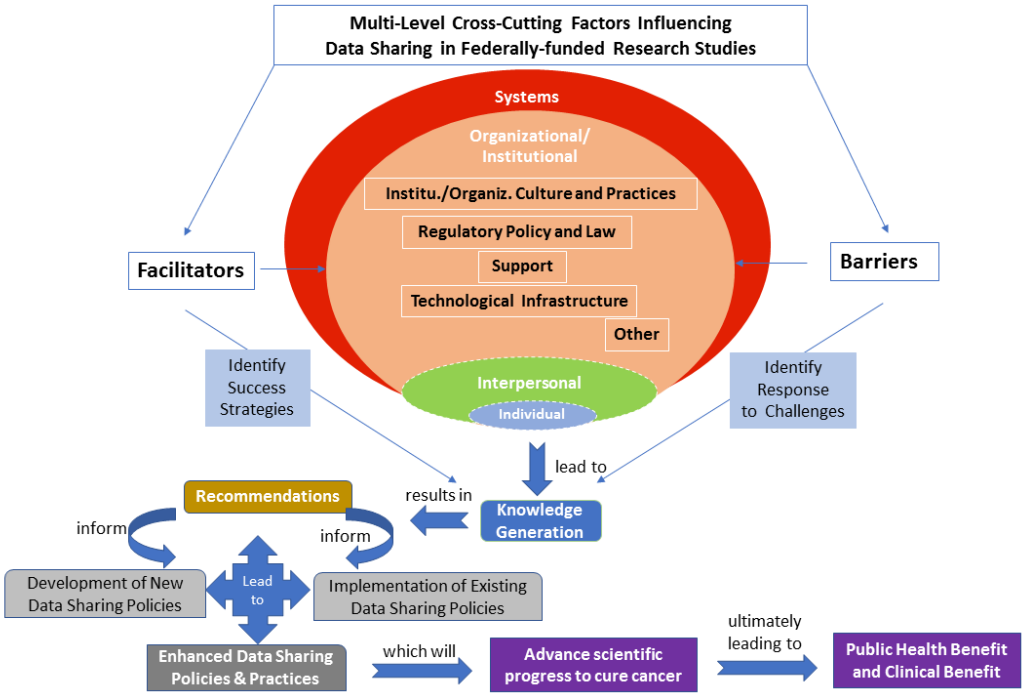

Figure 4: Initial Conceptual Framework: Factors influencing data sharing in federally-funded research

The ecological model is useful in framing the cross-cutting factors that influence data sharing among researchers at the systems level, the organizational / institutional level, the interpersonal level, and the individual level. In addition, it shows the interrelationships among

the multi-level factors which can act as facilitators or barriers to data sharing in biomedical research.

The focus of this dissertation research is on the organizational / institutional level factors which are outlined in white boxes within the peach circle of the framework. A review of the literature was conducted (peer reviewed literature and grey literature) as well as environmental scans as part of the coursework of this dissertation to help inform the development of the conceptual framework. Through conducting the environmental scan, different challenges experienced by investigators, program staff and leadership around data sharing emerged; some of which were echoed in the literature.

The environmental scan comprised of brief and informal meetings with NIH colleagues and individuals in leadership, and those who oversee research grants and are in direct communication with investigators. Their perceptions of data sharing and challenges with data sharing in NIH-funded research along with their ideas for facilitating sharing were discussed. Through participation in the EGRP-led Data Sharing Working Group, information about various data sharing activities and efforts across the institute to promote data sharing was observed. This also included participation in webinars on data sharing and open data science. The presentations and discussions contributed to knowledge of data sharing, the value of sharing and the implications of not sharing research data.

Another environmental scan conducted was through programmatic involvement in cohort and consortia-based research supported by EGRP. This provided the opportunity to brainstorm with colleagues on some of the challenges investigators are reporting and ideas for addressing those challenges, which are consistent with the literature. Learning about strategies to enhance the implementation of data sharing polices at different NIH-wide data sharing events provided

some insight on different models and strategies that could be adopted or modified for epidemiological cohort studies funded by EGRP. All these informal discussions and exploration of this dissertation research topic provided great insights on what others view as the problems and challenges with data sharing, and what their thoughts are on how to make data sharing easy.

Based on review of the literature and environmental scans, as well as systematic reflection on the themes or discoveries that emerged, the high-level factors (a priori) that function as barriers and / or facilitators to data sharing at the organizational / institutional level are as follows: 1) institutional / organizational culture and practices, 2) regulatory policy and law, 3) support and 4) technological infrastructure. These factors have already been described in greater detail in the first two chapters. There may be other factors, not part of this initial a priori list that could emerge during the data collection phase.

This conceptual framework shows that by identifying strategies that have led to successful sharing of data through learning from existing models / cases at the NIH with genomic data sharing as the exemplar, as well as identifying how the investigators respond to challenges with sharing data in public or controlled-access data repositories, new knowledge will be generated. For example, exploring the experiences with implementation of the NIH GDS policy at the NCI, depositing and accessing data in dbGaP can provide insights into factors that facilitate or hinder data sharing, and the opportunity to apply that knowledge with the intent to enhance data sharing practices among the research community in general.

Analysis of the knowledge generated will result in a set of recommendations for the Institute. These recommendations will inform the NIH Office of the Director, NCI and other Institutes at NIH on things to consider during the development of new data sharing policies, as well as the implementation of existing data sharing policies in the research studies funded by

NIH. The recommendations, if implemented, will lead to increased and enhanced data sharing policies and practices in general. The impact of enhanced sharing of cancer research data among NCI-funded researchers is the acceleration of the pace of scientific discovery and advancement in scientific progress towards preventing and finding a cure to cancer, and ultimately leading to improved health outcomes in terms of public health and clinical benefits.

III. STUDY DESIGN AND METHODS

A. Research Design

A case study method is appropriate for gaining an in-depth understanding of a real-life social phenomenon and the contextual conditions in relation to the phenomenon (Yin, 2009). For this dissertation research, a case study design was used to address the study questions related to the phenomenon and context around data sharing in biomedical research. The use of "how" and "why" in formulating research questions are one of the main criteria for a case study design (Yin, 2009). Therefore, the case study design was well suited for this dissertation research because the research questions addressed "how" organizational and institutional level factors are functioning as facilitators or barriers in data sharing among researchers.

Data sharing in biomedical research is a contemporary issue, one of the criteria for case study methods as defined by Yin (2009). It is an issue that has gained prominence on the national stage following the launch of former President Obama's *All of Us* Research Program in January 2015, former Vice President Biden's *Beau Biden Cancer Moonshot Initiative* in January 2016, and the 21st Century Cures Act signed into law in December 2016. These initiatives are in various stages of implementation by U.S. federal agencies to enhance the sharing of research data generated from federally-funded studies to advance scientific progress and benefit the public through novel scientific discoveries. Notably, journal editors and publishers have also instituted policies requiring researchers to deposit data in a public or controlled-access repository before their articles can be published.

To enhance data sharing policies and practices, it is of great benefit to first explore and learn about models of existing data sharing practices that have proven successful and try to understand what the facilitators and barriers to data sharing are. To this effect, this dissertation research examined existing NIH data sharing policies, specifically, the 2015 NIH Genomic Data

Sharing policy as an example of a successful model and used the knowledge gained from understanding the factors that influence the sharing of genomic data as well as from unresolved issues with genomic data sharing to inform efforts to facilitate and enhance data sharing practices in the field of epidemiology.

Therefore, understanding the organizational / institutional level factors that influence data sharing practices among researchers and how these factors facilitate or hinder data sharing among researchers was best explored through a case study design method. This required the use of multiple data sources for evidence, a unique strength of the case study design (Yin, 2009), such as: 1) review of documents and literature pertinent to policies and efforts around data sharing, and 2) interviews of key stakeholders who are influencing or are influenced by data sharing policies. This dissertation research generated a set of potential recommendations for opportunities to enhance sharing of federally funded research data in public or controlled-access databases.

Given the goal of this dissertation and the focus of the research questions on organizational / institutional level factors, the primary unit of analysis is the organization or institution, specifically, the Epidemiology and Genomics Research Program (EGRP), within DCCPS at the NCI. The EGRP has a large portfolio of over three hundred epidemiology and genomics cancer research grants. EGRP has been a key player in the implementation of the genomic data sharing policy at the NCI level. The embedded units of analysis include EGRP-funded investigators of genomic and epidemiology studies and NIH staff involved with data sharing policy development and implementation processes.

i. Case Selection: Rationale for Genomic Data Sharing as an Exemplar Case

This dissertation research is a case study of genomic data sharing as an exemplar for epidemiology data sharing policies and practices. As described in the previous chapters, the field of genomics/genetics has always been at the forefront of promoting open data access and data sharing to advance research progress in health and medicine. There is a culture that elevates the expectation to share genomic data and this dates back to the 1996 summit in Bermuda where leaders in the scientific community agreed to share genetic data generated from the Human Genome Project.

There has always been support from NIH leadership on data sharing and following the advancement in genomics and related technologies, the NIH has directed more resources towards the support of genomic data sharing, ensuring full protection of human data. There are data sharing policies that have been established by the NIH, starting with the 2003 general NIH Data Sharing policy which only applies to grants over $500,000 in total direct cost. This policy expects that final research data generated from NIH-funded studies should be made available while taking caution to protect the privacy of study participants.

NIH supports GWAS studies to explore and identify genetic factors associated with diseases and making the data available can influence health and strategies for prevention and treatment. Therefore, in 2007, NIH published the GWAS policy requiring researchers conducting GWAS studies to deposit both genotype and phenotype datasets in dbGaP, a controlled-access data repository supported by NIH. Several years later, in 2015, this policy was expanded to the NIH GDS policy, and the key difference with GWAS policy and the 2003 NIH Data Sharing policy is that the GDS policy specifically mandates broad and responsible sharing of large scale human and non-human genomic research data generated from NIH-funded studies

regardless of funding levels. The NIH Intramural Research Program Human Data Sharing (HDS) Policy also published in 2015 was developed to foster data sharing of human data generated from the NIH-funded intramural research; complementary to other NIH data sharing policies.

Much of what we know to be critical for success in data sharing exists in the field of genomics and the NIH GDS policy is reflective of these things. Unlike epidemiological data, there are resources, processes and systems created by NIH to facilitate the submission of and access to genomic data. These include: policies articulating clear expectations, including guidelines and timelines for submission in a repository, templates, specific measures and terms and conditions for secondary use of research data.

ii. Sampling Selection: Rationale for Sampling

This dissertation research employed purposeful sampling as a strategy for selecting participants who provided relevant data to answer the study questions. This was a deliberate and targeted approach that ensured that selection of the right type of information, individuals, right setting and context helped with meeting the goals of the study, in a way that "can't be gotten as well from other choices (Maxwell, 2012)." Purposeful selection helps in many ways to strengthen the validity of the study. Selecting participants from federal agencies and academic research institutions was done to help achieve representativeness of the population of researchers and administrators impacted by data sharing requirements of federally-funded research.

In addition, the findings from this case study helped provide comparisons and elucidate reasons for any differences or new discoveries. The criteria for selecting study participants for this study were grouped into two categories within the context of NIH as the organization and federal agency of interest, and its data sharing policies and practices. The study participants for

this study are both NIH-funded investigators and NIH staff.

NIH-funded Investigators

To keep this study focused and feasible for this dissertation research, the selection process began by identifying a single institute at the NIH and investigators or researchers with grant awards from that institute. NCI, the largest of the 27 institutes and centers at the NIH in size and funding, and a major player in the implementation of the NIH GDS policy was selected. The NCI is comprised of many divisions, programs, offices and centers and incorporating the entire portfolio of NCI for this dissertation was a nearly impossible task to achieve. Therefore, it was necessary to narrow down the sampling of investigators to a single program - EGRP, one of the four programs within DCCPS. EGRP was also the primary study site for this dissertation research.

EGRP is the largest funder of epidemiology and genomic research in the world; with a portfolio of more than three hundred grants. With extensive professional experience and work on EGRP research portfolio on cancer epidemiology and genomics, selecting EGRP as the primary NCI unit for this research was a reasonable choice. For these reasons, EGRP-funded investigators were determined to be an appropriate group of investigators to select and were relevant for this study. While it was equally an impossible task to include all EGRP-funded researchers in this study given the size of its portfolio, a sampling strategy was developed that allowed selection of investigators who helped address this dissertation research questions. In general, the EGRP-funded investigators selected as participants in this study are individuals who currently conduct genomic and / or epidemiologic studies and have some experience with and some knowledge of data sharing, including the deposit or access of data in controlled-access or publicly accessible databases or repositories.

Specifically, there were three groups of investigators derived from a few ongoing major scientific activities, priorities and initiatives in DCCPS and EGRP that align with the overall mission of the NCI / NIH. These groups were as follows:

1) New investigators invited to the *2017 DCCPS-sponsored New Grantee Workshop*

According to the NIH, new investigators are classified as: Early Stage Investigators (ESI), described as investigators who received their terminal research degree within the past 10 years and who have not received their first substantial independent and competing NIH Research Project (R01) grant.[21] NIH is committed to supporting Early Stage Investigators, and similarly, DCCPS values the contribution of new investigators because they tend to bring new perspectives and ideas to help advance research in cancer control and population sciences.

The *2017 DCCPS-sponsored New Grantee workshop* was held in September 18-19, 2017, at the NCI campus; the second time in two years that the division had convened new investigators who received their first NIH R01 grant from the DCCPS within the prior fiscal year (FY). The purpose of the workshop was to provide the new investigators with networking and collaborative opportunities to share and discuss ideas, meet with program experts throughout the institute and other colleagues from the scientific community, provide them with key information such as NIH-wide and NCI specific funding opportunities, scientific priorities, tools and resources that will help them become successful grantees, and ultimately help them advance their careers in cancer control research.[22] ”

New investigators invited to this workshop were considered a good group that would provide new insights and fresh perspectives on data sharing policies, practices and opportunities to advance science. A participant list from the workshop organizers was obtained, with about

[21] https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-101.html
[22] https://cancercontrol.cancer.gov/new_grantees/workshop-2017.html

forty-five investigators funded across the division who met the criteria for a new investigator and had received funding either in FY2016 or FY2017. Of the forty-five investigators, twelve of them were funded by EGRP (five attended the workshop) with the rest of the participants funded by the Behavioral Research Program (BRP), Surveillance Research Program (SRP) and Health Delivery Research Program (HDRP) in the Division of Cancer Control and Population Sciences at NCI. The new investigators of interest were those funded by EGRP, keeping in line with the overall research plan and research questions.

A one-page information sheet on this dissertation research was shared with the attendees, during a small informal meet-the-expert lunch session organized by EGRP program directors. The study plans were discussed briefly with the new investigators, and the intent to contact them in the future requesting participation in the study, pending IRB approval, was made known at that point. The recruitment strategy entailed emailing the EGRP new investigators who met the criteria, i.e. both those who attended the workshop, and those who were unable to attend the workshop.

Upon IRB approval, the investigators were sent an email to participate in the study. Investigators were given one week to respond. Those that didn't respond within this time frame were sent an email reminder and given an additional week to respond. The plan was to follow up with a phone call as a second reminder if no response was received at the end of the two week period. These investigators, though new NIH grantees, have some knowledge and limited experience with data sharing in dbGaP, and were a good group to include in this study.

2) <u>Experienced Investigators of a Large EGRP-funded Initiative - Genetic Associations and Mechanisms in Oncology (GAME-ON)</u>

To capture the perspective of genomic data sharing, a list of active scientific research initiatives in the EGRP portfolio was reviewed. Those that focused on cancer genomics and epidemiology with investigators considered to be more experienced were deemed of high interest. These are investigators who have over ten years of experience as NIH grantees and have successfully competed for multiple large highly collaborative NIH Research Project (R01) grants.

The Genetic Associations and Mechanisms in Oncology (GAME-ON) initiative was one of the initiatives selected as a source for recruiting study participants. The GAME-ON initiative is a "network of consortia for post-genome wide association research," funded by NCI in 2010 in response to an NIH issued Request for Application (RFA-CA-09-002) titled *Transdisciplinary Cancer Genomics Research: Post-Genome Wide Association (Post-GWA) initiative (U19)* [23]. It is one of EGRP's older initiatives with senior level and experienced researchers conducting large-scale genetic epidemiology studies. The goals of this initiative are to foster a transdisciplinary and collaborative approach in GWAS studies and to "provide a rigorous knowledge base that would enable clinical translation and public health dissemination of cancer GWAS findings." [24] The GAME-ON investigators are familiar with the requirements of the NIH data sharing policies given the nature of their research which includes both genomic and epidemiological data. Therefore, they were considered to be a good group to include in the study.

There are five cooperative agreement (U19) grants that are funded through the GAME-ON initiative and each grant has multiple investigators, including the Principal Investigator and co-Investigators. The recruitment strategy included emailing all the Principal Investigators and

[23] https://grants.nih.gov/grants/guide/rfa-files/RFA-CA-09-002.html
[24] https://epi.grants.cancer.gov/gameon/#background

co-Investigators who met the criteria, i.e. a maximum of 25 investigators, with information about this dissertation research study and request their voluntary participation. The investigators were given one week to respond. Those that didn't respond within this time frame were sent an email reminder and given an additional week to respond. The plan was to follow up with a phone call as a second reminder, if no response was received at the end of the two week period. These investigators are established and have been funded for the past eight years and have experience with sharing their data in dbGaP[25]. Given that each grant has multiple investigators, the recruitment goal was to reach at least five investigators who are focused on genomic research.

3) Investigators of a Large EGRP-funded Initiative - NCI Core Infrastructure and Methodological Research for Cancer Epidemiology Cohorts

Understanding data sharing practices as it pertains to epidemiological data in NIH-funded studies, which aligns with the goals of this research study can be achieved through the perspectives of investigators of epidemiology cohort studies. Cohort studies are one of the fundamental designs of epidemiology studies and the data collected from cohort studies have helped researchers to better understand the complex etiology of cancer, and have provided fundamental insights into key environmental, lifestyle, clinical and genetic determinants of this disease and its outcomes.[26]

EGRP provides support to twenty-nine cancer epidemiology cohorts which represents a significant percentage of the entire program's grant-based research budget. The EGRP-funded initiative titled, *Cancer Epidemiology Cohort Infrastructure: Core Infrastructure and Methodological Research for Cancer Epidemiology Cohorts (PAR-17-233)* provides infrastructure support to the core functions of existing cancer epidemiology cohorts and

[25] https://www.ncbi.nlm.nih.gov/gap
[26] https://epi.grants.cancer.gov/cohorts.html

methodological research. It does not support hypothesis-driven research projects because investigators have obtained other support for those through other mechanisms such as the investigator-initiated research projects (R01). There are twenty-four cohorts currently funded through this initiative.

The investigators of these cohorts have extensive experience with consortia research that involves the pooling of data from multiple large cohort studies to address research questions that can't be otherwise addressed in a single study or by a single cohort or entity. While this gets to the core of why sharing data is important and NIH's mission in advancing scientific discovery through collaboration, there are challenges with sharing data that are inherent in large prospective cohort studies that are worth exploring. Including these investigators in this study provided some insights on data sharing practices and challenges in large observational population studies such as cohort studies.

The recruitment strategy included first emailing all the Principal Investigators of the cohorts funded under this initiative, with information about this dissertation research study and requesting their voluntary participation. Investigators from the twenty-four cohorts whose research focus is on epidemiology, genomics or both were included. This would account for a maximum of 120 investigators. The investigators were given one week to respond. Those that didn't respond within this time frame were sent an email reminder and given an additional week to respond. The plan was to follow up with a phone call as a second reminder if no response was received at the end of the two week period.

The plan was to interview a minimum of 5 investigators on a first come first serve basis and also to sample until a point of data saturation was reached, i.e. when no new data, themes and coding from the interviews became evident. These investigators have experience with data

sharing practices in large pooling research projects across cohorts such as the NCI Cohort Consortium, which was created by NCI to "address the need for large-scale collaborations to pool the large quantity of data and biospecimens necessary to conduct a wide range of cancer studies."[27] Selection of investigators to include in this study is discussed in the analysis plan section later on in the chapter.

NIH Staff

NIH staff play a critical role in helping the institute achieve its goals and overall mission to improve health, lengthen life and reduce disease and illnesses in populations. They oversee the research and programmatic investments of the institute and are there to serve as good stewards of tax payer dollars. In the context of data sharing in NIH-funded research, program staff are responsible for developing policies, implementing the policies including troubleshooting and identifying ways to improve policies and processes, and help with facilitating the funding of high quality research studies that are reproducible. Therefore, NIH staff who represent diverse and complementary perspectives on the factors that facilitate or hinder the sharing of research data generated from NIH-funded studies, as well as those that could provide perspectives on opportunities for improving data sharing among researchers were included in this study.

The NIH staff included in this study are individuals that would help address the study research questions. These individuals are knowledgeable about data sharing policies, practices and the broader implications in general. They have some sense of what is feasible and implementable, i.e. what it takes to implement effective data sharing policies in NIH-funded research especially given the ethical, legal and sociopolitical issues in research involving human subjects. The staff have some understanding of the history and evolution of the NIH data sharing

[27] https://epi.grants.cancer.gov/Consortia/cohort.html

policies and could provide insights into the current data sharing practices, in addition to being open-minded in terms of ideas and opportunities for enhancing data sharing among NIH-funded researchers. Finally, the staff who are involved in the technical and informatics aspects of data sharing were included in this study because of their knowledge, expertise and experience with facilitating the submission of data and access to data in public or controlled-access data repositories such as dbGaP.

Specifically, the NIH staff included in this study fall into four groups based on the level of relevance and alignment with the research questions, and alignment with the overall mission of NCI and NIH. These groups are as follows:

1) Staff who hold leaderships positions at the NIH Office of the Director and other divisions at NCI. These individuals have research expertise in both genomic and epidemiology and can provide insights on the leadership implications of data sharing in research

2) Staff who are involved in the development of the NIH data sharing policies, which are coordinated within the NIH Office of the Director. These individuals coordinate the comments and responses from the scientific community on draft data sharing policies intended to inform the development of final policies.

3) Program staff at NCI involved with the day to day administration and implementation of the NIH GDS policy. These individuals serve as Genomic Program Administrators (GPA) and facilitate the registration of studies and submission processes for NIH designated controlled-access data repositories. They also serve as members and chairs of the NIH Data Access Committee (DAC) which reviews all requests from the scientific community for access to genomic data in dbGaP or other NIH-designated controlled-access data repositories. These

individuals interface directly with the investigators and provide some support to facilitate successful sharing in the repositories.

4) NIH staff who oversee the technical aspects of data sharing. This includes the creation and management of structural databases and informatics of the data repository infrastructure. These individuals have the knowledge and expertise on the types of challenges, from their perspective, that impacts the sharing of data in public or controlled-access databases.

B. Data Sources, Data Collection and Data Management

Data to be used in addressing the research questions in this dissertation research were collected through multiple data sources. These are namely through in-depth semi-structured interviews and document reviews. The measurement table in Appendix C maps these data collection sources and tools to the research questions for this dissertation.

i. Data Source and Data Collection

In-depth Semi-Structured Interviews

An interview guide was developed with questions and probes pertinent to the research questions, and used to interview a group of investigators funded by EGRP and NIH staff who represent diverse and complementary perspectives on factors that facilitate or hinder data sharing. The interview guide was modified as appropriate for each of the separate groups of individuals, with one for researchers and another for NIH staff (Appendix D). Interviews with the investigators were conducted by phone because they were geographically dispersed across the U.S., while interviews with NIH staff was conducted mostly by phone with a few conducted in-person.

As previously described, EGRP-funded investigators were key to addressing this dissertation research questions. Some of them are conducting research that is focused primarily

in epidemiology and some in both genomics/genetic epidemiology. New or early stage investigators while they may not have had experience with depositing or accessing data in NIH data repositories, could provide future perspectives on data sharing and help shape and change the culture. The experienced investigators had various levels of experiences with the data sharing policies. In some cases, compliance and enforcement had been a challenge, and this is where NIH program staff overseeing these grants could provide valuable insights. It was useful to compare both groups, as well as understand the facilitators or barriers to sharing epidemiological data generated from EGRP-funded prospective cohort studies.

The interview guides were piloted with two individuals – one NIH staff and one NIH-funded investigator who were not associated with this study but had similar background and experiences as the study participants. This helped provide additional clarity, flow, and focus around the interview questions, which were then further refined prior to data collection.

Document Reviews

Reports of Public Comments in Response to NIH Requests for Information

In addition to conducting in-depth interviews, a set of documents were reviewed as part of the data sources. Two of the documents were public comments in response to two NIH Requests for Information (RFI) on data sharing, relevant for addressing the research questions. NIH uses the RFI mechanism to solicit open feedback and input from the scientific community at large as it relates to new policy development at the institute or scientific priority areas of focus that the institute should be exploring or supporting.

The format of the RFI is typically a brief description of the purpose followed by specific questions that the institute would want the general public, including the research community to

address within a specific time frame. The responses are submitted directly to the coordinating office where all the verbatim responses are collated, compiled into one document and posted on the public website. The information includes the name of the responder, their organization or institution and their responses. This qualitative survey data is then analyzed by the lead office coordinating the RFI process (i.e. NIH Office of the Director) and a high-level summary report is produced for the NIH leadership to use in making decisions or determining next steps.

The two RFI documents reviewed for this dissertation were relevant to the research questions and illuminated some of the issues in NIH policy development on data sharing.

1) Archived public comments from the scientific community in response to a Request for Information (RFI) on *Strategies for NIH Data Management, Sharing and Citation* (NOT-OD-17-015)[28] published in the NIH Guide for Grants and Contracts in 2017. The purpose of this RFI was to hear from the scientific community on their thoughts for developing effective data sharing strategies, what the barriers and burdens associated with these barriers are to help NIH implement the 2015 NIH Plan[29] and policy on the management and sharing of digital-scientific data generated from NIH-funded studies. The RFI was released to the scientific community for comments November 14, 2016 – January 19, 2017 and at the end of this period there were ninety-five individual responses received[30]. The information NIH requested in the RFI are as follows:

- Section I. Data Sharing Strategy Development
  - The highest-priority types of data to be shared and value in sharing such data

[28] https://archives.nih.gov/asites/grants/04-28-2017/grants/guide/notice-files/NOT-OD-17-015.html
[29] https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf
[30] https://osp.od.nih.gov/scientific-sharing/nih-data-management-and-sharing-activities-related-to-public-access-and-open-science/

        - The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications
        - Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers
        - Any other relevant issues respondents recognize as important for NIH to consider
    - Section II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports (RPPR) and Grant Applications
        - The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing
            - Use of a persistent Unique Identifier within the data/software citation that resolves to the data/software resources, such as a Digital Object Identifier (DOI)
            - Inclusion of a link to the data/software resource with the citation in the report
            - Identification of the authors of the Data/Software products
            - Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately

- Consideration of ambiguously identifying and citing the digital repository where the data/software resources is stored and can be found and accessed

- Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications
- Any other relevant issues respondents recognize as important for NIH to consider

2) Archived public comments from the scientific community in response to a Request for Information on *Processes for dbGaP Data Submission, Access and Management* (NOT-OD-17-044)[31]. The purpose was for the NIH to receive input from the public on how to enhance data submission and access processes for the NIH National Center for Biotechnology Information (NCBI) database of Genotypes and Phenotypes (dbGaP), and how to manage the data in this centralized controlled-access database or repository. The RFI was released to the scientific community for comments February 21, 2017 – April 7, 2017 and at the end of this period, there were forty-seven individual responses received.[32] The information NIH requested in the RFI are as follows:

- dbGaP Study Registration and Data Submission
- dbGap Data Access Request and Review
- Policies for the Management and Use of dbGaP Data, including alternate controlled-access models; benefits and risks associated with the availability of genomic study

[31] https://grants.nih.gov/grants/guide/notice-files/NOT-od-17-044.html
[32] https://osp.od.nih.gov/scientific-sharing/genomic-data-sharing/

summary statistics; and benefits and risks of such reference use of dbGap for research participants, patients, and the scientific community

- General comments on any other topics with regards to dbGaP data for research participants, patients, and the scientific community

Reports from EGRP-funded Study on Assessing the Landscape of Data Sharing in Cancer Epidemiology Cohorts

There are three other internal reports that are relevant to this dissertation study that were reviewed as part of this study. These were developed as part of a qualitative study on the landscape of data sharing in EGRP-funded cohorts, conducted through a contract with the Science and Technology Policy Institute (STPI) from 2015 - 2016. The goals of this pilot study while relevant to the study goals in the context of data sharing in cancer epidemiology studies, there are major differences. This dissertation study explored implementation of GDS policy and lessons learned from it to inform what could be done with epidemiological data sharing. The methods and approach have been described in prior sections, but the study participants will include NIH staff and EGRP-funded investigators, including investigators of cancer epidemiology cohort studies.

This EGRP/STPI study was done in three phases with the first two phases focused on the perspectives of the cohort investigators' as the primary data generators, and the last phase focused on the perspectives of data requestors. The goals of the first phase were to gain a better understanding about data sharing practices within cancer epidemiology cohorts funded by EGRP, especially as it relates to epidemiological data and how NCI might support the sharing of epidemiological data. This was a pilot study of two selected EGRP cancer epidemiology cohorts.

The second phase included an extension of the pilot study in phase one, involving nine select cancer epidemiology cohorts. This phase included an assessment of data sharing practices from the nine EGRP-funded cohorts. The third phase of the study focused on examining the data sharing needs and perspectives of researchers who request epidemiological data from NCI-funded cohort studies. The contractor / research team had phone interviews with cohort investigators on their perspectives and experiences with requesting, receiving, and publishing shared data, and the creation and implementation of centralized epidemiology data repository. The findings from this pilot study was used to further compare and substantiate findings from this dissertation research study.

iii. Data Management

Data collected from the document reviews described above were organized in a matrix / table created in MS® Excel and saved on a secure computer. The data were linked to the study research questions for ease of coding and analyses. The interview data were organized by questions in the interview guide, which were tailored to the separate groups of study participants. The interview transcripts were uploaded into NVivo 12 plus software for analysis. The interviews were recorded with a tape recorder and digitally saved on a secure computer drive. The interviews were transcribed and organized by unique codes / identifiers, and saved on a computer. The tape recorder and other data collection materials and tools such as hard copies of transcripts and notes were locked in a secure cabinet file.

C. Analysis Plan

The goal of this study is to see what can be learned from experiences with sharing of genomic data as a case study and knowledge from secondary analyses of prior qualitative

surveys done at NIH. Then use that data to develop a set of recommendations that can inform data sharing in the field of epidemiology. The interviews and document reviews were part of the first phase of getting at the experiences and comparison of the findings. The RFI comments and the reports from the EGRP/STPI qualitative study were used to cross check and compare findings and discussed as part of the results in chapter 4. This dissertation research will help elucidate opportunities for enhancing data sharing in the scientific community, and how to develop effective policies around sharing data in public or controlled-access databases.

i. Interviews

Interview Guide as Analytic Framework

Two interview guides (Appendix D) were constructed with questions that elicited the data needed to answer the research questions with one specific to the NIH-funded researchers or investigators, and one specific to NIH staff. Except for a few questions that were unique to each group all questions in the guides were similar. This was useful for comparisons and triangulation across data sources. In the interview guides, all participants were asked to describe their research area of focus i.e. whether genomics / genetics, epidemiology or both. This helped with tracking the type of information they shared i.e. whether the data or information was reflective of experiences with epidemiology data sharing or genomic data sharing or a mix of both.

A few questions in the interview guides asked about the participants' perspectives on NIH's definition of data sharing and the timelines for sharing. A few minutes prior to that point in the interview, study participants were emailed statements (or given statements during the in-person interviews) on NIH's definition of data sharing, as well as a description of NIH's data sharing policies and the timelines. They were given a couple of minutes to read and think about the statements prior to the questions being asked.

Each of the study participants were consented prior to the start of the interviews via email a week prior to the scheduled interview. In addition, a brief introduction of the study that included consent language was read to the participants as part of the interview guide, prior to asking specific interview questions. The introduction gave the participants the opportunity to ask any questions about the study, to be assured of the measures in place to protect the privacy of any data or information shared, and finally a disclosure that this study did not represent the views of the NIH or NCI or EGRP but rather that of the Study Investigator. The interviews were all recorded on a tape recorder and transcribed.

Sampling Strategy and Analysis Plan

The sampling strategy for selecting study participants entailed using convenience sampling within purposeful sampling to get at the breadth of expertise needed for this study. This included individuals in each of the previously described sample groups whose research focus is in epidemiology or genomics or a combination of both genomics / genetics and epidemiology. The area(s) of research focus were captured during the interviews. For the new investigators, there were 12 EGRP-funded new investigators who met the criteria and were invited to attend the DCCPS New Investigator workshop held Fall 2017. All these investigators, including those who did not attend the workshop, were included in the sampling plan. For the GAME-ON initiative, there were five principal investigators and multiple co-investigators for each of the five grants funded under this initiative. The plan was to invite up to twenty-five investigators from this initiative (about 5 investigators from each of the 5 grants) with the intent to select 5 to participate in the study. For the Cancer Epidemiology Cohort (CEC) Infrastructure initiative, there are twenty-four cohorts and the plan was to invite up to 120 investigators (24

cohorts with about 5 investigators per cohort), with the intent to select 5 to participate in this study.

Given the unexpected level of interest from the GAME-ON and CEC investigators, the selection of participants was on a first come first serve basis, ensuring balance in the investigator's research area of focus (convenience sampling). The investigators were also interviewed until a point of data saturation was reached. For the NIH staff, the plan was to interview up to 15 staff who represented the four areas of expertise of interest for this study as previously discussed. As a result, the maximum number of individuals that were potentially contacted was 172.

Recruitment Strategy and Analysis Plan

An informational email invitation about this dissertation research study was sent out at the same time to all potential participants who met the eligibility criteria, i.e. NIH staff and NIH-funded investigators. In the initial email, interest in participating in this study was gathered and based on their responses, respondents were provided a few options for dates and time frames for scheduling interviews. Recruitment and data collection was planned for a three-month period, between Spring and Summer 2018, with some lead time built in and up to two weeks allowed for them to respond with their availability.

Interviews did not follow a particular order given the invitations were sent out at the same time. However, interviews with the investigators were intentionally scheduled before interviews with NIH staff in anticipation of challenges with investigators schedules closer to the end of the academic semester and summer break. An email response was confirmed prior to scheduling interviews. As interviews were completed, they were transcribed, with the transcript reviewed and coded using NVivo 12 plus software. In addition, a comparison between the

perspectives of new investigators and experienced investigators around key issues related to data sharing practices was captured during data analysis. The experienced investigators have been working in the field for a lot longer, have institutional history and familiarity with institutional and research practices and culture, and can contribute to knowledge around compliance with genomic data sharing policy and the impact it has had on their willingness to share data. See Appendix C for measurement table with analysis plan.

Memos as Part of the Analysis Plan

Following the completion of each interview, analytic memos were written after each of the interviews to capture the following: 1) the context around topics of discussion throughout the interview, 2) nuances in the way participants responded to the questions such as tone of voice, 3) major or common a priori and emerging themes from the interviews, 4) the relationships / interactions among the themes, 5) the implications for this dissertation research and for NCI, 6) any "aha" moments, and 7) any questions that may emerge from the interviews, 8) summary of comments related to experiences in the field of genomic and epidemiology, 9) comments related to experiences with genomic data sharing and epidemiological data sharing, and 10) facilitators and barriers in genomic data sharing that may be relevant or applicable to epidemiological data sharing. These memos were helpful for the data analysis, especially given the different perspectives of individual participants in the study.

Member Check as Part of Analysis Plan

A member check was done as part of the analysis process and included a presentation of findings from the interviews with select study participants. The purpose of the member check was to confirm with the participants that the key themes and comments captured during the interviews were accurate. This occurred after the interviews were completed and during the data

analysis phase. All respondents were sent an email requesting participation in the member check and those selected were the middle responders (i.e. not the first ones or the last ones to respond). This was a way to balance out the type and level of feedback provided on the findings, across study participants. The main themes from the interviews were synthesized by interview questions and discussed with select respondents, encouraging feedback, confirmation, validation or clarification of themes.

ii. Document Reviews

The public comments in response to the two RFIs described earlier were reviewed for this case study. A high-level executive summary of the responses from the RFI on *Strategies for NIH Data Management, Sharing and Citation* was compiled by the NIH Office of the Director and that for the RFI on *Processes for dbGaP Data Submission, Access and Management* is underway. A subset of the individual responses were reviewed and coded manually. A random sample and subset of the RFI responses were selected and coded for themes relevant to the research constructs, using the same coding schematic as the interviews. Given the length of the documents reviewing all responses was not feasible for this dissertation. One document is about 400 pages long with verbatim comments from 95 respondents, and the second document is 90 pages long with verbatim comments from 43 respondents.

In the RFI public documents, the type of information about the respondents noted in the documents include: name of the respondent, name and type of their organization, roles and research area of interest. This information is publicly available on the website, making the list of respondents easily accessible. The RFI process used a convenience sampling to gather comments; they published the questions on the website to the larger scientific community, giving

them a specific time frame to provide comments on the questions. A systematic approach was used to select a sample of respondents from the RFI documents for this dissertation research , using sampling methods described by Creswell (2014). Random sampling was used for sample selection for this study because it will not be feasible to review and code responses from a total of 138 respondents. This method ensured that each of the respondents has an equal opportunity of being selected (Creswell, 2014). This method also ensured that the sample was representative and generalizable to population of researchers who responded to the RFI.

Before selecting a sample, a spreadsheet of the RFI respondents / researchers was created, and then stratified and sorted by name of the investigator, institution, type of their institution, research area of interest. As part of my criteria for selection, no more than two people from the same institution were selected. Following this, a name from the list was randomly chosen as the start for assigning numbers. Based on the number of respondents on the list on each of the RFI documents, 1 out of every 4th person on the lists (approximately 20%-25%) was included in the study. Selecting every 4th persons for the RFI on *Strategies for Management* resulted in about 23 out of 95 individual respondents' (approximately 24%) comments coded. For the RFI on dbGaP, selecting every 4th person resulted in about 10 out of 43 individual respondents' (approximately 23%) comments coded.

The findings from this analysis and from the review of the three EGRP internal reports were used to compare and cross-check with themes to help increase the internal validity of the study. In addition, the coded themes were used to triangulate findings from the interviews. All coding of the documents was done by the Study Investigator of this dissertation research. The same method of coding of the constructs and related factors was used for all documents and interviews.

Pattern Matching

Recurring themes from the data collected across the various sources were reviewed, synthesized and grouped into categories for the analysis. Pattern matching was used in this case study to compare empirical data from the data sources, i.e. the document reviews and interviews, to show patterns of coded data. It helped to highlight how the themes were related to the predicted / theoretical patterns and constructs in the conceptual framework as well as with what is in the measurement table. Patterns were identified by looking at the common themes in the coded data such as technological infrastructure, support and policies.

Descriptive Analytic Framework

The interview guides tailored to the separate groups in the case study were useful for analysis as described above. This included new and experienced NIH-funded investigators of large NIH-funded initiatives focused on genomic and epidemiological studies and their experiences; and NIH staff on their experiences and perceptions of challenges with sharing genomic data in publicly accessible or controlled-access databases. The interview guide was used as a descriptive analytic framework to help categorize themes from the different interviews and guide the coding and analysis. A within-case analysis approach on each of the interviews and document reviews prior to doing a cross-case analyses across the different data sources was employed.

Coding

All interviews were recorded, transcribed and coded. The recordings were transcribed professionally with all personal identifiers removed prior to sharing the transcript with a second coder. The recordings were saved to a secure computer and will be destroyed after all research has been completed. Coding for a priori and emergent topics of the interview data was done in

NVivo 12 plus software and the document reviews done manually and organized in a matrix / table by relevant research questions and associated constructs. A code book (Appendix E) was developed to provide a description of the different codes to go with the analysis. Nvivo 12 plus software was used to organize the data and to run queries and frequencies. This was helpful in mapping relationships between the different constructs and factors.

Ten percent of the interviews was coded with a second coder using the same codebook. The codes in the codebook were first reviewed and discussed with the second coder to ensure a shared understanding of the meaning of the codes and how to apply the codes to the data. After co-coding three transcripts manually, a comparison of the coding revealed very few discrepancies, which mostly had to do with how the terms "facilitators" and "opportunities" were applied to the data. These were discussed and consensus was reached on the decision for how to code the discrepant data. This process helped increase the validity of the research findings.

Triangulation of Data

According to Maxwell (2012), triangulation is the process of "using different methods as a check on one another, seeing if methods with different strengths and limitations all support a single conclusion." Triangulation of the data helped with exploring relationships for convergence or divergence at multiple levels, across methods or data collection sources, as well as across different types of study participants. Triangulation across methods or types of data collection sources, i.e. in-depth interviews and document reviews, included exploring relationship across the following constructs (institutional / organizational culture and practices, policy, resources and technological infrastructure) to help corroborate findings across the data and see how each of the factors occurred as facilitators or barriers or opportunities to enhance data sharing.

Specifically, triangulation was conducted by data sources in the following ways: 1) by the types of investigators based on their classification as new or experienced investigators; 2) by investigator based on their self-identified research area of focus as epidemiologists and genomicists/genetic epidemiologists; and 3) by investigators compared with NIH staff. These all included a comparison of the experiences of the study participants in terms of their perceptions of facilitators, barriers and opportunities related to the constructs. A synthesis of the key themes that emerged across all methods / data collection sources i.e. the in-depth interviews and document reviews was conducted, as well as the themes from the comparison and contrasting of different experiences from all investigators and NIH staff.

D. Validity Considerations

The following tests were helpful in addressing any limitations and threats to the validity of the study.

Construct validity test

To counter or address validity threats in this dissertation research and reduce any biases in a specific method, multiple sources of evidence and data sources were used (interviews, document reviews) during data collection. The conceptual framework of this research is very explicit and grounded in literature. The interview guides were constructed in relation and alignment with the conceptual framework and constructs identified in the literature. The interview guides were also pre-tested with a few individuals who are similar to the study participants, for review and validation of the instrument.

Internal validity test

The use of an interview guide as a data collection tool addresses internal validity in terms of the consistent approach that was used for data collection in this study. There was less tendency for bias in terms of the individuals interviewed because there was a clear rationale specified as well as an explicit eligibility criteria for who to interview. During the data analysis, coding for constructs based on the conceptual framework was kept the same for the RFI documents and the interviews. In addition, cross-checking themes from the analyses with themes from existing reports and summaries helped increase internal validity.

The data was triangulated across multiple data sources to help eliminate biases of any specific method used, or any inferences made, as well as elucidate any convergence or divergence of evidence. During the interviews, deliberate care was taken to ensure that the comments and statements made by the respondents were clear, minimizing ambiguity in their responses. To further increase internal validity, a member check with select respondents was conducted after all interviews were completed, as described earlier. This check provided an opportunity to validate the findings and increase accuracy in the data being analyzed. Preliminary results of the findings were also discussed with EGRP leadership staff initially, followed with a presentation to a larger group of EGRP staff, with feedback and clarification integrated into the study.

External validity test

Testing for external validity is dependent on the study sample, which is NIH/EGRP-funded investigators. EGRP as the unit of analysis is a typical organization in terms of funding agencies. The mix of genomic and non-genomic research in EGRP makes the findings from this case study sample transferable to other agencies or similar research groups outside the NIH. The

findings or results from this study could inform and be transferable to data sharing practices in epidemiology and other federal or non-federal research groups.

Reliability test

To minimize any study biases and errors, analytic memos and reflective journaling were developed and used during data collection to keep a clear, detailed step by step systematic documentation of the procedures and logical tracking of steps taken in the study.

An interview guide was developed in a way that structured the types of questions and data collected and analyzed in alignment with the study research questions. Systematic reflective memos were developed after each interview to document various nuances in the way the study participants responded to the questions, e.g. tone of voice, comments, etc.

For these memos, a template was created that would include a set of prompts that would help me address the "What?", "So what?", and "Now what?" aspects of the information collected. The template included the following questions or prompts: 1) what were the major points that emerged from the interview, 2) What are the implications for NIH, 3) What are any primary / major take-aways or 'aha' moments?, and 4) Were there any emerging questions? To address researcher bias encountered during data collection from the study participants, comments made by respondents during the interviews related to researcher affiliation with EGRP/NCI/NIH were journaled as part of the analytic memos.

IV. RESULTS

This chapter is organized by data collection methods used in this case study: in-depth semi-structured interviews and document reviews. The presentation of the findings were framed around the three research questions and four main constructs. The first research question addressed the facilitators and barriers to sharing data in a public or controlled-access database or data repository; the second research question addressed the opportunities for improving the sharing of federally-funded research data in a public or controlled-access database or data repository; and the third research question addressed lessons learned from genomic data sharing practices that could enhance epidemiological data sharing.

The interview data from a total of 37 respondents as well as the data from the 5 documents reviewed were used to answer all three research questions. There were two institute-level (NIH) documents and three program-level (EGRP) documents that were identified, reviewed and found to be relevant to this dissertation research. The documents were used to further explore the first two research questions focused on facilitators, barriers and opportunities to enhance data sharing. The documents reviewed are:

- Compiled Public Comments on NIH Request for Information: *Processes for database of Genotypes and Phenotypes (dbGaP) Data Submission, Access, and Management*[33]
- Compiled public comments on NIH Request for Information*: Strategies for NIH Data Management, Sharing and Citation*[34]
- Three internal reports prepared for EGRP on an assessment of data sharing practices and processes in EGRP-funded cancer epidemiology cohorts (Table II).

[33] https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-044.html
[34] https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-015.html

TABLE II: DESCRIPTION OF DOCUMENTS REVIEWED

1. Compiled Public Comments on NIH Request for Information: Processes for database of Genotypes and Phenotypes (dbGaP) Data Submission, Access, and Management *(NOT-OD-17-044)*	Specific questions prepared by NIH, open for any member of the public to comment on
2. Compiled public comments on NIH Request for Information: Strategies for NIH Data Management, Sharing and Citation (Data Sharing Strategy Development) (NOT-OD-17-015)	Specific questions prepared by NIH, open for any member of the public to comment on
3. EGRP Internal Report #1 - January 2016. Evaluation Proposal of the NCI Epidemiology and Genomics Research Program (EGRP) Cohort Studies' Data Sharing Practices	Internal program reports from contractor based on specific tasks to evaluate data sharing practices in EGRP epidemiology cohorts
4. EGRP Internal Report #2 - July 2016. Findings from NCI Epidemiology and Genomics Research Program (EGRP) 9 Cohort Interviews on Data Sharing Policies and Practices	Internal program reports on semi-structured interviews with 9 cohorts to assess data sharing practices
5. EGRP Internal Report #3 - April 2017. Findings from Interviews with NCI Data Requestors on Epidemiology and Genomics Research Program (EGRP) Cohort Data Sharing Policies and Practices	Internal program reports from contractor based on specific tasks to evaluate data sharing practices in cancer epidemiology cohorts

The constructs that were considered to be key organizational level factors influencing the sharing of data in federally-funded research studies are: A) CULTURE AND PRACTICES, B) REGULATORY POLICY AND LAWS, C) RESOURCES, and D) TECHNOLOGICAL INFRASTRUCTURE. These are shown in the initial conceptual framework in chapter 3 (Figure 4). Within each of these main constructs are key factors that were developed through deductive coding as a means to help answer the research questions and confirm findings from prior research. The factors were developed a priori from findings from environmental scans and literature reviews, which helped with conceptualizing the initial conceptual framework for this

research. During the data collection and data analysis phase, an inductive approach was used to identify new factors that emerged during the analysis of the interview data, further informing the development of a revised conceptual framework. These are referred to as emergent factors.

A code list with definitions was generated prior to data collection based on established factors in the literature and through experiences in the field that influence data sharing among researchers. As new codes emerged from the data analysis, the codebook was updated with the emergent codes (Appendix E). To ensure consistency in the coding, the same code list was applied to both interview data and document reviews. For further validation of the coding, 10% of the of the interview data only was coded with a second coder.

As noted in the codebook, the terms *Facilitators, Barriers, and Opportunities* were defined to help make clear distinctions during coding and analysis of the data, and to help address research questions 1, 2 and 3. The code *Facilitators* was used in this study to describe factors that currently exist, are in place and working and therefore should be continued in order to achieve the goal of enhanced data sharing. The code B*arriers* was used to describe factors that are currently in the way that may likely hinder the ability to accomplish the goal of enhanced data sharing. The code *Opportunities* was used to describe new ideas or resources which might facilitate the goal of enhanced data sharing.

A. Semi-Structured Interviews

Thirty seven semi-structured interviews were conducted from April 17, 2018 – July 2, 2018, averaging 50 minutes per interview. Twenty five of the interviews were with EGRP-funded researchers and the remaining 12 with NIH staff. The researchers (also referred to as investigators or principal investigators (PI) throughout this document) included new investigators

and experienced investigators with research areas of focus in epidemiology, genomics and genetic epidemiology. Overall, 5 investigators identified genomics as their main research area of focus and 6 investigators identified genetic epidemiology as their main research area of focus. Given the low numbers compared to epidemiology (n=14), and the similarity in meaning of the research areas, a decision was made to combine the data from these two groups into one research area, "genomic/genetic epidemiology." This provided a richer and more meaningful analysis of data.

The NIH staff who participated in this study included analysts, scientists and director level staff across NIH and NCI. Collectively, the staff have many years of experiences in leadership roles and / or expertise in data sharing policy development, implementation and technical support; providing different perspectives and new ideas on opportunities for enhancing data sharing in NIH-funded research. TABLE III below describes the composition of participants in this study.

TABLE III: COMPOSITION OF STUDY PARTICIPANTS

No. Respondents	**New Investigators**	**Experienced Investigators**	Total
Epidemiologists	4	10	14
Genomicists / Genetic Epidemiologists	2	9	11
Total Investigators	**6**	**19**	**25**
Total NIH Staff			**12**
TOTAL			**37**

To protect the confidentiality of all participants, the findings are discussed in aggregate. The data from the investigators were analyzed by level of experience (new PI and experienced PI) as well as by their self-identified research area of focus (epidemiology, genomics / genetic epidemiology). The genomic/genetic epidemiology researchers from here on will be referred to as genomicists / genetic epidemiologists. As part of the final set of questions in the interview, respondents were asked to share some lessons learned from genomic data sharing practices that could be applied to epidemiology data sharing (research question 3). The analysis of research question 3 showed some repetition with earlier responses that elicited factors that they considered as facilitating or hindering data sharing, as well as opportunities to enhance data sharing.

For each of the factors described in the rest of this chapter, co-occurrence queries were run in NVivo across all factor codes to further explore patterns in the data and close relationships between factors, to help answer the research questions. The co-occurrence tables can be found in (Appendix F). The findings from this study are described below, organized by research question and framed by the constructs and related factors that facilitate or hinder data sharing. The factors identified as facilitators (from here on labeled as *Fac*) will be described first, followed by factors identified as barriers (from here on will be labeled as *Bar*), then those identified as opportunities (from here on labeled as *Opp*). Any emergent factors from here on will have '*E'* as the final letter in the code name.

Research Question 1: How do organizational level factors facilitate or hinder the sharing of research data in public or controlled-access databases or data repositories?

FACILITATORS – Culture and Practices

The construct, CULTURE AND PRACTICES, was defined as when the respondent discussed the culture, norms and practices of their academic institutions or NIH that may facilitate or hinder the sharing research data by NIH supported investigators in public or controlled-access data repositories. The a priori factors explored under this construct that facilitate data sharing are: *Intrinsic incentives* and *Culture differences in research fields*.

Intrinsic incentive - facilitator

The concept of incentives is generally considered to mean the same thing as motivation and was used in this manner to determine what types of things internally motivated the researchers to share their data. The factor, *Intrinsic incentive* as a facilitator for data sharing (*Intrinsic incentive-Fac*) was used when the respondents referred to sharing data for the advancement of science, knowledge gain, and as good citizens for the benefit of the public. It described the researcher's desire to do something good because it's the right thing to do. The external factor or extrinsic incentives that motivated researchers to share data are discussed as part of other constructs that will be described later in this chapter. There were 31 coding references for a priori code *Intrinsic incentive-Fac* mentioned by 22 respondents (59%).

TABLE IV: NUMBER OF RESPONDENTS WHO MENTIONED *INTRINSIC INCENTIVES* AS FACILITATORS

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total (25)
Epidemiologists (14)	1	6	**7 (50%)**
Genomicists/Genetic Epidemiologists (11)	2	6	**8 (73%)**
Investigators (25)	**3 (50%)**	**12 (63%)**	**15 (60%)**
NIH Staff (12)	[redacted]		**7 (58%)**
TOTAL (37)			**22 (59%)**

When asked what motivated researchers to share their data in public or controlled-access data repositories, more than half of the investigators and staff agreed that the personal benefit in their scientific research was a major incentive. The similarity in perspectives between investigators and staff is illustrated in the quotes below.

I think most of us [are] sharing the data to improve our capacity to understand the research question or improve our capability to do research. I think a minor part of it, NCI policy, is the data sharing requirement. (experienced PI, epidemiologist)

Well, I think there's a mix. There's some investigators that truly believe that data sharing is in their benefit because in the collective, if everybody's sharing, then everybody has more access to more data. There's also some investigators that don't see things that way … (staff)

In addition to personal benefit and the desire to advance scientific progress described in the previous quotes as incentives for sharing data, the respondents mentioned that the *Enforcement of policy* of the data sharing policy by NIH staff (a priori factor *Enforcement of policy* as a facilitator), as well as grant funding (a priori factor *Financial resources* as a facilitator), were incentives for sharing. Exploration of these factors, *Intrinsic incentives*-Fac, *Enforcement of policy-Fac* and *Financial resources-Fac*, (Table I, Appendix F), revealed similarities in perspectives between experienced and new investigators - epidemiologists (from here on stated as "epi" after respondent quotes) and genomicists / genetic epidemiologists (from here on stated as "gen/gen epi" after respondent quotes).

The perspective from new investigators, as illustrated in the quote below, and echoed by staff, corroborates the relationship between personal benefit (*Intrinsic incentive*) and *Financial resources-Fac,* as incentives for sharing. This depicts the perception of financial security among more experienced and well-funded investigators (compared to new

investigators) as related to the motivation or willingness to share data. Successful investigators from well-resourced institutions who have multiple large research project grants (R01), when compared to their less funded colleagues, were perceived to be more likely to share their data because their salary or academic achievement is not solely dependent a particular dataset and they may have other grants to support their researcher. It reinforces general sentiments across the respondents on the tension between financial constraints on the investigator end and the NIH mandate for data sharing.

... They've answered their research question, or maybe not answered, but they've got what they needed out of those data, and so they're much more willing to give them up, so to speak, to put them out to everybody else. I think that's key. And I would also be interested in seeing people who are currently R01 funded versus those who aren't. I think that those who have a really strong history of funding - they're not living paycheck to paycheck anymore. They're secure. They're doing fine. If somebody else wants to look at these data, great, because they've already got two or three things coming down the pipeline. So I think that makes them much more agreeable to share. (new PI, gen/gen epi)

Culture differences in research fields - facilitator

The a priori factor, *Culture differences in research fields* as a facilitator was used when respondents discussed the positive change in culture over time, including the sharing of data in the field of genomics compared to epidemiology as facilitating data sharing in research. There were 3 respondents (8%) who mentioned this factor including one staff and two senior investigators with expertise in epidemiology and genomic/genetic epidemiology. In the quote below, one of the investigators mentioned that researchers are more open to sharing data now compare to the past, and value sharing. It also illustrates convergence among the types of respondents with the theme of policy enforcement as the driving force behind this incremental change in culture and shift in thinking and attitude among researchers in different fields.

In addition, an experienced epidemiologist alluded in the quote below, to NIH's increasing recognition of some of the challenges investigators face with sharing data, thereby implying the need for opportunities to mitigate those issues. This could be done through changing the reward structure at institutions (*Reward structure changes needed-Opp* - defined as opportunities for addressing changes in institution's reward structure), something that was uniformly echoed by the two groups of investigators and staff as a major hinderance to sharing data in public or controlled-access data repositories.

Well, I think they [NIH] are listening to us and starting to get the idea that it's not as simple. … I think they're starting to appreciate that it's just not that easy. And I think investigators also are coming around. I know like in the old days, there was a very strong sense of - hey, this is my data, I worked hard to get it, why should I give it to you? What have you done for me? And I think that narrow attitude, I think, is really starting to fade and especially among the newer investigators appreciate the value in sharing, but just implementation is hard. (experienced PI, epi)

To support the comment from the experienced investigator, NIH staff respondents mentioned current efforts at NIH to understand the challenges with data sharing through workshops and targeted discussion with different individuals. In addition, NIH's process of soliciting feedback from the community on NIH's priorities for data sharing policy through the Request for Information (RFI) mechanism, is a practical attempt at *"trying to get with the public to educate within NIH, outside of NIH and stay aligned."*

FACILITATORS - Regulatory Policy and Laws

The construct, *REGULATORY POLICY AND LAWS,* was defined as any aspect of NIH data sharing policies, laws governing human subject research, participant privacy, including references to informed consents, Institutional Review Board (IRB) policies, and journal policies

for publications. The a priori factors explored under this construct as facilitating data sharing are: *Clarity of policy and Enforcement of policy*.

Clarity of Policy - facilitator

The a priori factor, *Clarity of policy* as a facilitator is defined as the clarity of NIH data sharing policies related to the timelines for when data should be shared, guidelines and expectations, communication and feedback about the policy requirements. There were 16 of 37 respondents (43%) in this study who mentioned *Clarity of policy a*s a facilitator (*Clarity of policy-Fac*).

TABLE V: NUMBER OF RESPONDENTS WHO MENTIONED *CLARITY OF POLICY* AS A FACILITATOR

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	0	3	3 (21%)
Genomicists/Genetic Epidemiologists (11)	1	5	6 (55%)
Investigators (25)	**1 (17%)**	**8 (42%)**	**9 (36%)**
NIH Staff (12)			**7 (58%)**
TOTAL (37)			**16 (43%)**

There were 44 references of this code in the data and about one-third of the investigators and more than half of the staff interviewed thought that the policy was clear in terms of the expectations, and that the timelines in the policy for sharing data were reasonable. The similarity in perspectives among the types of respondents on the policy timelines is captured by the quote below.

I think they're reasonable. The little challenge is that it's hard to know from here when the data's cleaned. You know, you've got to email the investigator and trust that they tell you .. So there's some of that. So I think the timelines are fine, it's just sort of the process of getting the data in the right databases takes a long time. (staff)

This quote further highlights strong relationship with co-occurring a priori factors – *Lack of clarity of policy-Bar* (defined as vagueness, ambiguity and lack of clarity of policy), *clarity of policy needed-Opp* (defined as opportunities to address the lack of clarity in the policy); *Training-Fac* (defined as existing training or education resources and materials) and *Clarity in submission and access processes-Fac* (defined as clear processes for data submission and access in data repositories) (Table III, Appendix F).

Enforcement of policy - facilitator

The a priori factor, *Enforcement of policy* as a facilitator is defined as when respondents discussed enforcement of data sharing policies by NIH staff as a way to enhance data sharing. This was an important factor under the POLICY construct, with 87 references of this factor mentioned by 26 of the 37 respondents (70%).

TABLE VI: NUMBER OF RESPONDENTS WHO MENTIONED *ENFORCEMENT OF POLICY* AS A FACILITATOR

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	1	5	6 (43%)
Genomicists/Genetic Epidemiologists (11)	2	6	8 (73%)
Investigators (25)	**3 (50%)**	**11 (58%)**	**14 (72%)**
NIH Staff (12)			**12 (100%)**
Total			**26 (70%)**

There were a lot of comments made among the investigators and staff related to their perception of enforcement of data sharing polices and the implications. The staff mentioned a wide range of enforcement strategies from as simple as an informal conversation between NIH and the university to withholding of grant funding as part of the terms and conditions of awards. While this is currently the case given that it's now included as part of the terms and awards of grants, it was not clear that the investigators were aware of this enforcement strategy.

The support from leadership (*Leadership support-Fac*) from both NIH and academic institutions is critical in establishing and enforcing rules and guidelines to share data for research purposes, while assuring the protection of human research data. In addition, leadership support of staff in the implementation of policy is also important. Staff were asked if they felt that they had the authority or capability to enforce the data sharing policy and the response was a mix depending on their level of experiences and role in the data sharing process. The underlying factor was the support of leadership, indicating the value and important role of leadership in policy enforcement.

I don't think PDs [Program Directors] can do it on their own – they have to have the backing of management because some people who object are concerned they're going to go higher up, and you have to have a consistent message. So I didn't feel like I had the authority to really press the case .. now it can be part of a funding decision.. I think it should be part of the funding decision. I think that's how you end up having the authority if you have people from higher administration saying, "Yes, this is important and this is how we're gonna deal with it. (staff)

Enforcement of policy-Fac was also found to be related to the factors administrative / technical resources (Inadequate *Admin/tech resources-Bar* – defined as the lack or limited resources in terms of personnel time, effort and technical support) and financial resources (*Inadequate Financial resources-Bar* – defined as the lack of or inadequate funding and cost). The investigators mentioned that while the enforcement of the policy is effective in pressuring

researchers to share their data in data repositories, there's the added burden of time and effort to prepare and submit the data as well as the added financial burden to support the process. This sentiment was echoed among the new and experienced investigators regardless of their research areas of focus, and illustrated in the quote below.

This factor, *Enforcement of policy-Fac*, was also observed in the document reviews (Appendix G).

FACILITATORS - Resources

The construct, RESOURCES, was defined as administrative, technical and financial resources for data sharing, including support from leadership and training as playing a key role in either facilitating or hindering the sharing of data in public or controlled-access repositories. Understanding these types of support and the extent to which they hinder or facilitate data sharing provides an opportunity for NIH and academic institutions to address these issues at the organizational / institutional level. The following a priori factors were used to capture the meaning of this construct, as facilitators: 1) *Administrative / Technical resources*; 2) *Financial resources*; 3) *Leadership support*; and 4) *Training*. These factors are described in detail below.

Administrative /Technical Resources - Facilitator

A priori factor, *Administrative and technical resources* as a facilitator *(Admin/Tech resources-Fac)*, was defined as when respondents described administrative and technical support including time and effort of personnel as facilitating data sharing processes. 29 of the 37 respondents (78%) mentioned *Administrative / technical resources* as facilitating data sharing, with 77 code references across the data. This was one of the most important key factors to

sharing data, as evidenced by the prevalence of the code across the different investigators and their self-identified research areas of focus and staff.

TABLE VII: NUMBER OF RESPONDENTS WHO MENTIONED *ADMINISTRATIVE / TECHNICAL RESOURCES* AS A FACILITATOR

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	3	6	**9 (64%)**
Genomicists/Genetic Epidemiologists (11)	1	9	**10 (91%)**
Investigators (25)	**4 (67%)**	**15 (79%)**	**19 (76%)**
NIH Staff (12)			**10 (83%)**
TOTAL (37)			**29 (78%)**

The data sharing process is a complex process that requires a substantial amount of investment in resources and manpower. Having adequate number of staff with the right expertise to perform administrative and technical tasks that support data sharing is critical to facilitating the sharing of research data in data repositories. The qualitative analysis shows that the code, *Financial resources-Fac*, co-occurred strongly with *Admin/tech resources-Fac* and also indicates convergence in themes among the different types of NIH staff represented in this study as well as the new and experienced epidemiologists and genomicists/genetic epidemiologists.

The close relationship between administrative resources and financial resources is illustrated in the quote below, from the researcher perspective which is also consistent with comments from staff. This quote points out funding to support the salary or personnel time and effort in their grant for the purposes of data sharing. The respondent mentioned that the administrative burden for data sharing that requires financial support such as personnel time and

effort involved in the data preparation, cleaning and documentation process could be alleviated with financial support.

Well, you know what? It's either give the investigators money or it's the NIH funds it and they have the people, and they do the work. They have to get on a plane and come down here and sort through files and do all that, [...], does it? We can give them access to our systems. It's a huge amount of time and money to comply with this. (experienced PI, gen/gen. epi)

The example above describes the level of frustration among investigators, emphasizing the lack of adequate administrative and financial support for investigators share data in data repositories. The sentiment here is that if adequate resources were available, it might be a lot easier to share data. This was echoed among the new and experienced as well as the epidemiologists and the geneticists/genetic-epidemiologists. Interestingly, from the perspective of a NIH staff, it was mentioned that there is an equally important need for administrative and financial resources on the receiving end so NIH can more effectively manage the data submission in the repository.

I think clarity in process is big as well as the resources, not only from the submitter's side, but also from the NIH's side. They have to dedicate more resources to facilitating the submissions, to making the infrastructure available in a way that's clear and not going to cause questions of "Well, what repository do I send it to? How do I actually pay for that repository?" (staff)

When respondents were asked about current factors that are currently in place facilitating data sharing, both the staff and researchers mentioned the existence of NIH data repositories and NIH's investment in the infrastructure. The following example emphasizes the close connection between the two factors and the value of leadership in supporting administrative and technical processes to enhance data sharing.

I mean the repository itself exists to share data. So at its centrality, the fact that we have a repository. And NIH invests in it, it's staffed, we have a governance structure around it. It's growing. (staff)

Further analysis of the data showed a strong relationship between *Admin/tech resources-Fac and several other factors - Clarity of submission/access process-Fac* (defined as clear guidelines and instructions on the data submission/access process as a facilitator) with representation from new and experienced epidemiologist and genomicists/genetic-epidemiologists; and *Expertise-Fac* (defined as technical expertise and knowledge of researchers and staff of the data submission or access process and of what resources are available). When respondents were asked what was essential for submission of data, these three factors were some of the most recurring themes.

I think there's two things. I mean one is having – or one thing: your level of expertise and understanding the process. And so, I mean having a programmer who can communicate with those at dbGaP for example, and understanding – who have experience in the process on how to do this. And the other one is on the informed consent end of things: to be able to explain if it's a collaborative study, to be able to explain to different institutions that don't have experience in this area about what's exactly required on the paperwork side. I think those are two things, and I mean I have two different people in each of these areas who can work with – one works directly with dbGaP and the other one works with the investigators to make sure that we have everything in place and it's quite clear to the investigators how they actually fill out these institutional certifications, or how they should be communicating with our IRBs to make sure it's done correctly. (experienced PI, gen/gen. epi)

Highlighted in the quote above is that not all institutions have the resources or expertise needed to handle the paperwork and also the process may not always be intuitive for the investigators in terms of the process and institutional requirements needed before they can proceed with the data submission. With adequate infrastructure, including staff with the proper expertise and well-functioning data repository, investigators may feel less administrative burden and more apt to share data in data repositories. The qualitative analysis showed that both the staff and investigators, both new and experienced epidemiologists and genomicists/genetic epidemiologists confirm the relationship between *Admin/tech resources-Fac* and the factor,

Repository capabilities-Fac (defined as the capacity, efficiency and adequacy of the repository) as seen in the quote below.

Well, simply the common repositories are a huge thing 'cause you don't have to go through an investigator. They're just there, right? I mean, there's a process, but you don't have to convince a colleague to, you know, take time to do this. So that's a big step so that having — again, having to go to an investigator, and usually they're happy to help you, but they're busy, so it might take time, or it might not be a streamlined process, depending on what they have, but once it's in a central repository it makes everything a lot easier so you don't have to deal with the investigator step. (experienced PI, gen/gen. epi)

This factor, *Administrative and technical resources* as a facilitator, was also observed in the document reviews (Appendix G).

Financial resources - facilitator

A priori factor, *Financial resources* as a facilitator *(Financial resources-Fac)* is defined as when respondents described the availability of financial resources such as funding support to do research or to hire personnel. 25 of the 37 respondents (68%) mentioned *Financial resources* as a facilitator for data sharing, with 57 coding references across the interview data.

TABLE VIII: NUMBER OF RESPONDENTS WHO MENTIONED *FINANCIAL RESOURCES* AS A FACILITATOR

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	2	7	9 (64%)
Genomicists/Genetic Epidemiologists (11)	1	5	6 (55%)
Investigators (25)	**3 (50%)**	**12 (63%)**	**15 (60%)**
NIH Staff (12)			**10 (83%)**
TOTAL (37)			**25 (68%)**

This was an important factor for enhancing data sharing, as indicated by the high prevalence of the code mentioned across the different respondent groups shown in Table VIII above. The following example indicates that personnel is needed to do the data sharing work and some form of funding is required to pay their salary.

Having money to have an employee whose role it is to do that to get the information into a meta file, to get the information in the right form so that it can go into a public use database or controlled-access database. Or to put the data in the form so that it goes to, it can go into a consortium along with the appropriate documentation. (experienced PI, epi)

There's a general sentiment from the investigators and staff that more funding will facilitate data sharing, quickly. One of the investigators begged for NIH to come up with solutions such as funding to support the effort of data sharing e.g. set aside funding for data sharing. Given the current climate of funding, investigators are constantly competing with their peers for a limited pool of funds. NIH has supported data sharing through release of funding opportunities specific for data sharing, for example, administrative data sharing supplements in EGRP[35]. Another example of how NIH is facilitating data sharing is through its investment in repositories at no cost to the investigators.

When asked what NIH was doing to facilitate data sharing, both the investigators and NIH staff mentioned leadership support as it related to the provision of financial resources and investment in physical resources and release of funding opportunity announcements. This was echoed by both investigators and staff and illustrated in the following example.

Well, I think that, you know, that many more funding announcements are going out requiring it. .... So I do believe that there is slowly but surely becoming greater recognition of what is needed, ... I think where they're putting it [data sharing] into their funding announcements that that's beneficial ... Oh, they're also making open available databases. So things like dbGaP. These are not of any cost to the public. So the GDC [Genomic Data Commons], they might get annoyed a little bit with the process, but at the end of the day, they're still publicly accessible resources and you know it's very widely used, so I do think NIH is doing

[35] https://grants.nih.gov/grants/guide/pa-files/pa-18-748.html

a good job of trying to create systems that allow the public access to data at no cost to themselves.(staff)

There were other comments from investigators that institutions should provide support beyond end of the grant and funding for investigator's time to do analysis so they could share data quickly. The factors, *Admin/tech resources-Fac* and *Leadership support-Fac* co-occurred strongly with *Financial resources-Fac.* Other co-occurring factors are shown in Table VI, Appendix F.

This factor, *Financial resources* as a facilitator, was also observed in the document reviews (Appendix G).

Leadership support

The a priori factor, *Leadership support* as a facilitator (*Leadership support-Fac)*, is defined as the support from the institution or organization's leadership including provision of resources and oversight to promote data sharing among NIH supported research. There were 44 coding references mentioned by 19 respondents.

TABLE IX: NUMBER OF RESPONDENTS WHO MENTIONED *LEADERSHIP SUPPORT* AS A FACILITATOR

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	2	8	**10 (71%)**
Genomicists/Genetic Epidemiologists (11)	0	1	**1 (9%)**
Investigators (25)	**2 (33%)**	**9 (47%)**	**11 (44%)**
NIH Staff (12)	[redacted]		**8 (67%)**
TOTAL (37)			**19 (51%)**

While both investigators and staff agreed that leadership support was an important factor in facilitating data sharing, more references to the code *Leadership support-Fac* was from staff compared to the investigators. This has to do with their insider knowledge of existing and ongoing efforts by NIH leadership to support data sharing, given data sharing is a priority for the Institute. 71% of the investigators that mentioned this factor were epidemiologists; an interesting observation that is worth exploring.

Interestingly, when respondents were asked what NIH is doing well to facilitate data sharing, just about all the investigators and staff provided positive responses as seen from the perspective of one staff respondent.

I think that NIH has made it clear that that is a high priority for them, so right now, as I mentioned, we don't really have a very effective data sharing policy; however, a lot of the programs within institutes and centers have made a step forward, same with journals, to be able to say that this is a high priority to us and we want the data shared. So I think making that clear, that this is a priority, and having the community recognize that is a big thing we've done. … the White House has extreme interest in open access and sharing of data, so not only are we feeling the internal heat for us to get a policy out, but also this administration is very interested in. so I think that's a positive thing. (staff)

In giving their perspective on what NIH is doing well to support data sharing, two of the new investigators, a genomicist/genetic epidemiologist and an epidemiologist provided positive feedback.

NIH has made it about as easy as it can possibly be (new PI, gen/gen epi)

NIH is doing a lot things well like the Cohort Consortium has been an incredible model of data sharing with a lot of good research coming out of the working groups, supporting the annual meetings and some infrastructure of the consortium – a good collaborative model where you still need to go to each individual cohort and request their approval to participate and request data directly from them. NCI has done so much through the Cohort Consortium and it's a great model of data sharing, and I don't know if the NCI considers that data sharing but I certainly do (new PI, epi)

On the contrary, there was one divergent perspective on this, from an experienced genomicist/genetic epidemiologist who did not think that NIH was doing anything well in terms

of data sharing. Part of the reason given was the perception of "*failure of NIH to really identify with the investigators to understand some of the challenges or issues around sharing data"* and the complexities of it instead of mandating sharing, and not providing the resources in the grant to support the activity. Although, some other participants expressed similar concern around the challenges of data sharing, they appreciated some of current and planned efforts by NIH to make data sharing easier for investigators.

Boy, I can't think of anything that they're doing well, to be honest. I think again, my view and I think it's shared by many of the colleagues is that NIH is not doing a good job. They're doing a good job of mandating, but they're not doing a good job of actually assisting facilitating. One other thing to throw in there. I think it's part of this is a disconnect in terms of the cultures. Many of the people at NIH don't understand, have been at NIH too long. They don't understand the realities of the extramural community, research community and the demands that are placed on them within their institutions and so on in terms of financial needs and support and so forth. So I think that is also a disconnect that I see. I don't think the people at NIH recognize the reality of what investigators are dealing with. (experienced PI, gen/gen epi)

Another area of divergence in perspective is among the NIH staff where two staff respondents had different views on how well-thought out the data sharing policy is. One staff described the policy as well-thought out and another staff participant doesn't think the policy was well-thought out. The difference in opinion is dependent on their roles, involvement, observation and experiences with data sharing at NIH.

The data analysis showed that *Financial resources-Fac* co-occurred strongly with *Leadership support-Fac*, along with other co-occurring codes as described in Table VII, Appendix F. One staff respondent mentioned that while NIH does invest in resources to support data sharing, there is probably an opportunity for NIH to evaluate how it invests in the data sharing infrastructure in a way that is the most cost-effective and beneficial to the researchers and the public.

I guess this sounds sort of self-serving, but I think that they invest in a lot of stuff. They do this with varying degrees of efficiency or probability of success. In a self-serving way, I'd say NCBI has succeeded because we have a mixture of researchers and engineers and we recognize what they think of many times as research questions are really engineering problems. (staff)

This factor, *Leadership support-Fac*, was also observed in the document reviews (Appendix G).

Training

The a priori factor, *Training*, as a facilitating factor (*Training-Fac*) is defined as training for researchers and staff to enhance knowledge and skills for effective and successful data management, submission and access. There were 22 coding references for *Training-Fac* and was mentioned by 9 respondents.

TABLE X: NUMBER OF RESPONDENTS WHO MENTIONED *TRAINING* AS A FACILITATOR

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	0	1	**1 (7%)**
Genomicists/Genetic Epidemiologists (11)	0	2	**2 (18%)**
Investigators (25)	**0**	**3 (16%)**	**3 (12%)**
NIH Staff (12)	[redacted]		**6 (50%)**
TOTAL (37)			**9 (24%)**

To help investigators and staff better understand data sharing processes and expectations, NIH has provided guidance on their website. However, the respondents indicated that NIH could provide clearer instructions and guidelines, standards, procedures for use of databases / repositories, flowcharts to show process, templates, and protocols including who to contact for

technical / administrative questions. It is not clear how many of the participants were truly aware of these existing resources, have used them or access them frequently. Some researchers may have limited experience and knowledge of data sharing processes, and some staff may have limited experience and training to implement the policy and understand the data sharing processes. The quote below describes some facilitators and opportunities to enhance knowledge and skills around data sharing, especially among novice users.

And I guess the other thing I would say that they've done is that there are YouTube self-help tools available, but they -- again an area for improvement is they tend to be most friendly to those who are high-volume users or very technically savvy with regard to the data submission process versus the occasional data submitter or the newbee to genomic analysis. So targeting our resources and -- targeting our resources to more novice users as well as increasing personnel and updating systems to make them more intuitive are all things that we are -- that need to happen and that are now in active discussions. I would say that, you know, the increase in quality and efficiency of the DAC system has been essential to improving access.

The are several other factors that co-occurred with *Training-fac* (Table VIII, Appendix F) but the most prevalent ones were: *Clarity of policy-Fac, Clarity of submission/access process-Fac* (19 coding references); *Expertise-Fac* (13 coding references).

FACILITATORS – Technological Infrastructure

The construct TECHNOLOGICAL INFRASTRUCTURE was defined as when respondents referred to the technical aspects and the expertise for sharing or accessing data in a data repository. This is related to the capabilities of the data repository, the knowledge and clarity of the data submission and access process, the management and preparation of data, including formatting and documentation. Understanding how these facilitate or hinder data sharing will provide insight for how to enhance data sharing practices. The following a priori factors were used to capture the meaning of this construct as a facilitator for data sharing. 1)

Analytic data complexity; 2) *Clarity of submission/access process;* 3) *Expertise;* and 4) *Repository capabilities*.

Analytic data complexity - facilitator

The a priori factor, *Analytic data complexity* is defined as when respondents referred to nuances related to the analytic dataset and variables, including the data format, standardization of data format, and documentation or annotation of data as facilitating data sharing *(Analytic data complexity-Fac)*. This includes comments on the complex nature of phenotypic or non-genomic data. There were 27 coding references of *Analytic data complexity-Fac* mentioned by 14 of the 37 (38%) of the respondents as facilitating data sharing.

TABLE XI: NUMBER OF RESPONDENTS WHO MENTIONED *ANALYTIC DATA COMPLEXITY* AS A FACILITATOR

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total (25)
Epidemiologists (14)	1	3	**4 (29%)**
Genomicists/Genetic Epidemiologists (11)	0	5	**5 (45%)**
Investigators (25)	**1 (17%)**	**8 (42%)**	**9 (36%)**
NIH Staff (12)	[redacted]		**5 (42%)**
TOTAL (37)			**14 (38%)**

Both epidemiologists and genomicists/genetic epidemiologists, as well as staff mentioned the importance of cleaned and well-documented data as essential for data submission and data access. One of the investigators mentioned the value of having some knowledge of the quality control process for the data deposited in the repositories. Both investigators and staff emphasized the value of well-cleaned and well-documented metadata for secondary analysis; a big part of the goal of data sharing in data repositories so that others may have access to the data and use it to

answer new research questions. In addition, the quote below depicts the concern for misuse and misinterpretation of data if there's no or inadequate data documentation provided to go with the data. It shows the relationship between the quality of data in the repository and indirect impact on a researcher's career which could influence their willingness to sharing data in data repository.

So whether there are guidelines or recommendations, I think that would be helpful in terms of sort of just knowing what is the minimum to get someone able to use a data set in the proper way I think without kind of misinterpreting or misunderstanding the data, and that is also sort of standalone in that you wouldn't have to always go back to the lead investigator or to the programmer? So I think part of it is both having a data set that is clean and usable and enough either documentation, But so like if this proper documentation is needed to ensure that others can use the data set to prevent misuse, misinterpretation or confusion. (experienced PI, gen/gen epi)

This factor, *Analytic data complexity-Fac,* was also observed in the document reviews (Appendix G).

Clarity of submission / access process - Facilitator

The a priori factor, *Clarity of submission/access process* as a facilitator *(Clarity of submission/access process-Fac)* is defined as when respondents referred to the clear administrative / technical processes for data submission and access from a data repository. This includes when respondents mentioned the availability of clear instructions and guidelines, templates, guides, templates and protocols, including who to contact for technical questions and for administrative processes both at NIH and at the academic institution. There were 92 coding references of *Clarity of submission/access process* as a facilitator for data sharing, mentioned by 27 of the 37 respondents (73%).

TABLE XII: NUMBER OF RESPONDENTS WHO MENTIONED *CLARITY OF SUBMISSION/ACCESS PROCESS* AS A FACILITATOR

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total (25)
Epidemiologists (14)	1	6	**7 (50%)**
Genomicists/Genetic Epidemiologists (11)	2	7	**9 (82%)**
Investigators (25)	**3 (50%)**	**13 (68%)**	**16 (64%)**
NIH Staff (12)			**11 (92%)**
Total (37)			**27 (73%)**

This was mentioned by both the investigators and staff as very important factor that influenced the success of data submission or access in a data repository. NIH has developed user guides and tools online on the dbGaP process to help facilitate sharing in data repositories. However one staff mentioned that these were likely less intuitive for the occasional users and designed for those that are more frequent users of dbGaP. They suggested "*targeting our resources to more novice users."* This includes instructions for downloading data, and recommendations for what data is needed and useful to submit to repositories.

The new investigators who didn't have a lot of experience with dbGaP had similar comments as the staff, to make the instructions clearer, emphasizing that although the instructions are available, there may be a need and opportunity to make it clearer and to provide training for users on the data sharing process, tools and guides to facilitate sharing. The quote below highlights that both researchers and staff will benefit from training.

While at the same time, if you get people who have made more than one submission, then it just takes minutes. We're trying to increase some of our guidance documents and materials on our website to be able to help people get over some of these speed bumps. So that's one big thing we're in the process of doing. I think one of the other things, in addition to educating the public, is that we're also trying to educate staff, particularly program staff or those who are involved from the submission side, or access side here to be able to have them well-trained and

understanding nuances and how to be able to kind of streamline these particular processes. So if we can train from the inside and the outside, I think that's one of our biggest things we're working on now. (staff)

NIH/NCI has also provided a few relevant templates, forms and described the processes for what is needed for data submission and accessing data. The question again is whether people are aware of these resources and if it's easy to find them on the website. Training and education could help increase awareness. Respondents also alluded to efforts made by NIH to increase training and awareness of relevant information and tools developed as facilitating data sharing and should be continued at different levels.

I think something that helps promote understanding throughout available data are meetings [at my institution], both on campus as well as that are open to all…. So we have all the meetings open so people should be aware and we'll say come attend this meeting, you can understand what it takes to organize a proposal and get the data. So I think those, you can disseminate that information, increase awareness about the opportunities in your academic presentations and on campus and out of campus as well. …I think the NCI also periodically has forums where they pull investigators into. I think those types of forums are important to bring the leaders in the field on specific topics to address the needs and enhance the discussion, the face to face discussions that will let you know this is what the possibility is if we are coming from individual cohorts and we've merged this. So I think those, sponsoring those types of forums are important as well and RFAs that are being sent out to encourage that. (experienced PI, epi)

Expertise - facilitator

The a priori factor, *Expertise* as a facilitator *(Expertise-Fac),* is defined as when respondents referred to researchers or staff expertise and knowledge of technical and administrative processes related to the data submission process. This includes knowledge of and experience with, and awareness of available resources, tools and systems that support data submission and access processes such as data preparation, access and formatting for submission. There were 51 coding references of *Expertise-Fac,* mentioned by 20 of the 37 (54%) respondents.

TABLE XIII: NUMBER OF RESPONDENTS WHO MENTIONED *EXPERTISE* AS A FACILITATOR

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total (25)
Epidemiologists (14)	1	3	**4 (29%)**
Genomicists/Genetic Epidemiologists (11)	2	6	**8 (73%)**
Investigators (25)	**3 (50%)**	**9 (47%)**	**12 (48%)**
NIH Staff (12)			**8 (67%)**
TOTAL (37)			**20 (54%)**

Across the data, staff and investigators including new and experienced epidemiologists and genomicists/genetic epidemiologists, indicated that having the right knowledge and expertise with the submission or access process will facilitate data sharing. The a priori factor, *Expertise-Fac* is closely related to *Clarity of submission/access process-Fac*, described in the previous section. The quote below illustrates both investigator and staff perspectives, indicating a learning curve for new users and clearer processes at the individual level and organizational level as key to successful data submission and access.

I think there's two things. I mean one is having – or one thing: your level of expertise and understanding the process. And so, I mean having a programmer who can communicate with those at dbGaP for example, and understanding – who have experience in the process on how to do this. And the other one is on the informed consent end of things: to be able to explain if it's a collaborative study, to be able to explain to different institutions that don't have experience in this area about what's exactly required on the paperwork side. (experienced PI, gen/gen epi)

Respondents indicated that more staff is needed to support data sharing processes at the Universities, especially where the research teams are lacking in the right expertise and knowledge about NIH data sharing policy requirements. From the perspective of the academic institution, in order for the researchers to meet the data sharing policies, they need to know what it takes and in some cases, the investigators rely on their data analysts or programmers and their

institution's research office for guidance on this. This does not always mean that these analysts and programmers and research officials always have the proper knowledge of what resources and data are available, and the right expertise needed for successful data sharing, e.g. knowing who the institution's signing official is.

In terms of support from the perspective of NIH, participants also indicated the need for more support from staff at the NCBI/NIH who manage the dbGaP process and who can assist investigators with the technical and administrative aspects of data submission and access. This would include people that could be called on if researchers needed to know where and how to deposit or access data e.g. IT support or the helpdesk.

Respondents indicated that there's a learning curve for most people who are new users or not frequent users of dbGaP and this could be intimidating, and probably instill reluctance to data sharing through the complicated system and process. Despite the availability of the tools, materials and resources to support data submission in a repository, of note is the level of awareness and usability of these resources to facilitate sharing. These provide an opportunity for training and education around tools and materials to help researchers at all levels feel more comfortable sharing data in repositories, and to increase awareness of existing resources to help facilitate data submission in repositories.

You know, there is a lot of how-to guides and videos, but, I don't know how much people read them, or if they know about them. But definitely, I think some of them have quite a bit of trouble, particularly in a first go around. It's not totally intuitive to them. … the one thing I will say for the dbGaP process, even though I think it's quite complicated, one thing that helps that process is that it's a defined process. There are all these materials that are developed, and people know about the resource. When you say dbGaP, the research committee knows exactly what you're talking about. So I think that it's a database that people know well. I don't know if people know this, is that you can actually email them when you have any problems, and they'll walk you through stuff. I'm just sure that they have a help email, and they do respond to it. So I do think there are these resources and it is a defined process, that all does help. It's just, particularly for new people, it feels pretty complicated. And there's different pieces and people do different parts and I think that can be a problem. (staff)

The example below indicates that clear guidelines and processes make it easier for researchers to understand the expectations and share their data, and for the staff, it helps with building their knowledge, skills and expertise around the accurate, consistent and effective implementation of data sharing policies among their grantees. This is needed for effective compliance and enforcement of data sharing policies for the benefit of the public. The quote below shows the close relationship between the factors *Clarity of submission/access process-Fac*, *Analytic data complexity-Fac*, and *Expertise-Fac.*

so if you get postdocs or students or the people who would normally be uploading this data, and they don't normally work with it, then now you're stuck. Because it's going to be rejected if it's not in the right format. ... But if there could be tools or good education support for uploading the data relatively easily, so even if you do require formatting that you have some sort of process in place that allows a more naïve user to be able to come in and walk through some sort of process that would allow them to format the data or answer the appropriate questions. Then I think that you would get a lot more data sharing, because sometimes I really think it's just physically being unable to upload, if you haven't used a particular software or system or format.(staff)

Repository capabilities - facilitator

The a priori factor, *Repository capabilities* as a facilitator (*Repository capabilities-Fac*) is defined as when respondents referred to the capacity, effectiveness, adequacy, and efficiency of the data repository as facilitating data sharing. There were 64 coding references of *Repository capabilities-Fac*, mentioned by 26 of the 37 (70%) respondents.

TABLE XIV: NUMBER OF RESPONDENTS WHO MENTIONED *REPOSITORY CAPABILITIES* AS A FACILITATOR

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total (25)
Epidemiologists (14)	2	4	**6 (43%)**
Genomicists/Genetic Epidemiologists (11)	2	7	**9 (82%)**
Investigators (25)	**4 (67%)**	**11 (58%)**	**15 (60%)**
NIH Staff (12)			**11 (92%)**
TOTAL (37)			**26 (70%)**

The analysis of the data showed a very close relationship between the factors *Repository capabilities-Fac, Clarity of submission/access process-Fac* and *Training-Fac.* This was evident in the responses from respondents when asked to describe some of the factors that are currently in place that are helping to facilitate the submission of data in repositories or databases. Most of the respondents mentioned the existence of a data repository as a major facilitator. Interestingly, the quote below from an experienced investigator emphasizes the benefit and ease of open data access via repositories, compared to the direct collaborative method which is not always ideal according to the comment below. This is a different perspective from the general negative sentiment around the capabilities of dbGaP in terms of the challenges with the data submission and access process in dbGaP.

Well, simply the common repositories are a huge thing 'cause you don't have to go through an investigator. They're just there, right? I mean, there's a process, but you don't have to convince a colleague to, you know, take time to do this. So that's a big step — again, having to go to an investigator, and usually they're happy to help you, but they're busy, so it might take time, or it might be not a streamlined process, depending on what they have, but once it's in a central repository it's certainly makes everything a lot easier so you don't have to deal with the investigator step. (experienced PI, gen/gen epi)

The existence of the NCI Data Access Committee (DAC) is one of the factors mentioned by staff as an important factor facilitating data sharing. The process for requesting data has been streamlined and made more efficient through a centralized coordinating system at NIH. Given the challenges mentioned by some of the respondents with accessing data, it will be a good opportunity to get the perspective of the researchers on the effectiveness and efficiency of the DAC in promoting data sharing.

I would say that, you know, the increase in quality and efficiency of the DAC system has been essential to improving access. (staff)

On the submission side, the example below reports on mixed reviews on the ease of data submission process from users of dbGaP. The quality control in place to assure high quality data is deposited and shared in dbGaP is invaluable. However this results in delays in the process and highlights opportunities such as increased administrative support at the NCBI to bridge the gap between assuring quality checks and increased data processing.

Some people say it's easy and some people think it's worse than donating a kidney. It depends on what they want. … We try to make it easy for everyone. Again, there's questions about quality … and [the] need to understand it and [the] need to have them fix errors. And some people resent that, some people were appreciative of that. Some people are not responsive. … overall, the compliance is pretty high .. If anything, people get frustrated with delays and waits (staff)

For enhanced sharing in data repositories, the importance of having adequate resources, automated systems with user friendly interface, adequate capacity cannot be understated. This was a shared feeling among the staff and investigators and some ongoing efforts in place at NIH that support this is shown in the following quote by a staff.

we try to automate or make things electronic as much as we can. So for the submission process, we now have a completely electronic submission portal where once you type in your grant number or whatever it is that you're funded through, it pulls your information, your name,

your study information, to be able to reduce errors but also reduce time that you have to type it in….We tried to standardize a lot of things so data that is submitted, it's, obviously as I mentioned, it's submitted according to the original informed consent (staff).

In addition, an experienced epidemiologist suggested a centralized system or a centralized repository that is managed by NIH, to help relieve investigators of the administrative burden from data sharing. This was compared to having a model like pub med central. The other factors co-occurring with *Repository capabilities-Fac* can be found in Table XII, Appendix F.

BARRIERS - Culture and Practices

Career Concerns - Barrier

The a priori *factor,* individual *Career concerns,* as a barrier for data sharing *(Career concerns-Bar)*, is defined as when respondents described individual researchers' concerns or perceived threats to their careers related to data misuse, misinterpretation, scooping and negative criticism or fear of others finding errors in their analyses. There were 59 coding references for *Career concerns-Bar*, mentioned by 26 of 37 (70%) respondents (70%).

TABLE XV: NUMBER OF RESPONDENTS WHO MENTIONED *CAREER CONCERNS* AS A BARRIER

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	4	9	**13 (92%)**
Genomicists/Genetic Epidemiologists (11)	2	3	**5 (45%)**
Investigators (25)	**6 (100%)**	**12 (63%)**	**18 (72%)**
NIH Staff (12)	[redacted]		**8 (67%)**
TOTAL (37)			**26 (70%)**

Overall, most of the respondents mentioned individual *career concerns* as a barrier to data sharing, although it was mentioned most by the epidemiologists (92%) compared to the genomicists/genetic epidemiologists (45%). Respondents including new and experienced epidemiologists and genetic / genetic-epidemiologists expressed concern that users who weren't part of a study would not fully understand the data generated from large complex epidemiology studies to be able to do good analysis. They fear that if other users don't have an understanding of the epidemiology data variables, study design, and nuances of the data, they are more likely to misinterpret the data if accessed directly from a data repository; which on some level they feel negatively impacts the career of the original data generator. One staff and an investigator describe this in the quotes below.

… there is a little bit of a cultural thing in epidemiology where you know people, epidemiologists can spend their entire career on sort of a give it like one large scale study in a given data batch. And so that makes them much less inclined to share because it is sort of individually more important to them. And so I think that's certainly part of it. (staff)

And there's one other thing that I think is really important to note when it comes to data sharing, and just uploading something to a network for sharing – there's often significant nuance about the studies and how they were collected: differences that may occur, issues of study design that somebody outside won't be able to comprehend, and so there's many advantages to sharing data and transparency and open access, but there are risks that people who don't understand the data can actually do incorrect analyses and generate incorrect results.(experienced PI, gen/gen epi)

The concern expressed by the investigators and echoed by staff, about the negative impact sharing their data could have on their career and in some instances their reputation in the scientific community from data misuse, misinterpretation and scooping is related to the lack of clarity or the vagueness in the policy, in terms of the requirements, expectations, processes or timeline for sharing data (*lack of clarity of policy-Bar).* This includes tension between knowing

when the data should be shared in accordance to the policy, and when they personally feel the data should be shared to avoid being scooped and to achieve their own goals, and their ability to ensure that the data shared is clean and reliable.

When are we supposed to put those data up? What if somebody wants to do something that we're actually already doing? How can it be quality control? So I am not at the point with an active grant that I really want to have people publishing off of it under the name of [redacted study] when I've never been able to check their numbers. .. I'm sure other studies that do a publicly accessible database had to deal with that. (experienced PI, epi)

So you know, I have mixed feelings about timing. But again, I also understand that to move science along, people bring different ideas, different approaches, different thinking. So if you make the data available, then other people can use their way. It is just horrible to be scooped, someone used your own data that you just killed yourself to collect. That is the problem.(experienced PI, epi)

The desire for individual level career advancement is tied to the institution's expectations for success which is part of the academic culture. The nature of the comments were generally related to the competitive culture of the scientific enterprise, with career advancement heavily dependent on peer reviewed publications, especially the desired to be the first to publish on their own data. Publication was a central theme echoed by both new and experienced investigators, and NIH staff. Both the epidemiologists and genomicists/genetic epidemiologists mentioned that they would want to be the first to analyze and publish their own data, and fear that if they don't their data may get scooped by others who could potentially analyze the data and publish on them. The following quote illustrates this sentiment.

Because really what you have to think about is the fact that all of the people who were involved in generating the data or paying for it or doing something with it, they all want to have the ability to go back and analyze the data and extract meaning. So, as scientists we're not rewarded for sharing; we're rewarded professionally for hoarding data and information. … If there's a dataset out there and I release it and if I generated a dataset, if my colleague down the hall takes that dataset and analyzes it and publishes it, he or she is going to get promoted before I will. Right? He or she might get a grant based on the analysis of this data that I could have gotten if I had done the analysis. So there are huge disincentives to data sharing.(experienced PI, gen/gen epi)

One interesting idea that was noted from the quote below has to do with how researchers are thinking about their data, and their perception of how that impacts their career. The notion of "control" of the data is something that is related to how they define or conceptualize data sharing. With epidemiology data, the investigators expressed preference for a collaborative model of sharing that would give them more control over how their data is interpreted or used.

we always want to get our studies analyzed first or published first before we share the data, so that's always one concern. And also just to make sure the data will be used correctly, because once it's out there, then we have no control of it. We just don't know how people are going to use it for. So ideally somebody will monitor this – I assume NIH – once we deposit all this epi data – it's like at NIH, someone's responsibility to make sure that that happens (new PI, epi)

The list of other co-occurring factors are found in Table XIII, Appendix F. This same factor, *Career concerns-Bar*, was observed in the document reviews (Appendix G).

Culture differences in research fields - Barrier

There are discipline or field specific differences in perceptions of data sharing, which may hinder or prevent sharing of research data in data repositories such as epidemiology compared to genomics. There are lots of data collected from large epidemiology studies that could be used to support trainees and junior investigators. With epidemiology data there are many variables from questionnaire data and epidemiology researchers are hesitant to share especially if they haven't yet analyzed all the data. There are so many research questions that can be asked and researchers fear that if put all data out before publication that others will scoop and publish before them. The respondents alluded to a negative perception of data sharing among researchers, highlighting the culture of research fields, specifically in epidemiology and genomics, as a barrier to sharing data. The factor that captures this concept is a priori factor,

Culture differences in research fields-Bar. There were 9 respondents (24%) who mentioned this factor as hindering data sharing.

TABLE XVI: NUMBER OF RESPONDENTS WHO MENTIONED *CULTURE DIFFERENCES IN RESEARCH FIELDS* AS A BARRIER

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	1	2	**3 (21%)**
Genomicists/Genetic Epidemiologists (11)	0	1	**1 (9%)**
Investigators (25)	**1 (17%)**	**3**	**4 (16%)**
NIH Staff (12)			**5 (42%)**
TOTAL (37)			**9 (24%)**

The example below illustrates differences between the way that epidemiologists and non-epidemiologists think about data as well as the complexity of epidemiology data, something that the investigators and staff agreed with. In the quote below, respondent mentioned that epidemiologists who collect data have the best understanding of the nuances of the variables and design and they fear that inaccurate analysis of the data will compromise the integrity of the data and the integrity of their career, with a ripple effect on their career progression. One staff corroborated this by indicating that this was part of the culture of epidemiology.

... I think there's a big difference. I think there are two groups of people. There are people who just analyze data, who have never collected data themselves. They're not epidemiologists. For them data is just data and you just crack through the data and analyze it. But there's more to it. And I think people who collect data, appreciate data in a completely different way… and there's a lot of intricacies to that process. And I think it's important that users of the data have a connection to the people who collected the data, who can explain the data, who can explain the design. And I think that's being lost by just being forced to upload data, epidemiology data to dbGaP. (experienced PI, epi)

Genomic / genetic data was considered to be more straightforward, compared to the complexity of epidemiology data which poses a problem to researchers in terms of complying with the data sharing policy.

.. *An epi study has tons of variables and, you know, is not funded to do one paper, and most has multiple papers, so main finding is pretty ambiguous. It's sort of — yeah, it [policy] reads like something that's not really tied to the real world of how epi studies are done. Takes a long time to do them. Once they're done usually you publish between five and ten years off of 'em because there's so much data in there, and the — of course you set it up to do it that way….. again, I think it's kind of a basic scientist view of release of data is just a spreadsheet of things that we're in, you know, Figure 1 of a paper, and generally these datasets are way more complicated than that.* (experienced PI, epi)

Lack of a reward system for data sharing - Barrier

The a priori factor, *lack of a reward system* for data sharing as a barrier *(Reward system-Bar*) was defined as when respondents mentioned that there were no systems or that they were not aware of systems in their institutions for rewarding, recognizing, crediting researchers for sharing data. Overall, there were 58 coding references for *Reward system-Bar*, with 26 of 37 respondents (70%) who mentioned that they were not aware of any rewards or incentives at the institutions for data sharing, and indicated that this was an opportunity for the institutions to create incentives such as the inclusion of data sharing in the promotion and tenure criteria, which currently is not the case. This would increase compliance with the data sharing policies, and also benefit researchers' careers and benefit the public.

TABLE XVII: NUMBER OF RESPONDENTS WHO MENTIONED *LACK OF REWARD SYSTEM AS* A BARRIER

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	3	8	**11 (79%)**
Genomicists/Genetic Epidemiologists (11)	2	8	**10 (91%)**
Investigators (25)	**5 (83%)**	**16 (84%)**	**21 (84%)**
NIH Staff (12)			**5 (42%)**
TOTAL (37)			**26 (70%)**

Although only one investigator, an experienced epidemiologist, indicated that their university was recently starting to consider data sharing in their promotion criteria, others indicated this was not the case at their institutions. A few experienced epidemiologists and genomicists/genetic epidemiologists mentioned explicitly that they had been on promotion committees and never heard data sharing being discussed during evaluation of their researchers, although there was one exception who mentioned their institution had recently integrated data sharing into their promotion and tenure process.

I just wrote a letter for promotion because I know they changed all the criteria. And it's interesting it wasn't at all listed. It's all about the PI and what they've accomplished or the investigator. There was nothing in the new criteria that addressed data sharing that I recall. (experienced PI, epi)

The respondents mentioned the lack of rewards, incentives or credit given to data sharers, as a big part of the problem in getting researchers to share their data, as indicated by the prevalence of the code across the data. Below is an example from a new investigator that

illustrates the absence of a reward system; also echoed by experienced epidemiologists, genetic epidemiologists, and NIH staff.

Sharing data does not benefit the data generator, in general. It's a good thing they do for the society. We don't have any rewards, it's a mandate. It's a responsibility. (new PI, gen/gen epi).

BARRIERS - Regulatory Policy and Law

Lack of clarity of policy - Barrier

The a priori factor, *Lack of clarity of policy* was one of the most prominent barriers to data sharing (*Lack of clarity of policy-Bar*) with a total of 97 coding references, mentioned by 33 of the 37 respondents (89%). This includes all the staff and epidemiologists and most of the genomicists/genetic epidemiologists. When the data was analyzed by the level of investigators, it showed most of the new and experienced investigators mentioned this factor as a barrier.

TABLE XVIII: NUMBER OF RESPONDENTS WHO MENTIONED *LACK OF CLARITY OF POLICY* AS A BARRIER

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	4	10	**14 (100%)**
Genomicists/Genetic Epidemiologists (11)	1	6	**7 (64%)**
Investigators (25)	**5 (83%)**	**16 (84%)**	**21 (84%)**
NIH Staff (12)	[redacted]		**12 (100%)**
TOTAL (37)			**33 (89%)**

This concept was most evident when respondents described specific aspects of the policy that they considered as influencing sharing of data by investigators in data repositories. The most

common themes included the vagueness of the data sharing policy and not knowing what was required, and the short timeline in the policy that they don't think allows enough time to clean and analyze datasets before it gets shared in the data repository

The following quote illustrates the concerns with the vagueness of the policy in terms of expectations and the ambiguity of what "main findings" or "timely release" or "final dataset" mean, given the different types of data. Some of these data, like epidemiology data have lots of variables with main findings ending up in more than one paper. This was echoed consistently across the different groups of respondents – the investigators and NIH staff.

At the same time, I think it's broad enough saying that the analytical data set has been finalized, maybe even less so in our genetic data, but our epi data – it seems like it's never finalized. You do a different analysis and you realize like, oh, there's missing data here. Oh, that's a strange skip pattern, or, oh, you know – there's constant data cleaning. So you could argue, if one would choose to – you could argue that your data's never finalized until you're retiring. But overall, I think it's very necessary. It's very reasonable to have these. But I think that it should be kind of a work in progress, and there may not be a one-size-fits-all for every situation. (new PI, gen/gen. epi)

While respondents generally agreed to the importance of a timeline as a guideline for when data should be shared, they agreed that the timeline for sharing data was too short and could vary depending on the type of data, especially the 2003 data sharing policy. Again, emphasizing that there's not a one-size fits all implying that there shouldn't be a single policy for the different data types. The quote below from a staff also describes the rationale for why the policy was developed that way, from the perspective of a federal policy maker.

... almost all policies across the government are written very broad, or kind of lofty, and it's intentional. And that is because we don't want to come back and in a year from now, have to construct new policy, because new studies have come out, or a new scientific discipline has come into scope, whatever it is ... So when we drafted the timelines for these, it was based on, at that time, community standards and what was acceptable. So it's easy for the 2015, for us to come up with somewhat of a timeline, because as I mentioned, it's only one type of data… So we have a general understanding in the genomic scale about how long after you finish producing data that you would ultimately be able to release it once it's cleaned and whatnot. So I think the problem is, which we also ran into with the 2003 policy, when you get beyond a single type of data, and

even there were issues in genomic data in the 2015 policy; not everything sits in a nice little box like that. So there's obviously projects, particularly longitudinal genomic projects or consortium projects, where it's very hard for them to meet certain timelines because their study just isn't built up that way. So I think that's an issue where, when a policy states a particular expectation for timeline (staff)

This example is one that acknowledges not just the challenges with having a single policy like the 2003 policy for different types of data / study designs, although there were also issues with the GDS policy, but also mentions the effort that has been made to address this issue to some extent through the addition of supplemental or appendix to the policy to further explain timelines for different types of data. It's important to determine if researchers are aware of this and considered this in their earlier responses or if this is an opportunity to increase awareness and dialogue with researchers on development or amendment of timelines especially for longitudinal studies. This could potentially have a positive impact not just on data sharing but also on how policies are written across the government – understanding that there's not a 'one-size fits all'.

Another related point considered important by the respondents was the inconsistent implementation of the data sharing policies across NIH. This is related to the factor, *Inconsistent enforcement of policy* as a barrier. A big part of how well researchers comply with data sharing policies has to do with communication at different levels, both internally and externally. There's variation across NIH in how policy is interpreted, implemented and what the expectations are for how NIH staff are to implement the policy was of concern to both new and experienced epidemiologists and genomicists/genetic epidemiologists.

I think the rule about posting up genomic data in dbGaP, I think that's a good rule. I think it works okay. I think the other policies that, I think there's just a lot of ambiguity about what's coming and what's required of us and different institutes within NIH have different rules. Like, we have a study with the National Institute of Mental Health and they have a different approach from, as far as I can see, from NCI, as far as what they require us to do. So I think

there's some non-uniformity and ambiguity and confusion amongst investigators. (experienced PI, epi)

Inconsistent enforcement of policy - Barrier

The a priori factor, inconsistent policy enforcement as a barrier to data sharing (*Inconsistent enforcement of policy-Bar*) was defined as when respondents discussed the lack of adequate or inconsistent enforcement of data sharing policy among researchers. This factor was mentioned by 17 of the 37 respondents (46%).

TABLE XIX: NUMBER OF RESPONDENTS WHO MENTIONED *INCONSISTENT ENFORCEMENT OF POLICY* AS A BARRIER

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	0	2	**2 (14%)**
Genomicists/Genetic Epidemiologists (11)	2	4	**6 (55%)**
Investigators (25)	**2 (33%)**	**6 (32%)**	**8 (32%)**
NIH Staff (12)	[redacted]		**9 (75%)**
TOTAL (37)			**17 (46%)**

The perspective of one staff as it relates to the enforcement of policy as depicted in the quote below confirms the absence of a uniform approach to data sharing across the institute, indicating an opportunity for standardization in the requirements. There are ongoing efforts at EGRP / NCI to standardize data sharing plans.

So I would start by saying that NIH in some ways does not have a universal unified approach to data sharing. In some ways the closest we have to that is our 2003 data sharing policy, which, you know, doesn't cover, you know, kind of the whole universe of NIH data, but certainly is fairly broad in its requirement for people to provide a plan for data management and sharing. So we have kind of a multiplicity of approaches when it comes to data sharing. (staff)

In addition, the perspective of a geneticist/genomic-epidemiologist in the quote below, further points out the challenges with current enforcement strategies or mechanisms by NIH.

… And then I think it's also very difficult to enforce in a way. I mean it's just kind of an ongoing battle and I've seen this type of problem happen in the past, … there's no way for the NIH to actually control that with the exception of large U19s or mechanisms, funding mechanisms by which the NIH administration or the NCI tracks the development of those grants, right? So these are mechanisms that are different from regular ones or regular projects that basically the NIH has no control. … I think what's more vexing is the fact that everybody knows the names of the investigators that basically never share data, never have and never will. And still when you read their grants they will just have that blurb about that they will share the data. How can you do it, you know. It's one of these things; what would you do? Would you now have an extra round of reviews to rate that investigator's compliance with data sharing policy? I don't know. Because also I would be very concerned if at some point people would do that with me because in many cases I would have my hands tied by somebody else or a consortium and then an investigator could actually be unfairly treated in a sense, "oh, this guy's not sharing data", even though he or she is trying their best. (experienced gen/gen. epi)

This quote highlights the negative impact that the perceived lack of control or limited and inconsistent enforcement of policy on all investigators could have on the perception of data sharing among researchers. It presents a good illustration of the relationship between individual, interpersonal and organizational level factors that are integral in enhancing data sharing practices among researchers. If some investigators are not held accountable, there is less motivation for other investigators to share their data.

The quote below illustrates divergence of thoughts from previous comments by investigators and other NIH staff around enforcement of the data sharing policies. This staff indicated the absence of enforcement, whereas others alluded to the inconsistent enforcement of policy. This further reinforces the need for NIH to evaluate current strategies and approaches for enforcement of data sharing.

They [NIH] have no enforcement. They haven't required as I suggested grantees to do certain things. There's no real capacity to hold people responsible to that, so the policy is rather

hollow in my mind. It's noble, but hollow. It's one of those things where the perfection – in a perfect world, that would be great, but it's not a perfect world that we live in. (staff)

These examples indicate an opportunity to revisit the way the policy is written by making the expectations and implementation processes more clear and consistent using strategies that are more effective than what is currently in place e.g. more education and better and targeted communication with key stakeholders like the researchers who are required to comply with the policies. This will be helpful for staff and researchers although may not be an easy thing to do.

There was one interesting comment from an experienced genomicist/genetic epidemiologist who mentioned not knowing how enforcement of the policy would occur when a grant ends. NIH has not thought through the impact of the regulations which includes financial impact (e.g. funding to hire people to upload data) and sustainability beyond life of the grant and maintenance (what happens after the grant has ended / maintenance of the expired grant)

I don't know anyone who has not complied. I feel like I know that we have been discussing some of our data sharing language that is on our website as part of our funding through our infrastructure grant, so I do feel like there is some sort of financial tie, but I don't really know what happens – once your grant is over and you've already been funded, I don't know honestly what kind of enforcement could happen afterward.(experienced gen/gen. epi)

This highlights either a gap in knowledge or communication or awareness of the data sharing process from beginning to end, clarity of the process in term of consequences or implications for not sharing, what the implications are for researchers, the institutions or NIH, and clarity in how the policy applies to grants in different stages especially the end stages. Of consideration is what efforts are currently in place or opportunities for researchers to share data in data repositories beyond the lifecycle of the grant. NIH has not thought through the impact of the regulations which includes financial impact (e.g. funding to hire people to upload data) and

sustainability beyond life of the grant and maintenance (what happens after the grant has ended / maintenance of the expired grant).

While the example below provides some insight on a strategy for getting data submitted before the grant ends, it is not part of the NIH data sharing policy but seems like an ideal or logical proposition to ensure the data from NIH funded data is deposited and shared. This doesn't account for the complexity of epidemiology studies whose research aims may encompass other related variables not already cleaned or analyzed.

... but if we're funding, I think basically they need to share all their data that's related to their aims whether it's published or not because sometimes you can't actually get things published within the grant period even maybe. I hadn't really thought about this before now, but I think that it should be the data that we use to address the aims, so by the end of the grant, if you haven't published some, by the end of the grant period you need to share the data that are being used to address the aims. (staff)

Here's another interesting observation on the relationship between organizational culture and policy enforcement. The culture of NIH was mentioned as a barrier to data sharing, and therefore has a negative impact on how staff are enforcing the data sharing policy. This challenges NIH or suggests clearer and more specific policies that are realistic, with a clear plan for how it will be implemented on the NIH end accounting for what it will take on the researcher end to comply e.g. funding, as well as identify how to create a system that is effective and functional. This also speaks to the leadership impact and leadership role in policy development and implementation. NIH could make bold moves in articulating changes to policies / mandates to ensure that they are implemented and researchers comply. For example NIH could mandate that a portion of a grant award (provide guideline / recommendation of level or how much) be allocated towards data sharing or require justification for not complying.

So I think it's partly a culture, and I blame the NIH for not taking the bold moves of the more long-term vision instead of the declaration: "Make it all available in 24 hours." That sounds great, but it's sort of like politicians. .....Those are easy things to declare, but you have to make tough decisions to start moving in those directions. It's a cultural thing – if NIH said, "This is so important that we're going to now require that all grants going out in 2018 – put aside one percent or two percent for data sharing." – That that's required, instead of saying, "We have this policy" because how do you pay for it? You're gonna have to make sure that they have the money to pay for it.(Staff)

In describing their experiences with enforcing the data sharing policy as NIH staff, the respondents mentioned the gap in knowledge of the policy implementation processes and what all the requirements and expectations are. This provides some insight into the complexity of the data sharing policy, process and implementation, and on the need for more training for staff to help them better enforce the policies as they deal directly with the grantees / investigators.

I think some of even just understanding; in the past, understanding the implementation. It's taken me some time to really understand how we're implementing and I feel like it wasn't necessarily clear what the timelines were to me, and what documents are needed and what my role was. So I think it's a complex policy, and the implementation's complex, and I think I felt like it took up some time for me to learn what I needed to do, was a challenge.(staff)

This factor, *Inconsistent enforcement of policy-Bar,* was also observed in the document reviews (Appendix G).

Privacy concerns - Barrier

The a priori code, *Privacy concerns* as a barrier to data sharing (P*rivacy concerns-Bar*) was defined as when respondents described concerns about potential risk and violation of participant confidentiality in a research study, as well as issues related to consent and requirements by Institutional Review Boards (IRB) which is set up to fully protect the privacy of human study participants. This factor was mentioned by 21 of the 37 respondents (57%).

TABLE XX: NUMBER OF RESPONDENTS WHO MENTIONED *PRIVACY CONCERNS* AS A BARRIER

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	3	4	**7 (50%)**
Genomicists/Genetic Epidemiologists (11)	2	6	**8 (73%)**
Investigators (25)	**5 (83%)**	**10 (53%)**	**15 (60%)**
NIH Staff (12)			**6 (50%)**
TOTAL (37)			**21 (57%)**

More than half of the respondents identified this factor to be an important barrier to data sharing given their concern with potential violation of confidentiality with their study participants, which could end up costing them and their institution a lot. Challenging is when request for data is limited by the data use limitations as specified in the informed consent. One staff mentioned that it could also depend on the way the IRB interprets the consent; their goal is to protect the privacy of the study participants, honoring their consent, as well as to avoid liabilities. However, this may pose a barrier to others being able to access specific data sets that fall outside of the data use limitation and use the data for additional analysis. Respondents agreed to the protection of research participant data and two new investigators – epidemiologist and genomicist/genetic epidemiologist expressed concern about how well study participants understand what they are consenting to in term of sharing their data.

... whether it was truly consented to be so restrictive or whether it's just that an IRB interpreted a consent to say that, we can only give access based on what the investigators say as to how the data can be used. And so I think what is often frustrating for people is they have an idea in mind of a project that they want to do, and there is a cohort that would be perfect. However, the data use that's acceptable on file does not jive with what they want to do. So

they're never going to get access for what they want to do unless they go back to those investigators and insist that groups be reconsented and we get a different kind of data use in place. … (Staff)

The NIH data sharing policy states that the sharing of data shall be in alignment with what's in the consent forms. The challenges and concerns expressed from the respondents indicates an opportunity for training / increased awareness and consistent communication and alignment of data sharing priorities and implementation within NIH and between NIH and academic institutions.

.. so that's a tension - where I am, you feel there's a left hand and a right hand at NIH. The right hand wants you to share data as much as possible, and then the left hand, the regulatory environment, says, "Wait a minute. Do you know where every piece of data's going? Do the participants, have they consented? Do they really understand? Has it got the appropriate protections?" And sometimes it doesn't feel like left hand and right hand are talking to each other because you sort of get conflicting things that you have to manage, but of course everybody wants to get the work done, so they sort of agree that it's a little schizophrenic at times. …'Cause I'm trying to comply with what NIH wants for data sharing, but I gotta comply with what my institution is being told by HHS in terms of data privacy, consent, data reuse, that sort of thing. (experienced PI, gen/gen. epi).

This factor, *Privacy concerns-Bar*, was also observed in the document reviews (Appendix G).

Definition of data sharing – Barrier (EMERGENT)

During the data analysis phase, a new code *Definition of data sharing* emerged as a barrier to data sharing, after discussion and reflection on why the investigators, mostly epidemiologist, showed preference for the collaborative model of data sharing. There were 108 references to the emergent code, *Definition of data sharing*, mentioned by all 37 respondents in the study who were asked to define or describe the concept of data sharing. The data was further analyzed and grouped into three main themes that described the different types of data sharing:

collaborative data sharing or sharing in consortia (36 references from 15 respondents (41%)); sharing data in a data repository or database (24 references from 13 respondents (35%); and sharing to advance science or for the good of the public (36 references from 25 respondents (68%)).

Overall more respondents defined data sharing to be sharing through collaboration or consortia, with the epidemiologists mentioning both collaboration and data repository more than the genomicists/genetic epidemiologists. More experienced investigators than new investigators mentioned data sharing through collaboration.

TABLE XXI: NUMBER OF RESPONDENTS WHO MENTIONED *DEFINTION OF DATA SHARING* (COLLABORATION) AS A BARRIER

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	2	7	**9 (64%)**
Genomicists/Genetic Epidemiologists (11)	0	4	**4 (36%)**
Investigators (25)	**2 (33%)**	**11 (58%)**	**13 (52%)**
NIH Staff (12)	[redacted]		**2 (17%)**
TOTAL (37)			**15 (41%)**

The collaborative model of sharing provides the investigator with more control over how the data is used and interpreted. This entails direct contact and involvement between the data requestor and data originator prior to access. This could result in joint publications but the emphasis by the investigators in this study was that they felt more comfortable that their data would not be misinterpreted or misused (related to the factor *Career concerns*).

This factor, *Definition of data sharing-Bar*, was also observed in the document reviews (Appendix G).

Of the 13 respondents who mentioned sharing in a data repository, most of them were investigators, with about half of the new investigators and experienced investigators, and mostly epidemiologists.

TABLE XXII: NUMBER OF RESPONDENTS WHO MENTIONED *DEFINTION OF DATA SHARING (*REPOSITORY SHARING) AS A BARRIER

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	3	5	**8 (57%)**
Genomicists/Genetic Epidemiologists (11)	0	4	**4 (36%)**
Investigators (25)	**3 (50%)**	**9 (47%)**	**12 (48%)**
NIH Staff (12)			**1 (8%)**
TOTAL (37)			**13 (35%)**

Data sharing defined as sharing to advance science or for the good of the science was mentioned by 25 respondents, half of the investigators mentioned this definition with overall more new investigators than experienced investigators, and more genomicists/genetic epidemiologists compared to epidemiologists. All staff interviewed mentioned this definition of data sharing.

TABLE XXIII: NUMBER OF RESPONDENTS WHO MENTIONED *DEFINTION OF DATA SHARING* (FOR ADVANCEMENT OF SCIENCE) AS A BARRIER

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	2	4	**6 (43%)**
Genomicists/Genetic Epidemiologists (11)	2	5	**7 (64%)**
Investigators (25)	**4 (67%)**	**9 (47%)**	**13 (52%)**
NIH Staff (12)			**12 (100%)**
TOTAL (37)			**25 (68%)**

Well, it would be - making the data from a particular study available to a broader research community so like the results from a particular study can be verified to ensure reproducible research. But also to give folks the opportunity to apply new methods, looking at it in a different way, maybe answering questions that weren't part of the original hypothesis. (experienced PI, gen/gen. epi)

There was divergent perspective in this type of data sharing described by the respondents which includes making all data including raw data available in a public or controlled-access repository while protecting the privacy of subjects contributing the data. The two comments on making raw data available as seen in the quotes below, came from experienced genomicists/genetic epidemiologists.

I think to me it's really making available all the primary data that is used in publication and in many cases use of the raw data would be better actually to make it favorable. (experienced PI, gen/gen epi)

The raw data is, it can be very useful. Often we deal with very highly processed data. (experienced PI, gen/gen epi)

This view is countered by one staff who advocated against sharing of raw data that wasn't linked to publication.

... because lousy data in, aka crap in, brings crap out. And I think that it's really important that the community have some understanding of what the nature of the data is and how

it's been handled, and its metadata, and failure to do that with rare exception. And there are people that could come and say, "I'd like the raw data because I want to do these things" and then there's a conversation that hopefully they'll be able to understand, reconstruct, or have a different way of analyzing, but the widely available use of data without QC issues and without tagging it to publications is challenging to me, and I don't think that that's a successful experiment. (staff)

The tension between wanting to share data in a way that assures control (i.e. through the collaborative model) and for the good of the science, whether that meets the NIH data sharing policy requirement is duly noted as important in existing data sharing practices at institutions, as well as the careers of investigators. When asked what their institution's norms or culture around data sharing was, one new investigator responded in the quote below:

So I would say that we are very open to the collaborative model of data sharing and we encourage it and we use it. I would say that's the culture here is positive towards the collaborative model of data sharing. And I also feel that there is a definite, like negative attitude towards public or controlled-access data sharing. I mean, definitely a lot of concerns. And some of the ones that I've voiced, I think that there's also some unvoiced concerns that are not, I'm not sure what they are, but I just sort of get this sense, the culture sense, that we don't want to do it. (new PI, epi)

This example provided some insight into the attitudes of investigators towards the sharing of data in databases or data repositories as mandated by the NIH data sharing policy and the norms of their institution around data sharing. It implies a key factor that could potentially influence investigators is the institution's culture / practices, as well as the concern for career advancement. This is corroborated by another new investigator, genomicist / genetic epidemiologist who alluded to the influence of culture norms on attitudes among investigators.

... would say my more senior investigators are less enthusiastic about it, and in a couple of cases may actually look for loopholes to get around or to minimize the amount of data that they have to share. But I would say kind of my more mid-career generation of people – we grew up in this world. It was 2003 I got my – when the first data-sharing policy came out. I got my Ph.D. in 2004. This has just kind of been part of my norm, always. (new PI, gen/gen. epi)

The researchers in this study were also asked about the culture of their institutions around data sharing and 17 of 25 respondents (68%) mentioned that their institution including the investigators were very collaborative and open to sharing data within their department or institution. This overlaps with the data from respondents who mentioned collaborative model of data sharing in their definition of data sharing.

TABLE XXIV: NUMBER OF RESPONDENTS (INVESTIGATORS) WHO MENTIONED INSTITUTIONS HAVE COLLABORATIVE CULTURE

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	2	6	**8 (57%)**
Genomicists/Genetic Epidemiologists (11)	2	7	**9 (82%)**
Investigators (25)	**4 (66%)**	**13 (68%)**	**17 (68%)**
NIH Staff (12)			**0 (0%)**
TOTAL (37)			**17 (46%)**

RQ 1c, 1d: BARRIERS - Resources

Administrative / Technical resources - Barriers

The a priori factor, *Administrative/ technical resources* as a barrier *(Inadequate Admin/tech resources-Bar),* is defined as when respondents described the lack of or inadequate administrative resources including time and effort of personnel and technical support. 30 of the 37 respondents (81%) mentioned *Administrative / technical resources* as a barrier to data sharing.

TABLE XXV: NUMBER OF RESPONDENTS WHO MENTIONED *INADEQUATE ADMINISTRATIVE / TECHNICAL RESOURCES* AS A BARRIER

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	2	9	**11 (79%)**
Genomicists/Genetic Epidemiologists (11)	1	9	**10 (91%)**
Investigators (25)	**3 (50%)**	**18 (95%)**	**21 (84%)**
NIH Staff (12)	[redacted]		**9 (75%)**
TOTAL (37)			**30 (81%)**

Sharing data is time intensive and requires resources. According to the participants, data sharing poses huge time constraints on the administrative and technical ends because it takes a lot of time and effort for programmers and analysts, working in conjunction with the researchers, to prepare the data files for submission in the data repository, ensure the data files are in the right format, manage the data requests from secondary users, properly document and annotate the data so that it is useful for reuse. Participants indicated that data submission and access in dbGaP are time consuming on both the investigator's end and also on the dbGaP management end; there's currently a backlog in managing the requests, as noted by an experienced genomicist/genetic epidemiologist. This also includes time and effort spent on data transfer agreements and consent forms. All the respondents in this study shared the same perspective about this relationship, including that the main issue with the data sharing process is not the data submission but rather the preparation of the data so the data / metadata is meaningful to a secondary user.

It's expensive to share the data. There's a real programming effort involved and I mean, steps to take, because you know, internal documentation is not the same as what's needed for external use. [experienced PI, epi]

The interesting aspect of this quote above, echoed by new and experienced epi and genomicists/genetic epidemiologists, and NIH staff, is that the respondent made a distinction in the level of effort dedicated to documentation of data based on whether the data was for internal use or if the data was for public use via publicly in controlled-access repositories. It speaks to the underlying issue and relationship with limited time and financial constraints. More importantly, for the researcher indicating that documentation for external use (i.e. sharing via repository) is a burden means that having a clean and well-annotated datasets with good documentation at the outset of the data collection phase may not always be the norm in research teams. This point was corroborated by one of the staff respondents. This is an opportunity for a culture shift in thinking.

I mean, you know, it's always just time and money. You know, who is available to do the deposit, you know, keeping track of when it needs to be done, you know, that it's just making sure that there's somebody familiar with the research who's able to do the work and knows how to do it and things like that. And sometimes that's a challenge if the grant has ended and you don't have the people employed anymore who worked on the data. You know, so there are issues like that, just having somebody who knows how to do it available at the right time, but I don't know how big of an issue that is, but certainly it's a limitation if the grant has ended, the person who does work with the data isn't there anymore or whatever, there's some issues like that, I think, that might arise. (experienced PI, gen/gen. epi)

So I think some of the biggest barriers to people who want to share are the fact that there are requirements for any database as far as the format that a data file needs to be presented in. There is metadata surrounding both the data files themselves and the patient that can be a hindrance. And so if a group, for whatever reason, has not generated data in that particular format, I think it can be a lot of work, and if you're not used to it, there's a large learning curve. It's not something that is easy to just look up on a YouTube tutorial. (staff)

This presents an opportunity for training and education on the requirements of data repositories, developing criteria to help with the standardization of data formats and to provide support for data formatting, to make it easier for investigators to submit their data. Most of the respondents interviewed did not have direct experience with the actual data preparation and upload or access of data from the data repository but may have overseen these activities through

their research team meetings. The investigators mentioned that they relied heavily on their analysts and programmers for these tasks and for them to convey any administrative or technical challenges experienced with the process or with the datasets or with the policy requirements.

As shown in TABLE XXV above, many of the respondent perceived the entire data sharing process as a huge administrative burden for the investigators and their research teams. They describe it as the more time spent on preparing data for submission was less time they had to work on their research, analyze their data and publish. As the main goal of academic researchers is career advancement which is measured by the number and quality of peer reviewed publications, many indicated their preference to focus their time and resources on doing research and publishing, instead of spending time and limited resources preparing datasets for others to use. This highlights the relationship between inadequate *Admin/tech resources-Bar* and *Career concerns-Bar*, as illustrated in the quote below by a genomicist/genetic epidemiologist. It also sheds light on the need for a reward structure that would credit or recognize researchers who share data, as part of their academic career evaluation.

It took us forever to get the study put together, the questionnaires, and then figure out recruitment and to accrue enough cases. We have critical mass now, but, I mean, it's so much blood, sweat, and tears. And then, for people who haven't done field work to just be able then to take the epi data and the genomic data and to analyze it, a bunch of us have talked about this. You know, we're kind of uncomfortable about it. (experienced PI, epi)

Corroborating this evidence is a comment by staff that highlighted the importance of proper data documentation for both internal and external users at the onset of data collection. This is an indication that there are varying levels of documentation by the investigators across academic institutions and that it will need a change in behavior for them to document early with the intent for deposit in a repository.

This is related to changing the culture of data sharing at institutions where it can be encouraged, rewarded and incentivized so that researchers can be more intentional about how they prepare their analytic datasets and variables. It will require some investment in resources. This provides an opportunity to encourage researchers to create a data dictionary early in the research phase, ideally before data collection with proper description, documentation and annotation of their data / variables. This will help with easy access and understanding of the data by data users.

This same factor, *Inadequate Administrative/technical resources-Bar*, was observed in the document reviews (Appendix G).

Financial resources - Barrier

A priori factor, *Inadequate Financial resources* as a barrier *(Inadequate Financial resources-Bar),* is defined as inadequate or lack of financial resources to support data sharing processes, including cost or funding for infrastructure and personnel. Almost 70% (25 of 37) of the respondents identified financial burden as a barrier to data sharing.

TABLE XXVI: NUMBER OF RESPONDENTS WHO MENTIONED *FINANCIAL RESOURCES* AS A BARRIER

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	2	6	**8 (57%)**
Genomicists/Genetic Epidemiologists (11)	1	8	**9 (82%)**
Investigators (25)	**3 (50%)**	**14 (74%)**	**17 (68%)**
NIH Staff (12)	[redacted]		**8 (67%)**
TOTAL (37)			**25 (68%)**

This factor *Inadequate Financial resources-Bar* was closely related to the factor, *Inadequate Administrative/technical resources* as a barrier. Staff mentioned that some smaller NIH institutes and centers lack sufficient resources to provide staffing support to assist the submitters one-on-one, compared to the larger NIH institutes and centers. Another comment from the staff perspective had to do with insufficient resources and investment in the biomedical engineering aspect of the technology at NIH e.g. dbGaP. This is an opportunity for NIH to evaluate the financial priorities and investments across the different key stakeholders, within and outside NIH, to help inform proper and adequate allocation or investment of resources to ultimately support / enhance data sharing practices among investigators.

NIH doesn't have a way to invest in engineering standards because we're not set up as an agency to do that. We don't have a big biomedical engineering contractor to hire. (staff)

It was clear from the comments from different investigators that some were aware that the data sharing policy allows data sharing to be included in the grant budget, while some weren't aware of this. The following quotes highlight this divergence in perspective which is due to lack of awareness and communication. They still perceived data sharing to be a major financial burden, especially with the standard NIH programmatic policy cuts applied to grant awards which reduces the overall budget of the grant regardless.

... the part about unfunded mandates and getting it done, that's a big deal because you're being asked to do things that may or may not be written into the grant. ...when you're talking about very, very tight budgets and trying to just get done what you can get done science-wise, putting into the budget the data sharing piece is hard because you're — you have a very small budget, and it's cut anyway, and then a mandate to do this data sharing, which either is not funded or takes away from some of the science you want to do. That's not a great incentive for investigators... Now you could argue that that is part of the science, but, you know, it's still to the investigator for their individual grant a burden. It's a financial burden or any effort burden. (experienced PI, gen/gen epi)

... you can't have an unfunded mandate of sharing when you can't put into grants the necessary resources to be able to make those data and the documentation. (experienced PI, gen/gen epi)

The respondents – new and experienced epidemiologists and genomicists/genetic epidemiologists indicated that funding resources are required for updating the data repositories to facilitate easier and more efficient submission and access to data. More on this will be discussed under the TECHNOLOGICAL INFRASTRUCUTRE construct.

It's a bit of a process. It's not just clicking on one button ... It is an online system at dbGaP so it's not like you're sending emails to people. The system is a bit antiquated and that really comes down to there's just not a lot of funding for it. So, if there was more funding, it could be more modern and sleeker.(staff)

This factor, *Inadequate Financial resources-Bar,* was also observed in the document reviews (Appendix G).

<u>Leadership support – Barrier</u>

The a priori factor, *Leadership support* as a barrier (*Leadership support-Bar*) describes the lack of support from the institution's / organization's leadership in terms of administrative, technical and financial resources, as well as oversight on the data sharing process. This factor was mentioned by 14 of 37 respondents.

TABLE XXVII: NUMBER OF RESPONDENTS WHO MENTIONED *LEADERSHIP SUPPORT* AS A BARRIER

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	0	3	**3 (21%)**
Genomicists/Genetic Epidemiologists (11)	1	6	**7 (64%)**
Investigators (25)	**1 (17%)**	**9 (47%)**	**10 (40%)**
NIH Staff (12)			**4 (33%)**
TOTAL (37)			**14 (38%)**

The analysis of the data showed that although the prevalence of this factor across the data was much lower than some of the other factors, the level of influence on the different aspects of data sharing was determined to be cross-cutting and powerful in enhancing data sharing. The perceived lack of appreciation by NIH of the time and resource intensive nature of data sharing was mentioned in the quote below by an experienced genomicist/genetic epidemiologist, as a negative leadership contribution to enhancement of data sharing. This hints at the opportunity for the NIH to foster deeper engagement and collaboration with the extramural community as they develop policies, as a way to bridge both cultures.

I think it's part of this is a disconnect in terms of the cultures. Many of the people at NIH don't understand, have been at NIH too long. They don't understand the realities of the extramural community, research community and the demands that are placed on them within their institutions and so on in terms of financial needs and support and so forth. …. I don't think the people at NIH recognize the reality of what investigators are dealing with. (experienced, gen/gen epi)

Other than limited provision of resources, one of the key challenges with data sharing at institutions mentioned by an investigator in the quote below, is the perception that data sharing is not a high priority for the institutions and they therefore may not be willing to invest their institution's resources e.g. administrative or financial support through internal departmental or

operational funds, to facilitate data sharing. Prioritizing data sharing is something that is clear at the NIH level as the funding agency but not as clear at academic institutions as illustrated in the quote below. It would require them to invest in technological resources and infrastructure to promote sharing of data in data repositories. This is contrary to the leadership role of NIH and the current administration as being supportive of data sharing in research funded by the NIH.

... and so the other piece of that, of course, would be if the institution wanted to invest in creating those data repositories and things like that, but my institution, I don't think would do that. You know, like, invest money in facilitating data sharing for individual investigators. (experienced PI, gen/gen epi)

Related to institutional barriers, which often are attributed to the leadership of the institution, one staff respondent mentioned that the conflict of interest and the issue around control of data within the institutions can prevent investigators from sharing their data even if they wanted to because the institution would not be supportive. This has to do with the perception of respondents that the institution could be too conservative in the interpretation of the consent form. This is related to caution on the institution's end to avoid legal issues and liability, but could result in more restrictive requirements for the consent forms, for example, by the IRB that may limit how the data is shared. The other has to do with conflict of interest which plays a key role in hindering data sharing as illustrated in the quote below.

... IRBs that often also have to take into account the representation of medical or scientific directors or CEOs/COOs/CIOs; they will often report things like, we prefer the data only be shared with not-for-profit use. And in my experience, that very rarely is actually written into informed consent and our certifications for data sharing are supposed to be based on consent language. But at the same time, the IRBs will take it a step further for an institute and say, "This is what we prefer for our institute," because they want to protect their own institute's interest. So it's so many conflicting interests. There are few levels at which you could be barred from sharing that may not actually be the investigator

Training - Barrier

The a priori code, *Training* as a barrier to data sharing (*Training-Bar*) was described as when respondent referred to the lack of training for researchers and staff to increase understanding of policy requirements and expectations for sharing data, build or enhance skills for the different aspects of data access, submission and data management, in order to produce meaningful datasets that can be reused. This also includes the lack of education, teaching or training tools, materials used to improve data sharing. This factor, *Training-Bar* was mentioned by 4 respondents, all of whom are NIH staff.

These quote below indicates that while guides and materials may be available to researchers and staff, they may not be aware of the existence of such tools or guides, and the materials may not be clear. This therefore warrants training around available or existing NIH resources and materials.

You know, there is a lot of how-to guides and videos, but, I don't know how much people read them. Or if they know about them. But definitely, I think some of them have quite a bit of trouble, particularly in a first go around. It's not totally intuitive to them. (staff)

On the other hand, there's training at the institution level for the administrative officials for data access and data submission processes, which are critical to the data sharing process and have huge implications. This could be an opportunity for a joint training or innovative mechanism to address training needs at both ends

RQ 1c, 1d: BARRIERS – Technological Infrastructure

Analytic data complexity - Barrier

The a priori factor *Analytic data complexity* as a barrier *(Analytic data complexity-Bar)* was mentioned by 21 of the 37 (57%) respondents as hindering data sharing. This was used

when respondents referred to analytic data format, standardization, and metadata documentation / annotation of datasets as hindering data sharing.

TABLE XXVIII: NUMBER OF RESPONDENTS WHO MENTIONED *ANALYTIC DATA COMPLEXITY* AS A BARRIER

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total (25)
Epidemiologists (14)	2	5	**7 (50%)**
Genomicists/Genetic Epidemiologists (11)	1	6	**7 (64%)**
Investigators (25)	**3 (50%)**	**11 (58%)**	**14 (56%)**
NIH Staff (12)			**7 (58%)**
TOTAL (37)			**21 (57%)**

Respondents described certain aspects of the capability, capacity, effectiveness and efficiency of technological infrastructure as hindering the sharing in data repositories. From the NIH staff perspective, not understanding what the submission requirements are for the database, the process and the proper data format was considered a challenge for the researchers. This view was echoed by researchers too. The examples below highlight the complexity of the data submission process, given the complexity of the data from large scale studies, as well as the different roles of the different individuals or entities involved in the process. This indicates an opportunity for clearly and well-defined processes and training.

So ease of upload would be helpful and help with formatting. So I think some of the biggest barriers to people who want to share are the fact that there are requirements for any database as far as the format that a data file needs to be presented in. There is metadata surrounding both the data files themselves and the patient that can be a hindrance. And so if a group, for whatever reason, has not generated data in that particular format, I think it can be a lot of work, and if you're not used to it, there's a large learning curve. (staff)

And I will say, it has - in trying to work through this process it's been rather difficult. I think the people at dbGaP try, but sometimes it's a real challenge to get data into the system too. And if you talk to people who submit data to dbGaP, it can be a real challenge. So, to try to facilitate

that with large scale projects now, NIH typically mandates some kind of data coordinating center. Part of their responsibility is to ensure that the data is actually formatted appropriately and submitted to dbGaP. (experienced PI, genetic/genetic epidemiologist)

One of the interesting comments from some of the respondents was concerns with the use of data that is deposited in a repository in the absence of collaborations with the data generators who can assist with understanding the nuances of the data and the study design. Part of this has to do with the quality of the data deposited (i.e. *clarity of the analytic data*), the ability of the secondary user to understand the data accessed, and the knowledge of the submission process (*Expertise*) which has specific data requirements in order to achieve optimal use. This is also related to the fear of loss of control which was a recurring theme around the career concerns of researchers (C*areer concerns-Bar*) as well as the differences in data sharing definition (*Data sharing definition -Bar*E)

My concern is just that data is not just data. When you just upload epidemiology data from a complex project like the [redacted] people could just use it, you know, it's not clear to me that investigators are going to understand what the data mean, how they were collected, what the study design was. And we actually had a very nice model that we observed in the last X years in the [redacted] whereby somebody would submit a proposal of what they wanted to do. It would be discussed by the PIs. In the olden days it was an independent advisory committee that reviewed all the requests to access the data or biospecimens. Then internally we always assigned a liaison who would work closely with this investigator, you know, to answer questions, to explain what the data mean. And I think this is now being lost by us having to just submit, upload the epidemiology data on this public database that anybody can use. I think that's really problematic, because I mean there will be no control whether people understand, investigators understand data correctly, whether they understand the design correctly. (experienced PI, epi)

There was a strong relationship between administrative/technical resources in how the respondents described the complexity of analytic data such as phenotypic data, as hindering data sharing, as well as the huge amount of time and effort on investigators and their research teams,

required for proper documentation of large datasets. However, they recognize that the process results in a more meaningful and accurate dataset for secondary use.

I think in the past, even as it's written now that the deposit is not that difficult, but I hear some discussion to add more phenotype information, and that becomes really burdensome for the investigators, and particularly some phenotype [unintelligible], socially related phenotype. It's not a crystal clear like a yes/no, how to type a variable, so that would take much more effort from our side. (experienced PI, epi)

This is also echoed by new investigators who echoed that this is a real issue for NIH and academic institutions to consider the amount of work it is to get data in the right format for submission and how this directly impacts compliance with the data sharing policy.

I think they request a huge amount of the metadata, so you have to do a lot of work to complete your submission. So again, many times, it is the effort it requires to the submission to kind of really make people do not want to share. (new PI, gen/gen epi)

In addition, another perspective from the NIH staff was that the challenge of maintaining phenotypic data, which tend have more variability compared to genomic data, in the repositories was concerning to investigators of longitudinal studies, specifically.

For those investigators that are doing the population research studies and are the PI's of those, I think their reluctance to share heavy covariate data is the fact that many of these studies are longitudinal in nature. … So, if somebody's a control today, they might be a case tomorrow. So, the genomic data is relatively stable, but the phenotypic data is not. And unless you're going to be going back on a pretty routine basis to update the phenotypes, which is a big resource burden for the investigator and for the NCBI curation staff, it's very difficult. So, a lot of our investigators now are submitting the genotypes and putting language in dbGaP saying, "If you want access to the phenotypes, contact us for the latest and the greatest and we're happy to work with you to get those data." But they won't deposit them in the repository because they become outdated very quickly. (staff)

This is an opportunity to evaluate mechanisms to enhance the different needs and nuances around different data types and the resources to maximize the sharing of all data types in the most cost-efficient way.

This factor, *Analytic data complexity-Bar,* was also observed in the document reviews (Appendix G).

Lack of clarity of submission/access process – Barrier

A priori factor, *Lack of clarity of submission / access process,* is defined as when respondents referred to the lack of clarity of technical and administrative processes for data submission and access in data repositories. This factor was mentioned by 20 of the 37 respondents (54%), mostly by staff, new investigators and genomicists/genetic epidemiologists.

TABLE XXIX: NUMBER OF RESPONDENTS WHO MENTIONED *LACK OF CLARITY OF SUBMISSION/ACCESS PROCESS* AS A BARRIER

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total (25)
Epidemiologists (14)	2	3	**5 (36%)**
Genomicists/Genetic Epidemiologists (11)	2	4	**6 (55%)**
Investigators (25)	**4 (67%)**	**7 (37%)**	**11 (44%)**
NIH Staff (12)			**9 (75%)**
TOTAL (37)			**20 (54%)**

In the quote below, the respondent mentioned that information or guidance on the dbGaP use was not readily available or not easily accessible from NIH staff who manage dbGaP, given their busy workload on the receiving end of the repository. This impacts the broad sharing of research data with the community.

I guess hindering compliance, I would say that I don't know that these warehouses have information, but I'm not sure it's that obvious how to deal with them, .. the people are — they're absolutely easy to work with, but my impression is that at least [with] dbGaP, they're very busy. So you submit something and six months later they ask you about it, and at that point, like, you don't even remember what you sent, and that this goes on and on. So it actually takes a long time to get the data in, actually to get it public after you submit it because I don't think it's that

straightforward a process. And maybe it's necessarily complex 'cause the data's pretty complex. (experienced PI, epi)

Along the same lines, the perspective of staff alludes to how difficult and complicated it could be for researchers submitting data, especially phenotypic data, in dbGaP compared to other data repositories. One staff respondent explained that this is partly due to the extensive curation of data required for dbGaP data processing as well as the overwhelming number of steps in the process that researchers and their institutional officials must follow for successful data submission. In addition, respondents mentioned that dbGaP system is old and not user-friendly, which poses a barrier to successful data submission.

We also are very strict about the phenotypic data having descriptions for all the coded values and things like that. On top of that for controlled-access, there's a certification of submission that needs to come from the signing officials at the submitting institute saying that it's okay to submit this data. There's a registration step to kind of get things set up and make sure the consent groups are coordinated. There's a data access committee that has to be understanding of what the data is -- that kind of stuff. So, there's some reasons why the controlled-access is quite a bit harder than just submitting like a sequence to Gen Bank or something like that. Well, I think it's, I would say, moderate. Like low, easy to hard. Some investigators find it very difficult. Others find it not so difficult. It's a bit of a process. It's not just clicking on one button. You actually have to have your signing official sign and you have to tell what you're researching. It is an online system at dbGaP so it's not like you're sending emails to people. The system is a bit antiquated and that really comes down to there's just not a lot of funding for it. So, if there was more funding, it could be more modern and sleeker. …. (staff)

The issue with data transfer agreements across institutions and the expertise required to accurately interpret the agreements is one that is related to this factor, lack of clarity of submission/access process, and has implications for how and when data from studies are shared with the public. One staff respondent says in the quote below,

But technology transfer and establishing the data transfer agreements can be challenging between institutes, depending on the lawyers that are working on the agreements. You never know -- somebody's interpretation of what's right and what's wrong or what's legal and what's not legal may be different from somebody else's. So, I think it would be actually good if you were

going to go down this road and get more information to talk to somebody in the tech transfer branch. ….. (staff)

The factor, *Privacy concerns* presents an area with big implications because of the relevance to informed consent forms, IRB requirements and protection of patient confidentiality in data sharing. Both researchers and staff emphasized the importance of being clear about what study participants consented for in terms of sharing their data, as well as adhering to the study participant's consent. The fear of breach in patient confidentiality makes researchers weary of sharing data in a public or controlled-access repository, even more of a reason to ensure that the data sharing processes are very clear at the individual level and more importantly at the institutional / organizational level. It is the responsibility of the institutions to ensure patient confidentiality is protected; which could be related to how the institution's culture around data sharing is shaped.

I'm just more worried about like so from an IRB perspective, is the primary investigator expected to get approval from their individual IRBs for releasing this data publicly, and does that need to be involved in the consent form? I mean all of those types of issues I think are important. (new PI, epidemiologist)

It did not seem entirely clear that researchers understood the limitations of data use as tied to the data and needs to be consistent with and specific for individual consent forms. The example below indicates that researchers may not have access to a specific dataset because of the limitations of that data use. While it is important to be consistent with the consent, it is equally important for researchers to understand why their requests to these types of data may get denied. The DACs are set up to ensure that requestors get data in compliance with what's in the data use limitations. Researchers not understanding this process can result in frustration which can hinder sharing of their data in a repository.

... whether it was truly consented to be so restrictive or whether it's just that an IRB interpreted a consent to say that, we can only give access based on what the investigators say as to how the data can be used. And so I think what is often frustrating for people is they have an idea in mind of a project that they want to do, and there is a cohort that would be perfect. However, the data use that's acceptable on file does not jive with what they want to do. So they're never going to get access for what they want to do unless they go back to those investigators and insist that groups be reconsented and we get a different kind of data use in place. So I think when people say they can't access the data, I think it's that they can't access the data for what they want to do with it. And it's not that they can't access it, but they can't get approval to access it because the use is not consistent. (staff)

Lack of expertise – Barrier

A priori factor, *Lack of expertise as a barrier (Lack of expertise-Bar)* is defined as when respondents referred to a lack of or limited technical expertise and knowledge of the data submission and access process. This includes their knowledge, experience with and understanding of available resources, tools, and systems that support data access and submission processes such as data preparation, and proper data and metadata documentation. This factor was mentioned by 19 of the 37 respondents (51%).

TABLE XXX: NUMBER OF RESPONDENTS WHO MENTIONED *LACK OF EXPERTISE* AS A BARRIER

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total (25)
Epidemiologists (14)	2	4	**6 (43%)**
Genomicists/Genetic Epidemiologists (11)	1	3	**4 (36%)**
Investigators (25)	**3 (50%)**	**7 (37%)**	**10 (40%)**
NIH Staff (12)			**9 (75%)**
TOTAL (37)			**19 (51%)**

Respondents described either lacking or having limited expertise and knowledge of technical and administrative processes of data sharing as required by NIH data sharing policy. This hinders researchers' ability to share data as well as NIH staff's ability to effectively implement the data sharing policy. One central theme echoed by both staff and investigators was that the database, dbGaP is not intuitive to use and is challenging to use especially for first time users. They mentioned that it is not easy navigating the dbGaP database and being able to easily identify the appropriate dataset of interest needed to answer research questions. It shows that there's a gap in knowledge of what data already exists or that's available, and how to access them.

I just found out about one study, this one study that has a website where you can access data about the transition from menopause and looking at nutrition and bone. And I think it's a wonderful resource for investigators, but I didn't know that it existed. And so I think that there may be a lot of data sharing going on that people don't know about. And I certainly -- there's a lot that I don't know about. The exception for me is the one that I know, but I really -- I don't know how extensively it's being used. (new PI, epi)

This is echoed by an experienced investigator, that it's difficult to know what's in dbGaP.

I think that's actually part of the challenge of data sharing, is just even finding what's out there. So — but I'm not sure that, you know, you said what we've learned from this - I'm not sure it's that easy to find out what's out there currently in dbGaP for it. Of course, I was gonna say the UK Biobank is — so it's easier because they have put together the whole thing. The dbGaP is, you know, by contribution, so it's more dissimilar. (experienced PI, epi)

The differences in perspective between the new and experienced investigators in terms of ease of access and use of data from the repository is likely dependent on the level of experience and expertise with the database. A new PI says *"it's not obvious how to access data from database"* while a more experienced investigator says it's *"very easy to use and can be accessed in one afternoon".*

This same factor, *Lack of expertise-Bar*, was observed in the document reviews (Appendix G).

Sub-Optimal Repository capabilities – Barrier

The a priori factor, *sub-optimal Repository capabilities* as a barrier *(sub-optimal Repository capabilities-Bar)* is defined as when respondents referred to the insufficient capacity, ineffective, inadequate and inefficient data repository as facilitating data sharing. This factor was mentioned by 27 of the 37 (73%) respondents.

TABLE XXXI: NUMBER OF RESPONDENTS WHO MENTIONED *SUB-OPTIMAL REPOSITORY CAPABILITIES* AS A BARRIER

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total (25)
Epidemiologists (14)	2	3	**5 (36%)**
Genomicists/Genetic Epidemiologists (11)	2	9	**11 (100%)**
Investigators (25)	**4 (67%)**	**12 (63%)**	**16 (64%)**
NIH Staff (12)	[redacted]		**11 (92%)**
TOTAL (37)			**27 (73%)**

This was identified to be one of the most recurring thematic factors across the data, among the different respondent groups, as hindering data sharing. When respondents were asked how easy it is to submit data in dbGaP or any other data repository, the investigators and staff mentioned the frustration with this, emphasizing that the database was "a real pain," "clunky" and mentioned that streamlining the processes would facilitate easy sharing.

dbGaP can be a real pain. So, I was involved in a study a few years ago that was a large collaborative multi-institutional study and samples were collected in a number of different institutes .... We tried to submit that to dbGaP, and honestly I've been held up for years because one of the organizations where samples were collected has not come through with their required documentation of IRB-approval for sample collection. So, and I have no leverage over them. Nobody in the project has any leverage over them. They just haven't delivered. And so, there's a dataset where we've made the processed data available to people, but we haven't been able to make the raw data available. And I will say, it has - in trying to work through this process it's been rather difficult. I think the people at dbGaP try, but sometimes it's a real challenge to get data into the system too (experienced PI, gen/gen epi)

The administrative/technical burden and lack or inadequate support /resources as a barrier related to the capabilities of the data repository is evident when respondents compared dbGaP to other repositories.

Oh, it's very hard. I mean, this is very time-consuming process because they have; I just want to say, the reason that we usually submit our data in dbGaP, and then move to the European data repository because they have less restrictions about what you need to do, the steps you need to do to submit data on their repository. So a lot had to do with how much effort it takes to share data with the global community. Sharing data does not benefit the data generator, in general. It's a good thing they do for the society. But even screening is a lot of work. People are hesitant [unintelligible] so much effort. So I think between dbGaP and EGA, EGA is an easier place to put your data in than dbGaP. (new PI, gen/gen epi)

I think that unfortunately actually dbGaP has suffered from a lack of resources and so a lot of their processes right now are outdated and their personnel numbers are low so that since many of the exchanges to support investigators in submitting data require back-and-forth communication or telephone conversations. Those just take a long time to happen because of the short staff. (staff)

Respondents reported that part of ease or lack thereof of data sharing among researchers is dependent on their level of expertise, knowledge and experience that they have with sharing data in a data repository such as dbGaP. Staff respondents confirmed that the dbGaP system is overwhelmed and causes delays in the submission process.

Well, again, resources to be able to do it and a system that is user friendly just to make sure that as you're uploading data, that it's clear, it's easy and it is not fraught with the

difficulties that, for instance, we routinely run into with dbGaP as we start to upload data and keep getting, errors keep happening during the upload which requires starting over and starting to re-upload again. So I think that platforms need to be put in place that make it easy and reliable to upload data. (experienced PI, gen/ gen epi)

Research Question 2: What are the opportunities for improving / enhancing data sharing?

OPPORTUNTIES – Culture and Practices

Culture shift in research fields - Opportunities

To answer research question 2 of this dissertation research, opportunities for changing the thinking and perception around data sharing in different research fields (*culture shift in research fields-Opp)* was explored through the data. This would enhance broad sharing of data in public or controlled-access databases. Culture shift as an opportunity for enhancing data sharing was primarily echoed by staff compared to the researchers. 8 of 37 respondents (22%) alluded to a few opportunities within the CULTURE construct and does not include epidemiologists as shown in the table below.

TABLE XXXII: NUMBER OF RESPONDENTS WHO MENTIONED *CULTURE SHIFT IN RESEARCH FIELDS*

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	0	0	**0 (0%)**
Genomicists/Genetic Epidemiologists (11)	1	2	**3 (27%)**
Investigators (25)	**1 (17%)**	**2 (11%)**	**3 (12%)**
NIH Staff (12)	[redacted]		**5 (42%)**
TOTAL (37)			**8 (22%)**

Staff respondents commented on the need to change the culture of how researchers perceive data sharing and do research by mentioning NIH's efforts *"to get researchers to start seeing data sharing as part of the natural research process as opposed to a separate activity*." They mentioned that NIH can't do this alone and would require the institutions along with professional academic organizations to help investigators see the importance of sharing data, and therefore start shifting their thinking and culture at institutions around data sharing. This is related to the opportunities for addressing a change in reward structure at institutions.

... I think that there needs to be a culture change to be able to have the universities incentivize their staffs to do this [share data]. We're only a piece in that and we're trying to do that, but if they can continue to do it, then that's definitely needed. So if it's, as I mentioned before, if it's that they make a statement about it is included in tenures, weighted just as heavily. Or you know, they give bonuses for people who share – whatever it is.(staff)

In the quote below, one respondent mentioned that providing targeted resources similar to genomic data sharing to support technological infrastructure needs of epidemiology data sharing would be something that NIH could do to facilitated data sharing in the field of epidemiology. Although there are existing databases for epidemiology data submission, it is clear that this respondent was not aware of what these repositories are. For example, there's NHLBI's supported repository for biospecimens and data (BioLINCC) and other ongoing efforts at the program level (in EGRP) to leverage existing databases such as dbGaP for epidemiology data sharing.

Yeah, I think that it would be actually useful to have a central repository for epi data like there is for genomic data, and it would have to have some pretty clear guidelines and rules for how those data can be used ... For ongoing cohort studies where there's a lot of follow up and things that take a long time. ... I think, is just that it's not part of the culture of epi studies, whereas I think it is a part of the culture of genomic studies that everybody knows the data need

to be deposited and they'll be public and all that sort of thing, and it's just become kind of a standard that people know to follow. I don't think epi studies have that in their culture, and I think there'd be concerned about that. Not to say it shouldn't be done, couldn't be done, but I think that's [not] just part of the culture for epi studies to be centrally deposited and allow anybody to use them for any purpose like the genomic data are. So I think that it would maybe take a culture shift or a bit from NIH forcing that to happen because it's just not the way things have happened for a long time. You know, the epi culture of data sharing is new, relatively new compared (experienced PI, gen/gen epi)

Respondents mentioned that data sharing may not be a top priority for their institutions and not something their institutions are willing to invest in. This is change that needs to happen at the leadership or institutional level to facilitate sharing of data; a systems level approach given the diversity of multiple key stakeholders involved in data sharing in federally funded research. To address this concern, one respondent suggested that NIH should recognize institutional leaders for data sharing occurring at their institutions, in addition to recognizing the researchers; reward their contribution by assuring funding.

Reward structure changes needed for data sharing - Opportunities

To address changes needed in reward structure for data sharing, 14 of 37 (38%) respondents provided some perspective on potential opportunities for researchers that could help with enhancing data sharing.

TABLE XXXIII: NUMBER OF RESPONDENTS WHO MENTIONED *REWARD STRUCTURE CHANGES NEEDED – OPPORTUNITIES*

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	1	2	**3 (21%)**
Genomicists/Genetic Epidemiologists (11)	0	4	**4 (36%)**
Investigators (25)	**1 (17%)**	**6 (32%)**	**4 (16%)**
NIH Staff (12)	[redacted]		**7 (58%)**
TOTAL (37)			**14 (38%)**

Getting credit, recognition and rewards from institutions for sharing data by making it count towards promotion and tenure is an incentive that was mentioned by both researchers and staff as an opportunity to increase data sharing given that this is not existent in academic institutions.

I think obviously more people would share if they got credit for it somehow; acknowledge how much they share. Maybe that would help.(experienced PI, gen/gen epi)

NIH staff corroborated the above comment from a researcher and indicated that there is recognition at NIH about this.

Some ongoing discussions amongst a lot of different scientific communities is how do you incentivize researchers. How do you incentivize people to share? And a lot of people believe that investigators being able to receive credit for their successful data sharing would help because then they would be able to use that in their tenure and promotion packages.(staff)

When respondents discussed incentives, they framed them as potential opportunities for institutions to enhance data sharing through changing the institution's culture and norms for academic credit. These comments indicated the lack of recognition by institutions of data sharers, and therefore an opportunity for institutions / organizations to enhance data sharing through reassessing their culture for rewarding data sharing, and how they evaluate the careers of researchers during the promotion and tenure process. Including data sharing as a criteria in the promotion and tenure criteria was a sentiment echoed by staff and new and experienced investigators.

Institutions themselves, I think I would love to see them incentivize sharing. In some way, whatever way that is, whether that's through giving of more research dollars or space or acknowledgments or accolades. But you know, a factored track to your tenure, or something along those lines would be fantastic if we started incentivizing the sharing from an institute level, because I personally see that as a potential barrier at this point and not a helpful space. So I

would love to make it a helpful space. I think that would be highly motivating, because I really do think that a lack of sharing comes from at least the perception that it would be problematic to someone's career.(staff)

And again this sort of comes back to the way our careers are evaluated. You know, our CV is number of grants, number of publications and so it's always sort of a race. If it's promotion criteria focused more on team science or collaborative science, and it didn't matter if we were middle authors in some big thing, then I don't think this would be as much of a concern from a scientist perspective. I mean if you share data I don't think it's going to be a major factor in your promotion process or not, which is a shame. I guess that's our culture to change.(new PI epi)

One strong theme across the data from investigators was for NIH to recognize that data sharing occurs through collaborations beyond just the data repositories and to give credit to the researchers for collaborating. This relates to the differences in how investigators and NIH are defining data sharing; investigators seemed to define data sharing to emphasize collaborative sharing, and NIH seems to emphasize sharing through a data repository as the ideal mode of data sharing.

Both experienced and new investigators with expertise in epidemiology and genetic epidemiology expressed the desire for institutions to recognize publications resulting from any level of involvement in large collaborative or consortia studies during the promotion and tenure review process. For such collaborative studies, the researcher may end up as a middle author instead of the highly coveted first or last author position on a paper, but their contribution should be valued equally.

Well I think there just has to be a way for the NIH to recognize that collaborations do exist outside of this funded mandate [to share data in data repositories]. I gave you one example of [redacted], but I can give you another dozen examples of what people emailed me. "Hey. I know that you're involved in that consortium. Can you tell us is this particular variant associated with this disease?" And we'll do the analysis and we'll share the results. We never said no to anybody for that. But there's no way to document it. There's no way to get credit for

that. I think that there's probably more than half that is being documented. Experienced PI, gen/gen epi)

I think one thing is what we just talked about, which is having those collaborations actually be part of the criteria for promotion, collaborative papers. You know, recognizing that even a middle author position on those papers still means you did a lot of work and you contributed to an important scientific issue. So right now I would say that those middle author papers count more abstractly in developing your professional network, but in terms of the count of publications you have, they don't count so much. So one thing would be to revising the promotion criteria, promotion and tenure criteria. (new PI, epi)

Respondents, mostly NIH staff with experience in policy development, implementation and leadership, mentioned current efforts and future opportunities by NIH in trying to support movement toward the change in culture at academic institutions around promotion and tenure.

Well, there's a shift as I mentioned and we're also trying to encourage a culture of data sharing, so the NIH has changed some of their evaluation criteria. So beyond just listing the publications, a lot of universities are now trying to include publications and products for tenure status. So what it allows is that you can put in a program that you developed or a technology that you've developed or in our case it would be a digital object identifier for data sets that you share. In VAs right now, the community has a push for all of these to be weighted equally for tenure status. So it looks like that's continuing across the world. If that continues, then that for sure incentivizes individuals to want to share so that they can ultimately get credit for it.(staff)

I will say that from the NIH point of view we have begun to allow for a citation of data resources that people have created as one of the things that are considered as part of the peer review process.(staff)

Another suggestion was that data sharing has to be a priority for the institutions where they address the concerns or fears of investigators related to credit and recognition for sharing data. Part of that priority involves higher level discussions with other organizations about how to make this a reality and potentially change the culture of promotion and tenure process at institutions. This is something that NIH is interested in. Participants were asked what academic

institutions and NIH could do to facilitate data sharing and prioritizing data sharing and collaborations through acknowledgment of data sharing in the tenure and promotion process, similar to that of publications were key.

In terms of how NIH could incentivize data sharers, staff respondents and investigators at all levels and with different experiences mentioned that NIH should reward data sharers through the grant review process by including it as part of the review criteria during funding decisions. This would mean that investigators who demonstrate that they share their data could be given "*a couple of higher points, better score, for following the policy" (staff).*

So if we're able to ask investigators when they come in for funding if they can tell us about their past history of sharing, and reviewers look at that and they view that favorably when they're trying to evaluate future funding decisions. Just ways that we can try to incentivize people to share.(staff)

The idea that this should not just be tied to the individual grant but also be reflective of the institution's support for data sharing was an interesting addition to the previous comment. An experienced investigator suggested that

Maybe a level of cut to your grant depends upon how well your institution is known sharing data. I don't know. (Experienced PI, gen/gen epi)

This same factor, *Reward structure changes needed-Opp*, was observed in the document reviews (Appendix G).

RQ 2: OPPORTUNITIES - Policy

Clarity of policy needed – Opportunities

In the data analysis, references by respondents to opportunities needed to improve the clarity of data sharing policy was coded a priori as *Clarity of policy needed-Opp.* This code was mentioned by 9 respondents (24%).

TABLE XXXIV: NUMBER OF RESPONDENTS WHO MENTIONED *CLARITY OF POLICY NEEDED - OPPORTUNITIES*

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	0	0	**0 (0%)**
Genomicists/Genetic Epidemiologists (11)	0	1	**1 (9%)**
Investigators (25)	**0**	**1 (5%)**	**1 (4%)**
NIH Staff (12)	[redacted]		**8 (67%)**
TOTAL (37)			**9 (24%)**

Increased clarity around policy guidelines and implementation processes are considered important to both the staff and investigators. This is in light of challenges described earlier in this chapter, related to the ambiguity in policy requirements / expectations. This concept was closely related to references made by respondents on the need for clearer guidelines with the data submission and access processes *(clarity of submission/access process needed-Opp)*, especially as it relates to requirements for specific types of data and the timeline for submission.

Yeah, I think that it would be actually useful to have a central repository for epi data like there is for genomic data, and it would have to have some pretty clear guidelines and rules for how those data can be used …For ongoing cohort studies where there's a lot of follow up and things that take a long time the question of when you deposit — you know, like, if you're following people for 30 years, when do you deposit the data? When is it done? You know, or do you just deposit the baseline data? … (experienced gen/gen. epi)

Change needed in enforcement - Opportunities

There were 15 respondents (41%) who described potential opportunities for improving the sharing of data through new strategies or approaches to enforcement of data sharing policy.

TABLE XXXV: NUMBER OF RESPONDENTS WHO MENTIONED *CHANGE NEEDED IN ENFORCEMENT*

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	1	2	**3 (21%)**
Genomicists/Genetic Epidemiologists (11)	1	3	**4 (36%)**
Investigators (25)	**2 (33%)**	**5 (26%)**	**7 (28%)**
NIH Staff (12)			**8 (67%)**
TOTAL (37)			**15 (41%)**

The a priori code, *Change needed in enforcement-Opp*, defined as changes needed in the existing strategies or approaches to enforcement of data sharing policies was used to capture this concept. The quote below from one staff shows that NIH leadership recognizes the need and is exploring other options or mechanisms that could potentially be more effective, but they weren't clear on what these options would be.

There are a lot of things we could do to put requirements around data sharing at a much broader level. We could do, you know we could decide we want to make a dbGaP for every data type to help, you know, sort of give researchers a place to put their data, we could pay for it all. You know, all of these are things we could do. Whether or not they're the right thing to do I think is very much this current subject of debate.(staff)

In addition, NIH staff mentioned ongoing efforts in terms of expanding the current policy requirements for data type with no budget limit and leveraging the 21st Century Cures Act signed into law in 2016. This will make it easier for enforcement and implementation by staff.

We also right now are working on drafting a new data management and sharing policy that has no monetary threshold and it's not for any specific data type in particular. And we're using the backing from 21st Century Cures, which enable the NIH director to require data sharing. … what we're trying to do going forward is that we're trying to have the investigator indicate their data management and sharing plan what their expected timeline is. And then before any funding decisions are made, that is potentially negotiated with the funding IC to be

able to determine is that an appropriate timeline, so that the institute or center that is ultimately funding them, has a say in it and they're not going to be upset three months down the line when they don't share it because they agreed to X, Y, and Z. but at the same time, it allows them the flexibility to be able to push back on the investigator and say, "Oh, no, no, no. We need it before any potential publication is made". (staff)

The variation in enforcement and implementation processes across NIH institutions and centers, and varying levels of communication and transparency was described earlier in this chapter by staff and investigators as a barrier to data sharing. There's a need for consistency in the enforcement of policy across NIH as well as across journals which don't all have the same requirement for data sharing. When respondents were asked what NIH could do to facilitate data sharing, they mentioned the need for consistency in the way that data sharing policy is implemented through standardized templates to help with uniform documentation of information in the data sharing plans, as well as being more proactive, upfront and forthcoming and to develop a plan for dealing with non-compliant researchers so enforcement is equal across the board.

I think NIH has been pretty good on this. I feel like they give notice, encouragement, and sometimes even pressure from NIH to deposit data on time. But we also noticed that there are some sites they are not really in compliance or delay [submission]. I'm not sure whether NCI has a policy to enforce this. I mean, should not say punish, but just to restrict those people to get further funding if they don't comply with these policies. Yeah. To put it that way. So I don't know what NCI or NIH is doing for when the people are not in compliance.(experienced PI, epi)

More effective and enhanced enforcement or strategies could be done through establishing and following through with clear consequences with investigators who are not complying with policy. One of the investigators suggested approach is mentioned in the quote below:

… I think sometimes it's, the thing that is really to identify some people who are repeat offenders and then crack down. That's the thing to do. But I think everything is working that way because also if the journals don't require now, the reviewers are asking for it, and you can still dance around the issue, but less and less I guess. (experienced PI, gen/gen epi)

The consequences would include barring them from accessing data from dbGaP, as well as withholding of funding as suggested by a staff– something that is currently in place but may not be enforced consistently or as broadly across the NIH.

So we have to monitor whether or not the grantees are submitting their data to the appropriate databases. So we actually currently are not supposed to award funding until they've provided proof that they can share that data and they've provided appropriate data sharing plans.(staff)

... I think prospectively, what I certainly would advocate for and hopefully others will as well, is from the outset, enforcing the policies, like actually enforcing the policy. Being a little stronger about it up front and not funding things that can't be shared a certain way up front. Because if you are going against the policy or you know up front that there is a potential that there is going to be a conflict there, then my personal feeling is it probably shouldn't be funded by public money. (staff)

When respondents were asked about what academic institutions could do to facilitate data sharing, the perception was such that institutions could provide the administrative support data sharing by providing dedicated support to their researchers. Similarly, there was indication for increased administrative / technical support on the NIH end to support the management or flow of data in dbGaP. This is related to leadership support at the institutions and indicates an opportunity for academic institutions to step in, in collaboration with NIH, to facilitate sharing of data among their researchers by providing them with the resources needed. This sentiment, illustrated in the quotes below, was echoed by staff and investigators. The NCI has recently created a central office, NCI Office of Data Sharing, to address some of these concerns, to help with more effective implementation of the data sharing policies.

Well, I think if NIH has a policy about data sharing, that the divisional institutions need to make sure that that's enforced but I'm sure that's not happening because I mean there's specific people to help you submit your budget and submit your progress reports, but there's no specific job that is looking at data sharing that I know about at least in my institution. (new PI, epi)

I think there needs to be more staff, definitely on the side of dealing with the databases and the creation of the data. I think there could be probably more of a tracking system, so we could know where things are in the process. (staff)

Addressing privacy concerns - Opportunities

The a priori code, *Addressing privacy concerns-Opp* was used to describe when respondents mentioned potential opportunities for addressing concerns of data privacy in public or controlled-access data repositories. This code was mentioned by 6 respondents (16%).

TABLE XXXVI: NUMBER OF RESPONDENTS WHO MENTIONED *OPPORTUNITIES FOR ADDRESSING PRIVACY CONCERNS*

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	0	0	**0 (0%)**
Genomicists/Genetic Epidemiologists (11)	1	3	**4 (36%)**
Investigators (25)	**1 (17%)**	**3 (16%)**	**4 (16%)**
NIH Staff (12)	[redacted]		**2 (17%)**
TOTAL (37)			**6 (16%)**

There's always a risk for violation of patient confidentiality in research and identifying ways to alleviate researchers' / institutional fears is important. Change in the informed consent processes / practices, and training / education on the relationship between consent and data sharing would be helpful.

One investigator suggested changing the process for consenting patients by making it explicit about the intent to share data during the consent process and use that as an opportunity to educate study participants about data sharing and what it entails. This might be a way to enhance data sharing on both the researcher end and the study participant end.

... certainly these days one should be consenting people with the notion that it should be going into some — that it will be shared, the data will be shared, but I don't know that people always do or have thought about consent and data sharing at the same time. Maybe they — and obviously they should, but I think having the participants know that their data will be shared widely, I don't know that that always happens in a consistent way with the data sharing policies. So that's a disconnect that might be thought about. You know, again, everybody should be aware of that, but I'm just not sure that it always happens. Or that it's happening in a way that the patient, the participant really knows that their data might just go out to the whole world in some way. I'm not sure, and I can see that — I've heard that as an issue only — you know, like, retroactively when somebody says, "Well, I can't really deposit my data because I didn't — or share my data because I didn't tell — the participants didn't consent to that." Again, prospectively going forward, they should be, but I don't know that that's always happening.(experienced PI, gen/gen epi)

When respondents were asked what NIH could do to facilitate data sharing among NIH funded researchers, one of the respondents mentioned the issue around consent and opportunity to allow for broader consent that will allow broader sharing in future research. This would include developing templates to be used to support broad sharing. The following examples representing staff and investigator perspectives both support previous comments by respondents on the limitation of data use that is bound by the participant consent forms and the data sharing policy and the need to develop a mechanism for broader consent for sharing data. Addressing this through updating the 2003 NIH data sharing policy, which is currently underway, could help achieve the ultimate goal of enhanced data sharing.

I think the most important is the consent and the broader the consent, the easier it is to share. So I mean I think with the repositories and with controlled-access mechanisms, there are processes in place to ensure security and confidentiality of the data, but I think the consent issue is the biggest issue. Because it's very hard to share broadly if that dataset can only be used for schizophrenia research, or if it can only be used with the investigators, the primary investigators of the study. But if there can be a more broad consent for future research purposes and broad sharing, then it is easier. So moving forward, that's what we're hoping people will do. But for the legacy datasets that exist, it's very difficult. Like you can't go back to Framingham 60 years and - you know. But I would say that's one of the biggest challenges. People want access and they want to get it fast, but it's just can't happen sometimes because there are limitations on the use of some of those datasets. (staff)

So it's kind of more complicated than just saying that the scientist does not want to share data of the bad persons, but a lot has to do with the overall consenting process, how that was set up. So I think being able to change this could be [unintelligible] is to create a good general template of the consenting process that will enable broad data sharing in the future.(new PI, gen/gen. epi)

This factor, *Addressing privacy concerns-Opp*, was also observed in the document reviews (Appendix G).

RQ 2: OPPORTUNITIES - Resources

Administrative / Technical resources needed - Opportunities

The a priori factor, *Administrative / technical resources needed-Opp* was defined as when respondents described some opportunities for addressing the administrative and technical barriers to data sharing and ways to improve data sharing among researchers. This factor, *Admin/tech resources needed-Opp was* mentioned by 18 of 37 respondents.

TABLE XXXVII: NUMBER OF RESPONDENTS WHO MENTIONED *ADMINISTRATIVE / TECHNICAL RESOURCES NEEDED*

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total (25)
Epidemiologists (14)	3	2	**5 (36%)**
Genomicists/Genetic Epidemiologists (11)	1	5	**6 (55%)**
Investigators (25)	**4 (66%)**	**7 (37%)**	**11 (44%)**
NIH Staff (12)	[redacted]		**7 (58%)**
TOTAL (37)			**18 (49%)**

NIH can enhance data sharing by providing resources to hire additional staff to provide administrative / technical support for data sharing processes. Some of the comments from the respondents addressed needs of the investigators on the submission side and that of NIH staff on

the implementation side. They mentioned the need for more staff and better systems and administrative / technical processes that make it easier for investigators to share data, and for NIH staff to better implement the policies. .

I think there needs to be more staff, definitely on the side of dealing with the databases and the creation of the data. I think there could be probably more of a tracking system, so we could know where things are in the process. (staff)

Similarly on the researcher end, respondents indicated the need for administrative support in the form of a programmer and technical support from the institutions and NIH, i.e. having staff at NCBI who manage dbGaP to assist with technical challenges encountered during the process. This was particularly important to both new and experienced investigators.

Maybe kind of somebody who's responsible to help us, like a programmer time. Somebody who can help us navigate through how this data sharing will be done properly, legally or anything. In terms of epi investigators, I think just having somebody who's always available to answer our questions, either at the NIH site or at [institution] would be great and helpful. new PI, epidemiologist)

An interesting comment from a new investigator and epidemiologist was that there are no defined roles for data sharing, whereas there are people who can assist with budgeting and submitting other materials such as progress reports at their institutions. The indication is that having a clear sense and knowledge of who to go to for the different aspects of the data sharing process would make it a lot easier to submit their data.

This factor, *Admin/tech resources needed-Opp*, was also observed in the document reviews (Appendix G).

Financial resources needed - Opportunities

The factor, *Financial resources needed-Opp*, was defined as when respondents described some opportunities for addressing the funding and financial resource needs of researchers and

staff to improve data sharing (*Financial resources needed-Opp)*. This was mentioned by 17 of 37 respondents.

TABLE XXXVIII: NUMBER OF RESPONDENTS WHO MENTIONED *FINANCIAL RESOURCES NEEDED*

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	0	4	**4 (29%)**
Genomicists/Genetic Epidemiologists (11)	1	4	**5 (45%)**
Investigators (25)	**1 (17%)**	**8 (42%)**	**9 (36%)**
NIH Staff (12)			**8 (66%)**
TOTAL (37)			**17 (46%)**

The respondents mentioned the need for funding and financial resources in terms of salary support for personnel, to help facilitate and enhance data sharing, as well as funding to be able to do analysis and share data quickly, even before the publication of main findings, as being critical.

Analysis of the data showed a direct relationship between the codes *Financial resources needed-Opp and Leadership support needed.* In the example below, the respondent noted the lack of priority given to data sharing by institutional leadership and the lack of financial resources at the institutions which directly hinder data sharing. This case presents an opportunity to explore how to work with academic institutions to prioritize data sharing.

If the institution had the resources to cover the data sharing, that would be terrific. But I can tell you in the reality of things, if I've got $100 that I can spend on research, taking some of that $100 to put data in a data-sharing commons? Not gonna make the top 1,000. It doesn't benefit the institution at all to do that. There's no return on the investment ever .. (experienced PI, gen/gen. epi)

The respondents, both researchers and staff, alluded to the need for financial incentives / resources as important in data sharing. The culture of data sharing at institutions is one that does not incentivize data sharing and as a result impacts how well data is shared by the researchers. This presents an opportunity for institutions to revisit their reward structure or systems, as previously described earlier in this chapter. Suggestions were made for NIH to incentivize data sharers through providing funding, as well as explore different options that will expand the level of funding for data sharing.

The code *Analytic data complexity-Bar* co-occurred with *Financial resources needed-Opp* and is illustrated in the example below where a staff who acknowledged that a big bulk of the burden with data sharing has to do with the complexity of the analytic dataset and preparation needed for meaningful / useful sharing, and suggested including funding for data sharing into grants. This was an emphasis among staff to have data sharing costs explicitly accounted for in the grant application, so it's clear, and made known to investigators, as some may not be aware of this. This is happening but may not be well-known among the investigators and staff.

There's definitely a lack of funding. There's kind of two things: one is that it does take some effort to get your data in an archivable, consumable form. And so, the NIH really should be -- because they have this mandate, they really should be building this into the funding explicitly. I think it's implicit because of the policy. But there should be a line item in every grant, saying X number of dollars for the data sharing staff is paid to you. That would help. (staff)

This factor, *Financial resources needed-Opp*, was also observed in the document reviews (Appendix G).

Leadership support needed - Opportunities

The respondents described some opportunities for increased leadership support to improve data sharing *(Leadership support needed-Opp).* There were 11 respondents who mentioned this a priori code.

TABLE XXXIX: NUMBER OF RESPONDENTS WHO MENTIONED *LEADERSHIP SUPPORT NEEDED - OPPORTUNITIES*

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total
Epidemiologists (14)	1	4	**5 (36%)**
Geneticists/Genetic Epidemiologists (11)	1	2	**3 (27%)**
Investigators (25)	**2 (33%)**	**6 (32%)**	**8 (32%)**
NIH Staff (12)	[redacted]		**3 (25%)**
TOTAL (37)			**11 (30%)**

A comment from one staff questioned the support by NIH leadership of the value of epidemiology data sharing, indicating the perception of a lower priority at NIH compared to genomic data sharing. This seems like an opportunity for leadership to increase investment in resources to support non-genomic data sharing, communicating its value to the scientific community.

It sheds light on the limited or inadequate support from NIH leadership on sharing of non-genomic data, which affects the implementation of the data sharing policy by staff who deal with grantees on a day to day basis. Making the priority clear with targeted resources for epidemiology data is likely to result in a ripple effect that will empower program staff to better enforce the policy thereby increasing increase compliance among epidemiology researchers. The recent effort to update the 2003 NIH Data Sharing policy is an attempt to address these concerns

at the institute level, although some programs are already implementing some aspects of this among their grantees.

I think we talked about buy-in. So I think, to me, so [redacted] as a genetics person, so there's been like from the top, there's been that emphasis for sharing, and you know he pushed the quick timeline. Where's the support for the 2003 policy or the non-genomic, like if you just go to Epi? Generally we say "Epi studies". Where's that coming from? It's as though that it's two different classes of information. So that kinda gets to my theme of like we don't have much training in the 2003. There's just so much that doesn't seem like it's coming from the top so much, where the genomics policy, everybody knows they came from the top. I think what we need is if we think that the non-genomic Epi data is just as valuable, we need to know that the entire NIH community needs to know it, and the research community needs to know it because right now it does seem like two different classes of information. And we need to make sure that we can support, you know, put our money where our mouths are, and support it, you know, literally put our money where our mouths are. (staff)

One investigator mentioned not knowing what happens to data beyond the end of the grant and what the implications might be for the researcher and NIH. Currently there is no funding to support the submission of data after a grant ends; this is considered the responsibility of the investigator to ensure that all data are submitted within the given timeline. Respondents indicated that this was challenging and that NIH's provision of funding will ensure sustainability of ongoing submission of data in the data repository even after the grant funding ends. The suggestion was for NIH to support data sharing beyond the life of the grant i.e. after the study has ended by providing the funding for that in the grant.

<u>Training needed – Opportunities</u>

Understanding what types of training opportunities and education needed by researchers and staff to increase knowledge and skills was considered to be important for enhancing data sharing.

The a priori code used to describe this concept was, *Training needed-Opp,* mentioned by 15 of 37 respondents.

TABLE XL: NUMBER OF RESPONDENTS WHO MENTIONED *TRAINING NEEDED-OPP*

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total (25)
Epidemiologists (14)	1	4	**5 (36%)**
Geneticists/Genetic Epidemiologists (11)	0	2	**2 (18%)**
Investigators (25)	**1 (17%)**	**6 (32%)**	**7 (28%)**
NIH Staff (12)			**8 (67%)**
TOTAL (37)			**15 (41%)**

When the respondents, both researchers and staff, described some of the challenges with sharing data, part of that was the lack of knowledge about the processes involved in data submission / access at the NIH and at their institution. Training needed to effectively and successfully submit data in data repositories came up in the discussions and it was often tied to the need for increased clarity in the submission of data in data repositories. This example reinforces the importance of providing materials and tools for researchers and staff to use to help with understanding the data sharing process, requirements and expectations. It also mentions that there is an opportunity to learn from existing databases and tools to inform how NIH could enhance its data repositories like dbGaP.

But if there could be tools or good education support for uploading the data relatively easily, so even if you do require formatting that you have some sort of process in place that allows a more naïve user to be able to come in and walk through some sort of process that would allow them to format the data or answer the appropriate questions. Then I think that you would get a lot more data sharing, because sometimes I really think it's just physically being unable to upload, if you haven't used a particular software or system or format. (staff)

So whether there are guidelines or recommendations, I think that would be helpful in terms of sort of just knowing what is the minimum to get someone able to use a data set in the proper way I think without kind of misinterpreting or misunderstanding the data, and that is also sort of standalone in that you wouldn't have to always go back to the lead investigator or to the programmer? So I think part of it is both having a data set that is clean and usable and enough either documentation, whether that is a video or a webinar type of thing, and something that has been tested. … I think there is something about sort of communication and whether it's education or another piece of it that I think is missing that could be helpful we could learn from something – I think there are databases out there, like NHANES or even SEER – I think the other thing is the direction to really be able to go is something where it would be just a little bit easier, where there would be tools that are embedded where you might be able to answer like 80 percent of the questions that people have of the data. (experienced PI, gen/gen epi)..

One staff mentioned that training of the administrators at the institutions was as equally important as training for the researchers so that they are also aware and have a clear understanding of the complicated process as well as their institutional processes. This includes understanding the process for data transfer agreements, and approvals by their institutional signing official. In addition, training to establish consistency in the interpretation of the data sharing policy within the institution, between the IRB and investigators, especially as it affects the review of consent forms in the context of data sharing.

... I think the hard part is in getting the approval and that's because it is not only NIH that has to approve it, but it has to go an institutional signing official for review. And that's an unfamiliar step for signing officials to approve a data access request. … And so investigators, I think, submit these requests and it goes off into administrative land within their own university and it never gets approved and then they think it's our fault. It's something that they have to educate their own signing officials about, or NIH has to do more outreach to administrative representatives and do policy education there.(staff)

One comment from staff was that academic curricula should include data sharing as part of their training. This should be done early in their training, similar to bioethics training, and data sharing can be included as part of professional development training / requirement at universities. This elicits the need for a culture change at the institutions around data sharing.

*… and then I think some of it is an issue of socialization of the concept, .. What I meant by a socialization process is it's something that has to be brought up when we're being trained. Just now, you get training in bioethics, for example – I would say this is something important to be training in as well.(*staff)

Despite the various types of materials, tools and guidance that respondents mentioned as currently facilitating or that could facilitate data sharing, one staff noted that some researchers don't plan ahead or don't always anticipate journal requirements for publication such as the accession number obtained once data has been submitted. They need to know that the process for data submission could take a long time, therefore needing a better understanding of how the process works and be more proactive.

There's submission instructions. It's all like this is what you've got to do. It's all public. We have web, You Tube tutorials on how to do it. But if they don't start looking at the problem or the task until the very last minute, I'd say they will be surprised if you can't finish on time because you did not - you know when you submit a paper that, it's in the publication guidelines, what they're going to ask you for - accession numbers. So when they come back and ask at the end, you shouldn't be surprised because you agreed to this when you submitted your article to peer review. When the people read it, they're busy. I get it, they're busy and it's not a priority for them. (staff)

In addition, the communication of existing resources, tools and findings on the use and value of data can be improved across the scientific community. The example below indicates a lack of awareness of such efforts and existing resources.

Maybe having a workshop and also some ideas about what the data has been used for, various research. Maybe highlight some important findings that people have done using these shared resources. I think that would be of interest to us, to see, okay, what has been done, what's out there, what findings, that kind of stuff, and for like interesting findings. You know, researchers love to read, like, oh, you know, or give us some ideas about, well, we can actually use this repository data to do this kind of research, innovative, creative, that kind of stuff. So maybe this [publications or findings on data use] is something that you can share among the funded researchers once in a while, like, " hey, look at this; you know, this repository has been used to find this very cool research." (new PI, epi).

Communication strategies on data sharing policies vary across the NIH institutes and centers. Therefore this seems like a practical opportunity to enhance publicity around policies and communication methods to get the word out on requirements / expectations. One participant suggested NIH send reminders to investigators to submit data (which NIH is currently doing) and to read updates to policies when they come out even before they receipt of notice of grant award.

For staff, training and education is important to help staff be able understand some of the nuances involved in the processes and to help investigators navigate through the problems both on the submission and access side. Training would also help them understand what the research study challenges are around data which will help to inform methods and approaches to policy development and implementation. A foreseeable challenge with the training is conducting it in a way that's consistent across the ICs.

This factor, *Training needed-Opp,* was also observed in the document reviews (Appendix G).

RQ 2: OPPORTUNITIES – Technological Infrastructure

Addressing analytic data complexity - Opportunities

The respondents described opportunities to improve data sharing through addressing some of the challenges they identified with the complexity of analytic data as it relates to the submission / access in a data repositories (*Addressing analytic data complexity-Opp)*. This code was mentioned by 5 respondents including 4 staff and 1 experienced epidemiologist.

The respondents suggested that streamlining the data sharing process was very important, especially with standardizing data formats. Requiring that the data be collected in a similar format will help improve sharing in data repositories

So I think some of the biggest barriers to people who want to share are the fact that there are requirements for any database as far as the format that a data file needs to be presented in. There is metadata surrounding both the data files themselves and the patient that can be a hindrance. And so if a group, for whatever reason, has not generated data in that particular format, I think it can be a lot of work, and if you're not used to it, there's a large learning curve. It's not something that is easy to just look up on a YouTube tutorial and say, "Oh, now I can do this process and upload my data." … And so any databases or repositories that have a system that intakes the data in a broader range of formats and with less mandated criteria is going to allow for easy sharing. (staff)

There was suggestion for NIH to look across other existing databases as a model (e.g. PubMed Central, National Health and Nutrition Examination Survey (NHANES) and SEER to learn how to best enhance data sharing by improving/ standardizing the format for data for easier submission and access. An example was to make the database analogous to PubMed Central, where investigators could submit their data and have a centralized system of formatting to standardize it.

It's a streamlined process with an easy interface, easy to use interface. And then not having to have the data in any particular form. So not having the depositor have to pre-process the data. Kind of like Pub Med Central, right? You deposit the manuscript however you have it and the National Library of Medicine puts it in the right format so everything is consistent. You see, that's actually a really good analogy. Submit it as you have it. It's clean, you submit it as you have it and then there's some centralized system, person, it could be artificial intelligence, I don't know what it is. But there's an algorithm that can then take whatever, whether it's a particular dataset format, like SAS or data or even a flat file or a series of flat files. And then however the variables are configured and that has a way of standardizing it and then making it available. (experienced PI, epi)

Clarity of submission/access process needed - Opportunities

The respondents mentioned opportunities to improve data sharing through addressing lack of clarity in the data submission and access process in a data repository (*Clarity of submission/access process needed-Opp,* mentioned by 14 of 37 respondents).

TABLE XLI: NUMBER OF RESPONDENTS WHO MENTIONED *CLARITY OF SUBMISSION/ACCESS PROCESS NEEDED*

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total (25)
Epidemiologists (14)	1	2	**3 (21%)**
Genomicists/Genetic Epidemiologists (11)	1	3	**4 (36%)**
Investigators (25)	**2 (33%)**	**5 (26%)**	**7 (28%)**
NIH Staff (12)			**7 (58%)**
TOTAL (37)			**14 (38%)**

Training and education around data sharing processes and practices were highlighted as areas respondents considered to be critical in addressing the issues with lack of clarity of the data submission and access process.

I don't think that we do a good job of understanding how long that takes, so even if you think that you budgeted for some time, I think it's not just programmer time or investigator time. I think there is something about sort of communication and whether it's education or another piece of it that I think is missing that could be helpful …. and maybe provide good examples or good tools or good templates for what makes good data sharing. (experienced epidemiologist)

From the perspective of a new investigator, there are basic things that would help with clarifying the data submission / access process that could directly contribute to improved sharing

among researchers. These are centered around administrative and technical support and training as essential, with provision of guidance documents, workshops and reminder emails. They acknowledged helpful communication and support from NIH staff, however, mentioned that more could be done by both NIH and the academic institutions to facilitate sharing

I think to be able to ask questions, you know, like someone who is very responsive at NIH in terms of helping us navigate through how we should do this is very helpful….I guess we just need to create like a flowchart, maybe, or some kind of process in place, but if it's kind of study-specific [then] what needs to be done when we start doing the research and how we will, in the end, post the data or share the data – would be great. …. Maybe kind of somebody [at academic institutions] who's responsible to help us, like a programmer time. Somebody who can help us navigate through how this data sharing will be done properly, legally or anything…. [NIH to provide] guidelines in terms of how we can do this step by step, maybe, I don't know – I don't know if there's a workshop or something for a young investigator or even the established one who doesn't know how to share data yet – and then point person to help us put the data set together – it's just like somebody who's available for us to ask questions then. (new PI epi)

This factor, *Clarity of submission/access process needed-Opp,* was also observed in the document reviews (Appendix G).

Expertise needed - Opportunities

The respondents mentioned opportunities to improve data sharing through addressing lack of expertise and knowledge among researchers and staff around the data sharing process (a priori code, *Expertise needed-Opp,* mentioned by 6 of 37 respondents).

TABLE XLII: NUMBER OF RESPONDENTS WHO MENTIONED *EXPERTISE NEEDED*

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total (25)
Epidemiologists (14)	1	0	**1 (7%)**
Genomicists/Genetic Epidemiologists (11)	1	2	**3 (27%)**
Investigators (25)	**2 (33%)**	**2 (11%)**	**4 (16%)**
NIH Staff (12)	[redacted]		**2 (17%)**
TOTAL (37)			**6 (16%)**

Addressing the expertise needed to enhance the sharing of data in a data repository would require training of the investigators and their teams, including their institutions' administrative / research offices (related to the factor, *Training needed-Opp*). One staff alluded to the need to increased level of expertise among program directors who may not be familiar with genomic data. Not doing this could potentially have a negative impact on how their grantees share data.

... But if you get a program director or program officer that doesn't often work with genomics, or never has worked with genomics, and they're coming into it the first time, and they have no idea how to be able to educate their grantees on the process. They don't know about it themselves. ..(staff)

When asked what was essential for successful data submission and data access, one staff mention the having people who are experts in bioinformatics as critical, in addition to other materials provided by NIH.

I mean, truly, having a help desk of people who work with the data, know how to, if they are accessible to users, then I think that's good. Tutorials or educational material is good. There are groups working on tools and pipelines to make this sort of data intake easier, so I do think that these are issues people recognize and are trying to improve. So I know that's certainly happening with NCI's Genomic Data Commons. (staff)

The quote below presents an opportunity for leadership to make change by increasing investment in education and levels of expertise in order to have a positive impact on data sharing among researchers and the public.

And our office is certainly going to hope to help play a role in educating the public in trying to alleviate concerns at all levels, whether that's the patient concern, whether it starts there; whether it's investigators doing it. They are going to have opportunities to actually have a career. But trying to balance out what everyone's concerns are and just educate around that and also implement the policies in such a way that we can help to alleviate some of that fear and concern and make it, like you say, a win-win for both sides. (staff)

Addressing repository capabilities - Opportunities

The respondents described opportunities to improve data sharing by addressing the limitations in the repository capabilities such as limited capacity, inefficiencies and ineffectiveness of the repository *(*a priori code, *Addressing repository capabilities-Opp;* mentioned by 17 of 37 respondents).

TABLE XLIII: NUMBER OF RESPONDENTS WHO MENTIONED OPPORTUNITIES FOR *ADDRESSING REPOSITORY CAPABILITIES*

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total (25)
Epidemiologists (14)	1	2	**3 (21%)**
Genomicists/Genetic Epidemiologists (11)	1	4	**5 (45%)**
Investigators (25)	**2 (33%)**	**6 (32%)**	**8 (32%)**
NIH Staff (12)			**9 (75%)**
TOTAL (37)			**17 (46%)**

One of the most recurring themes across the data, from staff and investigators, was that dbGaP is not user-friendly and is challenging for investigators to submit data; accessing data

from dbGaP was also considered problematic, though not as much as the submission process. To address this issue, the respondents suggested improving the design of dbGaP so it streamlines the processes, making it easy to use and to facilitate easier and faster submission and access. This includes enhancing the dbGaP website so that the information is easy to find, and a user testing of the databases or websites before it gets rolled out to ensure that any issues with the data format and structure which could impede sharing are addressed. Usability testing will need to involve the users or target audience early in the development / design stages of the database.

Automation of data submission processes and materials will not only enhance the database but reduce time and increase consistency, in addition to improving the way program level staff are able to track data submission / compliance with policy. Enhanced electronic systems to improve transparency and to track things in the system could provide a full picture of the entire process, which will give a good sense of how the program staff can better facilitate sharing. The development of tracking systems to determine the status of data submissions, where the datasets are in the process prior to submission, and to be able to know where delays are in the process on both ends can facilitate sharing. From a staff perspective, as described in the quote below, there are some challenges with implementation of the data sharing policy that could be addressed by creation of an automated data tracking system.

... not being able to track data that has been submitted and the status in order to follow up with investigators as needed. Rather, staff have relied on the investigators to let you know if they submitted their data or not. Staff would make this request via emails directly to the investigator (staff)

This would require new investment in "*engineering standards"* and technology, the right expertise for "*thinking through the experimental design and understanding the structure of the data collected*" for submission, designing the system and aiming for interoperability with other

databases. In addition, ensuring that the system is set up in such a way that vets access to data in order to protect privacy or confidentiality of data in the database was considered important.

... you don't want NIH necessarily designing these sorts of things, you want the sorts of people who, from Silicon Valley who design sites like Amazon and Google to do this, because I think again it is ... that's what you hear I think the most is- so the easier, the most logical it is, the more it is thought about in terms of in relation to the experimental design, because in some ways if you know the structure of how the data needs to be submitted it should make sense in terms of the design of the experiments of the data collection, so it a very smoothly interoperable and you don't have to reenter your data multiple times. (staff)

Other suggestions mentioned by investigators around opportunities to address the repository capabilities and enhance data sharing in repositories are illustrated in the following two quotes. This includes suggestion for a model that prevents data from getting scooped, where the "*data gatherers*" are separate from the "*data analyzer*"; and another example that refutes the "*one-size fits all*" model of dbGaP.

It is just horrible to be scooped, someone used your own data that you just killed yourself to collect. That is the problem. Which then, that's my question about whether we should have data gatherers and data analyzers. There's some large studies that are organized that way. If you think about [redacted] which is an HIV cohort, they have the data coordinating center and they have the other groups that may be more about generating the research ideas. So it is something to be thought about.(experienced PI, epi)

... instead of building a data model that that captures everything, you can build a whole series of lightweight data models that capture the key elements of this study or that study. And then at a high level instead of merging datasets, merge the data fields if you want to query across multiple datasets... provides a great deal more flexibility in terms of bringing in new and unexpected types of data (experienced PI, gen/gen epi)

Increasing efficiency through centralized systems or centrally coordinated systems for different aspects of data sharing was one of the resonating themes from staff and investigators. For example, centralizing data access requests across NIH would increase efficiency, and also for institutions to provide central support at the institution level to handle all data submissions, relieving investigators of the burden. This is illustrated in the two examples below.

I touched on it a little bit about kind of making the data access side more efficient by merging data access committees. And I think if NCI can do it with their three different, well actually going to be four different data access committees because they're establishing a fourth for Kids First, I believe.(staff)

Yeah, I think that there's an opportunity for improvement there. If an academic institution had some extra money or something that could add to education of researchers. Maybe a core data center that would do the submission would make it easier. That's my idea. If a large institute had multiple researchers that you might be able to justify that their core data center would be responsible for their submission. (experienced PI, epi)

Enhancing the interoperability of databases or data repositories, and increasing the ability of investigators to easily know what datasets are available that could help their research was considered important to improving data sharing. When asked what factors are essential for successful data sharing, one staff mentioned:

But when groups are able to see in some way, everything that's available, of course you can search by name within the dbGaP database, but I actually don't believe it's all that easy to search just by some different pieces of metadata. Do you want all studies that have exome data that are looking at samples that had x number of clinical factors? You might not even know everything that you're looking for. So more of the databases that incorporate tools that can help an investigator visualize data sets that they may have never even had on their radar that are going to be helpful to their research.. (staff)

This factor, *Addressing repository capabilities-Opp*, was also observed in the document reviews (Appendix G).

RQ2: OPPORTUNITIES – OTHER (EMERGENT)

Addressing data use, cost and value – Opportunities (Emergent)

The factor, *Addressing data use, cost and value* was one recurring theme that emerged from the data that described the gap in knowledge among the respondents on the use, benefit and value of data shared in a repository, and the cost for sharing data through a repository (emergent code, *Addressing data use, cost and value-Opp*, mentioned by 9 of 37 respondents).

TABLE XLIV: NUMBER OF RESPONDENTS WHO MENTIONED *OPPORTUNITIES FOR ADDRESSING DATA USE, COST AND VALUE*

No. Respondents	**New Investigators (6)**	**Experienced Investigators (19)**	Total (25)
Epidemiologists (14)	2	1	**3 (21%)**
Geneticists/Genetic Epidemiologists (11)	0	1	**1 (9%)**
Investigators (25)	**2 (33%)**	**2 (11%)**	**4 (16%)**
NIH Staff (12)	[redacted]	[redacted]	**5 (42%)**
TOTAL (37)	[redacted]	[redacted]	**9 (24%)**

The respondents including staff with experience in policy, leadership and implementation suggested that these be addressed through targeted analyses. In particular, one of the new investigators mentioned the importance of doing an analysis to have a better understanding of what the data in the repositories have been used for and the types of research that have emanated from use of such data. This will inform how researchers could use the data accessed from the repositories for their own research and potentially provide some perspective on the value of the data in the repositories. It could also help justify the amount of effort involved in sharing data in databases or data repositories.

Maybe having a workshop and also some ideas about what the data has been used for, various research. Maybe highlight some important findings that people have done using these shared resources. I think that would be of interest to us, to see, "okay, what has been done, what's out there, what findings", that kind of stuff, and for like interesting findings. You know, researchers love to read, or give us some ideas about [how] we can actually use this repository data to do this kind of research, innovative, creative, that kind of stuff. (new PI, epi)

There are papers that have been published on the use of data in NHLBI's BioLINCC daa and biospecimen repository (Coady et al, 2017) but this respondent was not aware of this

resource and requested information such as these be shared with funded researchers often. This is an opportunity for NIH to increase communication and awareness of existing resources at NIH.

Related to the value of data shared in the data repository, an experienced genomicist/genetic-epidemiologist and staff both asked about the use of data in the repository, and the value of sharing data via repository, respectively. This is an opportunity to further explore in future studies, data use and NIH investment in data sharing.

…. So what I don't know is, are these data used maximally? I know what the intent is, but is this data sharing method successful at actually having people [maximize] that [data]? (experienced PI, gen/gen epi)

Because, you know, there's some data that may be generated and it may be useful to share for a short period of time and then maybe after that, it's not and then what do you do? So as I said, I think trying to figure out the value of sharing the data, as well as how, what the metrics are to assess that value are probably important. How do you measure successful data sharing? (staff)

One staff indicated that a way to appreciate the value of data sharing and further enhance data sharing among researchers would be through a qualitative research / evidence-based analysis to show that the public understand benefit and risk of data sharing, their consent is truly informed and that they support data sharing. This provides an opportunity for more communication about benefit and risk of shared data, ethical implications, similar to mentioning de-identified data.

I think one thing that would encourage people to share is if there were a set of qualitative research and evidence that the public and our participants do understand what data sharing really is and they understand their informed consent. Because we can call it informed consent, but we don't have necessarily know if it's really informed, what are the ethics involved. ... If participants have fewer concerns than what we assume as investigators, maybe the investigators will sort of come around. So if the NCI or anybody else has actual like data to support that participants and the public are pro-data sharing in public or controlled-access databases, I think

we as scientists would feel better about it… Everybody needs to understand both the benefits and the risks. (new PI, epi)

The idea for a cost-benefit analysis to be done was an important suggestion that came up during the interviews. A couple of the staff respondents echoed similar comments around challenge with finding the right balance between the cost of data sharing for funding agencies, institutions and their investigators, and the value or benefit gained from sharing data in a way that promotes science in the most effective and efficient manner. This further emphasizes the complexity of this phenomenon.

And I think one of the interesting questions from my point of view is what is the, you know, is the juice worth the squeeze ratio when it comes to data sharing? Because I think on one hand there is broad agreement that we need to do more collectively as an enterprise, we need to do more in data sharing. I think there is also general agreement that you don't want to share all the data, because that's kind of crazy. .. But where is the sweet spot in the middle that really allows data sharing in a way that generally facilitates science? … .And so making sure that the value and what you get out of data sharing is commensurate with whatever the cost or burden that is going into it is super important to how people feel about data sharing (staff).

The example below also describes the need for a cost-benefit analysis that assesses the value of the data submitted and used for secondary research – and how the data is being used. This may help investigators and staff understand if this is worth the investment, and also to readily identify which of the datasets are the most used and supported.

Additionally, I think there is a cost benefit analysis that still needs to be done with regard to, you know, "your data meets a certain criteria, therefore you shall share". It doesn't necessarily take into account if that data is actually valuable for secondary research purposes. So, we're mandating that the data be made public, but not necessarily accounting for whether it's being used. And how to cull that data if it's not so that we're not wasting resources. I think there could be some sort of analysis done to say, "These are the datasets that are often used. These are the

types of datasets we should focus on, prioritizing submission for, and figure out a way to kind of meet in the middle of cost and value but leaning towards value overall. (staff)

Another suggestion for an empiric research was to look at the impact of data sharing on investigators' development in terms of addressing the barriers related to cultural concerns and career of investigators

You know, I think in that regard, people who might be willing to, you know, do some research to take some quantitative looks at what effect if any does it have on investigators' development. How often are individual groups or, you know, which I guess it's really hard to tell, but some more empiric research on looking at these concerns if we are at a point that we have data and metrics, appropriate metrics to be able to begin to measure that so that the fear can be countered with information - with data, exactly, so that we can [unintelligible] the scientists what they understand and relate to (staff).

B. Document Reviews

Document #1: *Compiled Public Comments on NIH Request for Information: Processes for database of Genotypes and Phenotypes (dbGaP) Data Submission, Access, and Management (NOT-OD-17-044)*[36]

The NIH Request for Information (RFI): dbGaP, was published February 21, 2017 – April 7, 2017 in the NIH Guide, to solicit feedback from the public on "any opportunities or challenges well as potential areas and opportunities to improve understanding, efficiency, or transparency of the processes associated with the following topic areas: 1) dbGaP Study Registration and Data Submission; 2) dbGaP Data Access Request (DAR) and Review; and 3) Policies for the Management and Use of dbGaP Data - Alternate Models; Benefits and Risks w genomic study summary statistics**;** and Clinical Use of Genomic Research Data Maintained in

[36] https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-044.html

Controlled-Access in dbGaP. The comments from the public were compiled into a single document and posted on the public website.

The most relevant topics to this dissertation research were the first two. Therefore, data analysis was conducted on the comments from respondents on challenges and opportunities to improve data submission and data access in dbGaP. The same coding schematic and definitions used for the interview data was applied to this data for consistency in the analysis. The reference to challenges by the respondents was considered as barriers, and "opportunities to improve" was considered as facilitators or opportunities to improve or enhance data submission and data access in dbGaP.

There were 43 individuals who responded to the RFI on dbGaP from different types of organizations such as universities, health care organizations, non-profit organization, etc., and with different roles and levels of experiences with dbGaP. Random sampling was used to select a sample of respondent for this research. The list of respondents was recreated from the pdf document available on the website, in MS Excel, and sorted by institution name listed in alphabetic order. The criteria used for selection was that no more than one from same institution would be selected, and included a balance between self-identified novice and experienced dbGaP users to provide different perspective in the analysis, and include respondents that provided a response to at least one of the questions. Respondents were selected by every 4^{th} person and then by every 3^{rd} to adjust for the different levels of experience. Respondents were excluded if they did not provide a response to at least one of the two questions of interest for this study, and if they did not indicate their level of experience with dbGaP because level of experience was going to be used in the analysis / triangulation of the data, similar to the method used for the interview

data. After the inclusion and exclusion criteria were applied, 10 participants with different levels of experience and use of dbGaP were selected.

In describing their levels of experience with dbGaP, 4 of the 10 respondents self-identified novice users (i.e. used once or only a few times), 4 of the 10 respondents self-identified as experienced users (i.e. used many times over the course of several years), and 2 reported that they had never used dbGaP but did provide their comments on the questions. In describing the primary purpose of dbGaP use, 2 of the respondents used dbGaP primarily for study registration/data submission only; 5 of the 10 respondents used it for data access/download only; 2 of the 10 respondents used it for both data submission and data access/download; and one respondent used it for browsing unrestricted access study information. The different roles and experiences captured in this sample, will provide rich and diverse perspectives that will either corroborate or contradict the findings from the interview data, or contribute new knowledge and insights on the challenges and opportunities for enhancing data sharing through data repositories like dbGaP.

The analysis of the data from this sample showed the most prevalent construct was TECHNOLOGICAL INFRASTRUCTURE, followed by RESOURCES, then REGULATORY POLICY AND LAW (Appendix G). These will be described in more detail below, highlighting the facilitating or hindering factors, as well as opportunities as they relate to data submission and data access, and described by level of experience and primary purpose of dbGaP use.

A. POLICY

Addressing Privacy concerns – Opportunities

One novice user mentioned some ideas for addressing issues that may arise when institutional certifications, which are tied closely to the informed consent forms and hence the privacy of participants, can no longer be provided by a University. Centralizing informed consent review could increase efficiency and improve the data submission process.

The required Institutional Certification ensures consistency of data sharing with the participants' informed consent. However, obtaining Certification is problematic for completed studies where IRBs are no longer active. To circumvent this, dbGaP should consider informed consent review by the appropriate NIH IC's IRB, or the NIH OHSRP for non-NIH funded studies. Another consideration is to establish a central NIH ethics review board to streamline submission. Prospectively, the single IRB model proposed by the revised Common Rule will improve data sharing efficiency. (Novice user)

B. RESOURCES

Administrative / Technical Resources – Facilitator and Barrier

There were 2 novice and experienced users who mentioned that administrative staff support helped facilitate the submission of data in dbGaP. Two novice and experienced users also mentioned that the process of accessing data in dbGaP, including adhering to IRB requirements was too time consuming, posing an administrative burden on the investigator. They also mentioned that it took too much time and effort, requiring the right expertise such a programmer or bioinformatician to help improve data sharing among investigators. This was similar to the findings in the interview data.

the process is very smooth and logical. Especially the dbgap rep. are very helpful and informative in response to our questions (novice user)

The requirement for review by institutional officials is a huge waste of time, since my IOs are busy with other tasks. Takes hours/days to access one study. Simplify the process! (novice user)

Training needed - Opportunities

Comments from one of the respondents who had never used dbGaP indicated the need to provide training and communicate expectations and requirements around submission and access in dbGaP. This was also echoed in the interviews.

Do the various funding institutes at times have their own additional requirements? We received a request from one institute in the past to designate a particular "consent category" (from a list provided by the IC) in addition to the information provided on the standard Institutional Certification form. It would be helpful to either have a standard process across all ICs, or if there are differing requirements, a clear way to communicate those requirements to PIs and IRBs.

C. TECHNOLOGICAL INFRASTRUCTURE

There were 6 of the 10 respondents who mentioned the construct *TECHNOLOGICAL INFRASTRUCTURE,* as important in data sharing through dbGaP. The different factors under this construct are explored in more detail.

Analytic data complexity – Barrier

This code *Analytic data complexity-Bar* was mentioned by two novice users where described that the challenge with accessing data in dbGaP as being related to the analytic data file and the capability of the database - the file names are large that hinder access and also slow down the download of data from the database.

since all downloaded data are scripted, sometimes it is technically difficult for general researcher to download or open the files after downloading. Some of the files also have very large names that one time, windows (windows 7) was not able to work on the file or re-name it. I have to relocate the file and re-name it in another folder. (novice use)

Addressing Analytic data complexity - Opportunities

This code *Addressing analytic data complexity-Opp* was mentioned by one respondent. To address the challenges with non-standardized data formats in dbGaP, one novice user mentioned the need for data standardization and formatting for easier submission in the database. This was a similar sentiment echoed in the interview data.

Different research institutes generate data using different sequencing machines in varying data and metadata formats. To maximize reuse, dbGaP should utilize common data elements to enable data harmonization and standardization, similar to the Global Alliance for Genomics and Health community. To ensure better data discovery, access and citation, dbGaP should consider Digital Object Identifiers for all data assets. For distributed computing of large scale data, dbGaP should consider tools that allows maintenance of data structures in distributed dynamic and/or unreliable (e.g., user desktop) environments.

Lack of clarity of submission and access process – Barrier

The code, *Lack of clarity of submission and access process-Bar*, was used to describe the concept of the lack of clarity in the submission and access process in dbGaP. This was mentioned by 2 respondents, one experienced user and one novice user who used dbGaP to access data for browsing. The example below highlights the lack of clarity in the data access request process as it relates to where data sits or is housed. This was a similar sentiment across the interview data.

In my experience, dbGaP has been the place to request and obtain permission to access controlled data, but the data has always been somewhere else -- eg at the TCGA DCC, at CGHub, or at the GDC (since last year). So the notion of "downloading data from dbGaP" seems odd to me, although I have seen buttons and menus and such on the dbGaP website related to browsing and downloading data. I think that it would probably be helpful to clearly distinguish between DAR functions and where the data actually can be accessed. With the move towards the cloud, the data may also be available in multiple locations and a mechanism to be able to reference/find data in a variety of locations would be helpful -- eg if a user is looking for a particular WGS bam file, it would be useful to know if it exists in Google Cloud Storage, Amazon S3, etc in addition to at the "official repository" which might be the GDC in Chicago.(experienced user)

Clarity of submission/ access process needed - Opportunities

The code, *Clarity of submission/access process needed-Opp,* was used to describe opportunities for addressing the lack of clarity in the submission and access process in dbGaP. There was mention of the need for increased clarity in both the submission of data in dbGaP as well as the access of data from the database. This was echoed by 9 respondents including 3 novice users and 4 experienced users of dbGaP who primarily use dbGaP for data access only, for both access and submission, and for browsing; and 2 users who never used dbGaP. One of the examples provided is similar to a theme in the interview data around the lack of clarity with the data access process, lacking uniformity across the institute. One of the experienced users mentioned:

Each DAC seems to act independently, without consistent turnaround time or consistent consideration of application content. A more uniform guideline on how decisions are made would reduce confusion and improve review efficiency.(experienced user)

From the novice user perspective, which corroborates that of the experienced user described above, the submission process could be more efficient and suggested streamlining the submission and access process, especially for non-NIH investigators who plan to submit their data in dbGaP and those who don't have electronic Research Administration (eRA) accounts. This could be an opportunity to look at other models to help increase efficiency in the process and promote broader sharing of data, including increasing access to junior investigators.

To address these challenges, dbGaP should consider additional methods for registering a submitter (institutional email/ORCID account, etc.), and facilitate submissions without requiring approval by NIH IC Officials, but requiring non-NIH submitters to link the submission to a peer reviewed publication. Not mandating the eRA account and NIH IC sponsoring requirements will improve efficiency and expand dbGaP considerably. If de-coupled from NIH IC, submitters from non-NIH institutions should bear the administrative costs for submission as a donation to NLM…. access to this tool also requires an eRA login. *This limits graduate students, post-doctoral fellows, and other young investigators to those affiliated with a PI who is an eRA account holder (novice user)*

One interesting comment from an experienced user related to access was the need to streamline the approval process for data access requests so that there's one centralized approval process. This could increase efficiency in the access process.

There are so many committees. I can understand the need that a specific committee is needed to review one dataset, but it would be helpful that the approval process is centralized, especially for a continuing use of the several datasets. In other words, one oversight committee may be able to review and approve a renewal request based on the annual report.(experienced user)

Another interesting point made by one of the respondents who had never used dbGaP had to do with the need to keep study participants in the loop on their results and on the project status. This is all about communication of results and engagement of study participants as a means to enhance data sharing. This could be done through including this as part of the data submission and access process for dbGaP.

Repository capabilities – Facilitator and Barrier

The comments around the capabilities, capacity, efficiency and adequacy of the repository was one of the most prevalent themes among the respondents sampled for this study. There were 4 novice and experienced users who mentioned these as barriers to submitting and accessing data in dbGaP. The codes, *Repository capabilities-Fac* and *sub-optimal Repository capabilities-Bar* were also mentioned in the interview data.

Poorly designed process. Needs to be completely revised and redone to simplify the process (novice user)

Addressing repository capabilities - Opportunities

There were two experienced dbGaP users with some suggestions for how to improve dbGaP functionality, and hence the increasing efficiency in the process for submission and access from dbGaP. For example:

I would suggest an option to add studies, variables, and datasets to a cart or basket while browsing. I often find myself navigating between two copies of dbGaP in different browsers: one for browsing individual studies or variables and another for either selecting those variable datasets from those studies in the file selector for download. It would be great to be able to add items to a cart and then have ready access to just those items from the download page

Document #2: Compiled public comments on NIH Request for Information: Strategies for NIH Data Management, Sharing and Citation (Data Sharing Strategy Development) (NOT-OD-17-015)[37]

The NIH Request for Information (RFI): Strategies for NIH Data Management, Sharing and Citation, was published Nov 14, 2016 - January 19, 2017 to solicit public comment on "data management and sharing strategies and priorities in order to consider: (1) how digital scientific data generated from NIH-funded research should be managed, and to the fullest extent possible, made publicly available; and (2) how to set standards for citing shared data and software." The comments from the public were compiled into a single document and posted on the public website.

Specifically, in Section I of the RFI, the *Data Sharing Strategy Development* section, NIH asked the public to comment "on any of the following topics to help formulate strategic approaches to prioritizing its data management and sharing activities: *The highest-priority types of data to be shared and value in sharing such data; The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications; Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers; and any other topics respondents recognize as important for NIH to consider."* In Section II, NIH asked for

[37] https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-015.html

comments on a variety of topics related to the Inclusion of Data and Software Citation in NIH Research Performance Progress Reports (RPPR) and Grant Applications.

The topic in this RFI that was the most relevant to this dissertation research was the topic on *Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers.* Comments from respondents were analyzed using the same coding schematic and definitions used for the interview data. The reference to "mechanisms to overcome these barriers" was considered as opportunities to improve or enhance data sharing in NIH funded research.

There were 95 individuals who responded to the RFI from different types of organizations such as universities, health care organizations, non-profit organization, for-profit organizations, etc., and with a variety of domain of research most important to respondent of their organization. Random sampling was used to select a sample of respondents from this group for this research. The list of respondents was recreated from the pdf document available on the website, in MS Excel, and sorted by institution name listed in alphabetic order. The minimum inclusion criteria used for selection was that no more than one individual from same institution would be selected, and respondents must provide a response to at least the question on Barriers in section I, given this is the only question to be used in the analysis.

Respondents were selected by every 4^{th} person. After the inclusion and exclusion criteria were applied, a total of 17 respondents from different organization and diverse domains of research areas of interests were included in the analysis. Some of the domains of research areas included basic sciences, sleep, public health, health economics, cardiovascular disease, etc. The analysis of the data from this sample showed the most prevalent constructs were

TECHNOLOGICAL INFRASTRUCTURE followed by RESOURCES, POLICY then CULTURE (Appendix G). These are described in more detail below.

A. CULTURE

Career concerns - Barrier

The code that captured the concept of career concerns related to data sharing was *Career Concerns-Bar*. This was mentioned by 2 of the 17 respondents to the RFI on NIH Data Sharing Strategies. The respondents echoed a similar theme with the interview data.

On the other hand, data sharing reduces investigator advantages when applying for grants and limits protection of publication opportunities for the research team, students and colleagues.(biomedical research)

Reward structure changes needed - Opportunities

The code that captured the concept of opportunities for addressing the lack of rewards at the institution or organizational level was *Reward structure changes needed-Opp*. This was mentioned by 5 respondents who were mostly in the biomedical research field and echoed a similar theme with the interview data. They mention allowing the opportunity to incentivize data sharing through funding which is not currently happening to the extent desired, and with equal recognition of publication of data as that given to manuscript publications. Changing the reward structure at institutions could change the culture of sharing research data at institutions. This theme converged with the interview data.

Rewarding and promoting best practices for data publication as much if not more than article publication would be ideal (biological sciences)

B. POLICY

Enforcement of policy – Facilitator and Barrier

The codes used to describe the concept of enforcement of data sharing policy as a facilitator and barrier respectively were *Enforcement of policy-Fac (*mentioned by 2 respondents) and *Inconsistent enforcement of policy-Bar* (mentioned by 1 respondent). The respondents echoed a similar theme with the interview data, indicating that data sharing policy promotes data sharing because data is tied to publication (biomedical sciences). In addition, a challenge with sharing has to do with the lack of consistency in the implementation and interpretation of the polices across NIH, academic institutions and other organizations.

Inconsistent data sharing policies between Institutions, funding agencies, and publishers (population health/economics)

Privacy concerns – Barrier

There were two respondents in biomedical research, who described their concerns with participant confidentiality and potential breach in privacy with data sharing in data repositories. This is a barrier to researchers sharing data and was also expressed by respondents who were interviewed for this study. The code used to describe this factor is *Privacy concerns-Bar*.

Several constituents have proposed that secondary research conducted with patient-level data should be independently reviewed for scientific merit as a condition of access. This point emphasizes again protection of risk to research subject confidentiality where identifiable data necessary for analysis, or where there is potential for re- identification. (biomedical research)

Addressing privacy concerns - Opportunities

There were 2 respondents who suggested ideas for addressing some of the concerns related to patient / participant privacy which tend to hinder sharing of data because of the potential risk

with breach in patient privacy in research studies. The code used to describe this is: *Addressing privacy concerns-Opp*. This same factor was observed in the interview data.

Crafting protocols for addressing privacy and related issues can be a time- consuming task for the PI and associated research team on a project, and this can draw resources from other aspects of the research. If the NIH could specify acceptable common protocols, and offer a "safe harbor" to researchers who used any of these protocols, that would reduce the burden of creating archival data sets. (population health/economics).

C. RESOURCES

Inadequate administrative/technical resources – Barrier

The code *Inadequate Administrative/technical resources-Bar* was used to describe the comments of 3 respondents who mentioned administrative and technical resources not being available and the burden on investigators as hindering data sharing, respectively. This was also observed in the interview data.

For data sharing to advance, research sponsors and institutions must commit resources. It is not clear that the public or political leaders … appreciate the additional burden and cost of creating usable shared data resources.

Administrative/technical resources needed - Opportunities

The code, *Administrative / technical resources needed-Opp*, was used to describe administrative and technical support that could help with alleviating challenges experienced by investigators with data sharing. The goal is to make it easier for them to access or submit data in repositories. This was mentioned by 3 respondents and observed in the interview data.

For data generators, an important impact of increased reporting and storage of data is the changes to routine and workload that open data mandates have on individual researchers and institutions. … suggests governing bodies may reduce the burden for individual researchers

by: 1) promoting curated, redundant and maintained collaborative IT and software solutions and supporting services offering economies of scale. (biomedical research)

Financial resources – Facilitator and Barrier

There were 6 respondents who mentioned that funding was a key factor for successful data submission. They also mentioned that researchers should be encouraged to include in their grant applications funding for data sharing. This was also echoed by the respondents interviewed for this study. The code used to capture this concept was *Financial resources-Fac.*

Commitment of funds are required for data preparation and curation of databases … (cardiovascular health)

In the population field it has been standard practice to share data for many years, and PAA believes that this practice should be expanded to other fields. In order to do this [share data] it is important that researchers are advised and encouraged to include funds for the preparation of data sets in their budgets … (population health)

There were also 6 respondents who mentioned that the lack of adequate funding hinders data sharing by investigators. This was also echoed by the respondents interviewed for this study. The code used to capture this concept is *Inadequate Financial resources-Bar.* The example below highlights some of the major barriers to data sharing, one of which is funding. This is a recurring theme across data sources and groups of respondents participating in this study.

The biggest barriers are funding support, data curation, and incentives to the data creators to store their data.(biomedical research)

Financial resources needed - Opportunities

Respondents emphasized the need to address the financial challenges related to data sharing costs. The code used to describe this concept, mentioned by 2 respondents is *Financial resources needed-Opp.* This was also observed in the interview data.

An interesting suggestion was to have more grant funding dedicated to support infrastructure as opposed to the scientific research aims. This is given the comments by respondents that a majority of the burden associated with sharing data has to do with the data curation, preparation and infrastructure to support it, but that NIH should consider implications for sustainability of funded infrastructures for data sharing in the long-term.

Most databases … are funded through a combination of multiple research grants, which have a typically short duration, low success rates and by their nature focus primarily on the development of new research activities rather than service provision. … recommends the development of more fit for purpose funding schemes developed with the intention of supporting the operations of services rather than purely new research activities….. and to understand the challenge of the long-term resource implication that comes with data infrastructures. (biological sciences)

Leadership support – Facilitator

Similar to other data sources, the respondents to this RFI (3 of 17) alluded to the support from institutional / organizational leadership e.g. NIH or academic institutions, as an important factor in facilitating data sharing among investigators. This is through the provision of resources – administrative, technical or financial resources needed. The code *Leadership support-Fac* was used to describe this concept and was also observed in the interview data.

For data sharing to advance, research sponsors and institutions must commit resources. It is not clear that the public or political leaders, who increasingly support or call for data sharing (and other "transparency") appreciate the additional burden and cost of creating usable (biomedical research).

Training needed - Opportunities

There were 4 respondents who mentioned of the need for training, education and communication both internally and externally on available tools, software, platforms and systems to help with data sharing. Developing and enhancing existing training guides or materials and

increasing awareness of existing materials will increase knowledge, skills and expertise and overall, help with the different stages of data sharing. The code used to describe this was *Training needed-Opp* and it was observed in the interview data.

Outreach to the research community about the importance of data access, code access, and data archiving would improve the diffusion of best practices. NIH could develop training modules, like those about responsible conduct of research and human subject protection, to inform researchers about archiving principles and options, and to highlight ways to create reproducible data sets and to maintain a code archive. Presentations could be included in professional meetings to further raise the visibility of these issues (population health/ economist)

D. TECHNOLOGICAL INFRASTRUCTURE

Analytic data complexity – Facilitator and Barrier

The codes, *Analytic data complexity-Fac* (mentioned by 3 respondents) and *Analytic data complexity-Bar* (mentioned by 1 respondent) were used when comments were made around the importance of data documentation, formatting, etc. for submission of data in data repositories. This data corroborates information obtained from the interviews in this dissertation research. They addressed the need for proper data documentation early in the process as a good way to reduce administrative burden and also improve quality of data shared, as well as tools in place to facilitate this.

In terms of preparation necessary to share data, software packages and platforms now exist that allow researchers to document this step in the research process as it occurs, so that once the data collection is complete, all relevant metadata are automatically created. For NIH policies on data sharing to be successful, researchers must be trained to use such software, so that the burden on PIs of engaging is data sharing is minimized… Documenting decisions throughout the research process, rather than after publishing results, significantly reduces the burden and results in higher quality documentation. This requires education and guidance for researchers at the beginning of new studies.

In addition, the indication that researchers were not documenting their data or providing metadata along with the submission of their datasets is one that is echoed across the data sources.

Much data used for health research has inadequate metadata. Producing the metadata needed for archiving can be expensive, (population health)

Addressing analytic data complexity - Opportunities

There were 4 respondents who discussed opportunities for addressing some of the challenges researchers face with the analytic data that is submitted in the repositories. The example below mentions one of the ways for addressing the issue of data format, which is not just standardizing the data format but also standardizing written informed consent for ease of data sharing and access. The code used to capture this concept is *Addressing Analytic data complexity-Opp.*

IRB/ethics-- standardized written informed consent for data deposit for all prospective and registry studies. Standardized clinical data format--NIH should adopt or help organize an established format for clinical data. (biomedical sciences)

In addition, more indexing to facilitate access was recommended to facilitate identification and access and reuse of data cited in articles.

We recommend more extensive subject-matter indexing within and across repositories to facilitate discovery and reuse. … Today, investigators typically learn about relevant datasets by word of mouth and then search in repositories using the dataset author's name or a specific DOI. (Digital Object Identifier). As data and articles become more connected, investigators will also be able to identify a data object for re-use through an article citing those data. DOI enables this by linking data to articles. (Program evaluation)

Lack of clarity of submission/ access process – Barrier

The lack of clarity of the submission/access process of data repositories is a strong theme across the data sources. Examples were mentioned in other data sources, including the interviews. The code used to describe this concept, mentioned by 2 respondents was *Lack of clarity of submission/access-Bar*.

In my experience, dbGaP has been the place to request and obtain permission to access controlled data, but the data has always been somewhere else -- eg at the TCGA DCC, at CGHub, or at the GDC (since last year). So the notion of "downloading data from dbGaP" seems odd to me, although I have seen buttons and menus and such on the dbGaP website related to browsing and downloading data. I think that it would probably be helpful to clearly distinguish between DAR functions and where the data actually can be accessed. With the move towards the cloud, the data may also be available in multiple locations and a mechanism to be able to reference/find data in a variety of locations would be helpful -- eg if a user is looking for a particular WGS bam file, it would be useful to know if it exists in Google Cloud Storage, Amazon S3, etc in addition to at the "official repository" which might be the GDC in Chicago.

Clarity of submission/ access process needed - Opportunities

The lack of clarity in the submission and access process of data repositories is a strong theme across the data sources. Examples were mentioned in other data sources however, there were two respondents who mentioned interesting ideas for addressing this barrier to data sharing. The code used to describe this concept, mentioned by 7 respondents, was *Clarity of submission/access process needed-Opp.* This factor was also observed in the interview data

… *it would be helpful for investigators to use workflow tools which are embedded within the research data life cycle. Such tools would serve to prompt investigators to do the things necessary to make data as useful as possible and also make these activities easier. We find it instructive to think of GitHub as an example of such an embedded workflow tool. Although GitHub does not itself house data (relying instead on Zenodo), it can be used for code creation, testing, and version control throughout the lifecycle of software development – ie, it is embedded in the processes preceding the act of sharing of code. This interplay between the GitHub platform and the everyday tasks of coders could be envisioned for data producers, thereby promoting the sharing of the resulting dataset. … (population health/economics)*

Lack of expertise - Barrier

The limited or lack of knowledge or expertise among researchers with different aspects of the data sharing process is a recurring theme in this study. The code, *Lack of expertise-Bar*, was used to capture this factor, and was observed in the interview data. This example below from one respondent illustrates how the lack of expertise is a barrier and the implications in data sharing.

Individual researchers do not generally have the experience or expertise to document, store, and disseminate the data that is collected in projects. Imposing this burden on individual

projects is burdensome and inefficient and leaves data spread across a wide variety of locations in inconsistent formats. Individual researchers may also be inexperienced in how confidentiality can be protected while providing maximum possible access to the data… (population health)

Repository capabilities – Facilitator and Barrier

The capability of a repository or database is critical to successful submission of good quality data. There are costs associated with large data files and therefore need to be taken into consideration in terms of the repository capacity to store different types of large files, the impact on the user end during the download process. There were 4 respondents who mentioned this factor as a barrier (*sub-optimal Repository capabilities-Bar),* and 1 respondent who mentioned this factor as a facilitator *(Repository capabilities-Fac*). The exact same factors as barriers and facilitators were observed in the interview data.

[dbGaP] not user friendly - too complicated and rigid and too much burden on submitters (scientific researcher)

Access - not easy or user-friendly (red tape with process; 'not clear which subset needed to request / download') (scientific researcher)

Addressing repository capabilities - Opportunities

With the data deluge of the 21st century, more sophisticated options for data storage and analysis, for example in the cloud, will help with the capacity and functional issues of existing data repositories. See below an example that illustrates this There were 5 respondents who provided suggestions for addressing this issue. This was also consistent with the interview data. The code used to describe this is *Addressing repository capabilities-Opp.*

Subject-level repositories should be chosen whenever possible due to well-established curation practices. .. Since storage costs can be significant with large datasets, we recommend proposals to be accompanied by data management plans that highlight the size of the total data collection and the biggest expected size of a single file. For files of size greater than 2 GB or collections greater than 20 GB, we recommend requiring that researchers discuss their datasets with their repository of choice to confirm the deposit can be accommodated and to determine the

projected costs of its preservation over time. For longitudinal studies, it is important to carefully reflect on the variables that will be studied. We recommend those variables to be listed in a data management plan prior to the start of the project. Simple spreadsheets are not effective tools for managing these kind of data due to the number of entries over time and the lack of version control or logging. For such studies, we recommend databases be used, and associated costs be accounted for. (biological sciences)

Documents #3, #4 and #5: EGRP internal reports on evaluation of data sharing in epidemiology cohorts

Three separate but related internal reports on results from an external evaluation conducted between 2015 and 2017 of data sharing practices in EGRP funded cohorts were reviewed and determined to be highly relevant to this dissertation research study. The evaluations were conducted by a federal contractor through interviews with cancer epidemiology cohort PIs (experienced and early stage investigators) and their administrative/research staff. These documents were reviewed and coded separately. Given the number of overlapping codes across the reports, the findings will be grouped for a more focused discussion of the analysis.

- *Report #1 - January 2016. EGRP Internal Report #1 – Evaluation Proposal of the NCI Epidemiology and Genomics Research Program (EGRP) Cohort Studies' Data Sharing Practices*
- *Report #2 - Jul 2016. Findings from NCI Epidemiology and Genomics Research Program (EGRP) 9 Cohort Interviews on Data Sharing Policies and Practices*
- *Report #3 - April 2017. Findings from Interviews with NCI Data Requestors on Epidemiology and Genomics Research Program (EGRP) Cohort Data Sharing Policies and Practices*

A. CULTURE

Career concerns – Barrier

When the cohort teams were asked about their perception of central repository for epidemiology data, some of them mentioned that they were concerned because they didn't know how their data would be used of potential misinterpretation and misuse and mischaracterization of their data. In addition, they expressed some fear that if they shared their data in a repository, others may analyze and publish on their data before they've been able to analyze it, which could be threatening to their career advancement.

B. POLICY

Data sharing definition - Barrier

The cohorts mentioned that they shared data both within and outside of their institutions and / or network, although sharing was mostly done with their network. This corroborated the findings in the interview data where investigators mentioned that sharing was mostly within their departments or institution. This implies that the collaborative model of sharing was more in line with their cultural norms as it relates to the definition of data sharing

Privacy concerns – Barrier

Some of the cohorts mentioned concerns with potential breach of patient privacy if the data is deposited in a central repository. Some of the cohorts did not feel comfortable sharing their de-identified data in a repository. In addition, there was some differences among cohorts in what the IRB approves in terms of the consent form and data sharing. Some reported that their consent forms as is would allow for sharing because NCI was named a potential data recipient. In contrast, others mentioned that historically, their IRBs would not allow for their participant

data to be shared in a repository unless explicitly stated, or otherwise may require reconsenting of the participants.

C. RESOURCES

Administrative / technical resources – Facilitator

The cohort research teams mentioned that they would be willing to share data if they had the administrative support and if the data sharing process wasn't too burdensome.

Administrative / technical resources – Barrier

There was a consistent theme in the documents on the administrative burden of the data sharing process, such as uploading data into dbGaP as being time consuming. In particular, one of the cohorts mentioned that with the collaborative model of data sharing, the cohort PIs and their teams spend time with data requesters on the analysis of their data and review of publications, to ensure that there are no misinterpretations of their cohort data. This highlights their perception or experience with data sharing as being time intensive.

Inadequate Financial resources - Barrier

The PIs indicated that depositing data into a centralized repository would require resources not budgeted for in their grants. Some were aware that they could include a budget for data sharing in their grants, but mentioned that it was challenging for them to estimate what the cost of sharing data would be. Specifically, "*predicting the number of data sharing requests for any given year and the increasing costs associated with data sharing over the long term*" (report#1) was mentioned as challenging. This was interesting in that some of the investigators (and staff) interviewed as part of this dissertation study also did not seem to be aware that investigators could include data sharing costs as a line item in the budget. On the contrary, this

report shows that the cohort research teams were aware, just that it was challenging to come up with an estimate.

D. TECHNOLOGICAL INFRASTRUCTURE

Analytic data complexity - Barrier

In one of the reports, cohort research teams were asked how the broader epidemiology community viewed data sharing, and their response was that it was very complex, more so than other types of data. In addition, they mentioned that compared to genotype data, it was more difficult to harmonize phenotypic data, and their reluctance to depositing in a central repository like dbGaP was based on the fear that others may not understand the study design and nuances around collection of epidemiology data, which is important for analyses. This was also observed in the interview data.

Lack of clarity of the submission / access process – Barrier

Two of the cohort teams mentioned that uploading data into dbGaP was not only time consuming but that it was not easy to understand the process of doing that. This was also observed in the interview data.

Sub-optimal repository capabilities – Barrier

The cohort teams when asked about their experience with using the database, dbGaP, mentioned that dbGaP was not user-friendly and not efficient given the long delays with the process on the dbGaP's staff end. One comment around the use of data repository was that while it was good to share epidemiology data as broadly as possible, "*a centralized repository would be infeasible and an impractical use of resources*" (report #3). This alludes to the variability in

the different types of research and data types and challenges with conforming to the specifications of a single repository. This was observed in the interview data.

Addressing repository capabilities – Opportunities

When asked what capabilities a data repository would need to make it useful, the cohort research teams mentioned the following: communication with cohorts; summary statistics; advanced research tool to easily find cohort data; harmonized data; standardized formatting; clearly defined processes to request and receive data; standardized variables; linkable data sets and downloadable data. These are related to other a priori factors mentioned in the interviews and other document reviews for this dissertation research – *Analytic data complexity* and *Clarity of submission/access process.*

V. DISCUSSION

The sharing of human research data is a complex phenomenon that has been at the forefront of biomedical research since the 1990s, with the launch of the Human Genome Project, followed by the 2003 Fort Lauderdale agreement to share genome sequencing data from studies while protecting the identification of research participants. This has continued through the 21st century with NIH being at the fore front in the development and implementation of policies to foster the sharing of data generated from NIH-funded research.

With the deluge of data generated from scientific research, the rapid advances in technologies, and the tight and unpredictable funding climate, researchers are under a lot pressure to keep up with the rapid pace of science and also comply with data sharing policies from funding agencies and science journals, which to date has seemed to be very challenging (Borgman, 2012). Sharing data generated from research studies promotes high quality original research results that are validated, reproducible and replicable. This becomes even more important to ensure rigor in research methods (Harris, 2017), maximizing the value, access and use of data generated from federally-funded research studies to increase the pace of scientific discovery and improve clinical and public health outcomes.

NIH's guiding principle around data sharing states that "All data from NIH funded research should be made as freely and widely available as long as doing so safeguards the privacy of participants and the confidentiality and proprietary nature of the data." (NIH Data Sharing Policy, 2003). However, not all researchers are meeting this requirement for several reasons which were explored in this dissertation, and will be discussed in detail later on in this chapter. Therefore, understanding the multi-level factors influencing data sharing in federally-funded research is key to improving broad data sharing in research.

There are studies that have published on different factors impacting data sharing in general, including individual level factors. The goal of this dissertation research was to identify organizational level factors that facilitate or hinder the sharing of research data in public or controlled-access data repositories, and to identify opportunities for enhancing data sharing practices among NIH-funded researchers whose research areas of focus are in epidemiology, genomics or genetics. Given the history and culture of sharing in the field of genomics / genetics, this dissertation explored lessons learned from genomic data sharing practices that could enhance epidemiology data sharing, which is currently lagging behind.

A. General Discussion

There have been a few studies done on data sharing though most have been done through surveys of the broader scientific community and in different fields of research. A case study of genomic data sharing practices, a field with a culture of sharing, was the primary method used in this study to gain an in-depth understanding of the phenomenon of data sharing in biomedical research, particularly in research supported by the NIH, and how that could be applied to other research areas such as epidemiology. This qualitative research method provided a deeper exploration and understanding of similar factors identified in the literature as influencing data sharing (a priori factors) as well as the interrelationships between those factors. It has also helped uncover new factors or ideas that emerged from the research as important to answering the research questions of this study (emergent factors).

Although this dissertation research was primarily designed as a case study, during the later stages of the research, the standard model of case study research was combined with action research. This dissertation research is an example of a case study that is embedded within action research. Action research employs a repetitive, reiterative and interactive process for solving

adaptive problems, in collaboration with stakeholders to achieve organizational change goals. This involves several cycles in the action research steps: "planning to take action; taking action; assessing the effects of having taken action; reflecting on the implications of those effects; reevaluation and possibly modifying either the implementation plan or the goal or both; and starting again at the beginning of the succeeding cycle by taking the next action step, and so forth" (Marquardt, pp. 138-139).

Action research was not the original intended design for this study. However, the benefit of engaging leadership and staff in the development of this research, and incorporating feedback on study interpretations and conclusions from the senior leaders and other stakeholders at the program level to help address this complex phenomena was invaluable. This entailed inter-reflective thinking with senior leaders whose insights and perspectives on the data were integral in identifying key challenges during data analysis. This collaborative process with stakeholders was useful in gaining a better understanding of the underlying problem with data sharing in research and in identifying potential solutions and opportunities for change that are both beneficial and pragmatic for the organization.

Thirty-seven in-depth semi-structured interviews with NIH staff and investigators, and a review of five public and internal documents addressing data sharing in research were conducted. The questions asked during the interviews were open-ended questions that elicited a wide range of responses around factors that facilitate or hinder data sharing. Generally, these questions asked the participant to describe: their experiences, perceptions, thoughts about data sharing practices and policies among researchers and their institutions' norms, factors currently in place or essential for data access and submission, challenges experienced, opportunities for NIH and academic institutions to facilitate data sharing, and how lessons learned from genomic data

sharing could be applied to epidemiology data sharing. There were very slight differences between the interview guide used for NIH staff and that used for researchers (Appendix D).

The discussion of results are based on major themes that came out of this research, related to facilitators, barriers and opportunities for enhanced data sharing; addressing the first two research questions. Similar themes around facilitator, barriers and opportunities were observed when respondents described lessons learned from genomic data sharing practices, the third research question in this study. The co-occurrence analysis conducted in this study, i.e. looking at patterns and relationships between factors, was very informative in highlighting the relationships between the different factors; understanding that they are not totally separate; and identifying some of the most prevalent factors, concepts that needed further exploration, and areas with big implications in practice.

Both data collection methods - interviews and document reviews, corroborated findings in the literature and elucidated new findings not captured in the literature. These methods, in addition to integration of feedback from this case study practice site (EGRP) through action research helped elucidate opportunities for addressing challenges with data sharing and for improving data sharing practices in research. In addition, two new factors not previously identified in the literature or included in the original conceptual framework were found to be important and critical to conceptualizing and addressing this phenomenon of data sharing in biomedical research. These findings, including the lessons learned from genomic data sharing were used to develop a set of recommendations for enhancing data sharing among NIH-funded researchers (Appendix H).

Revised conceptual framework

The initial conceptual framework for this research showed four main high-level a priori factors at the organizational level that were hypothesized to influence data sharing in federally funded research studies. These are referred to as the main constructs of the study: CULTURE, POLICY, RESOURCES, and TECHNOLOGICAL INFRASTRUCTURE. A revised conceptual framework (figure 5) was created to capture key findings from the data collection and analysis of this dissertation research.

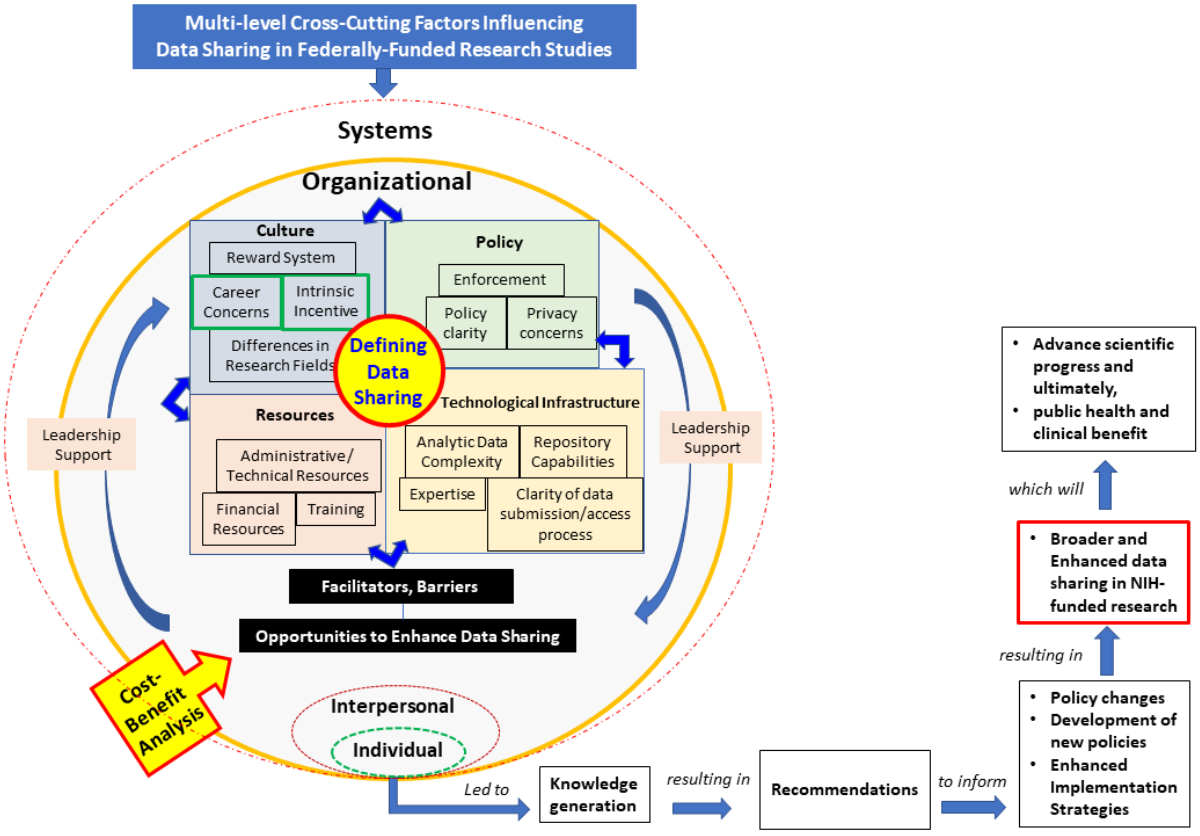

Figure 5: Revised conceptual framework: Factors influencing data sharing in federally-funded research

The revised framework shows the individual factors – both a priori and emergent factors, as influencing the sharing of data in public or controlled-access data repositories. During the analysis of data collected from in-depth interviews with researchers and NIH staff, and from the documents reviewed, the relationship between these factors became more evident. Unlike the initial conceptual framework, the revised framework shows the relationships and inter-dependence between the separate constructs and individual factors, which helped better understand the phenomenon and strategies for addressing it.

The identification of the a priori factors confirmed the findings from previous research and environmental scans conducted prior to data collection and analysis of this dissertation research. Each of the larger boxes in the revised framework represents one of the four main constructs – as in the original framework, with each of the constructs connected to each other. Within each of the constructs are smaller boxes that represent the a priori factors as facilitators and / or barriers or opportunities, and they highlight the inter-relationships between factors within and across constructs.

The CULTURE construct is comprised of organizational level and individual level factors – *Lack of a Reward system, Culture differences in research fields,* individual *Career concerns,* and *Intrinsic incentives.* The POLICY construct is comprised of *Policy clarity, Enforcement* and *Privacy concerns factors,* all of which are closely related to each other and with other factors across other constructs. Within the TECHNOLOGICAL INFRASTRUCTURE construct are the *Analytic data complexity, Repository capabilities, Expertise, and Clarity of data submission / access process* factors, all of which are also closely related to each other and with other factors across other constructs.

The RESOURCES construct shows the relationship between *Administrative/technical resources, Financial resources and Training. Leadership support*, which also falls under the RESOURCES construct is depicted as the driver of change and illustrated the two larger blue arrows outside of the construct boxes. Support from leadership at funding agencies and academic institutions is integral for driving change at the individual, organizational and systems levels, especially given the hierarchical structure of many organizations.

One of the new factors that emerged from this data is *Definition of data sharing* as a barrier, which was determined to be key to understanding the underlying challenges with data sharing in research. This is shown in the revised framework as the yellow circle at the intersection of the four constructs. Without a shared understanding and communication of what data sharing means, both within the research community and across NIH, there continues be a gap that fosters reluctance for data sharing; hence a low compliance with the NIH data sharing policies. The second factor that emerged from this research, and determined to be an important opportunity for improving data sharing was *Data Use and Cost-benefit analysis*. This is illustrated by the yellow box at the bottom left of the figure. The a priori and emergent factors, their roles and relationships with other factors, their similarities and differences, and how they impact data sharing in NIH sponsored research are discussed in more detail by construct and related factors.

Knowledge generated from the analysis of this study, through a deeper understanding of factors that facilitate or hinder data sharing, and opportunities for enhancing data sharing, resulted in a set of recommendations that will inform and improve data sharing at all levels – at the individual, organizational and systems levels. It could impact changes in data sharing policies, the development of new policies and enhanced strategies or approaches for policy

implementation. As a result, the immediate goal for broader and enhanced sharing will be achieved ultimately leading to advanced scientific discoveries, public health and clinical benefit.

Discussion of Research Findings

There are multi-level cross-cutting factors influencing data sharing in federally-funded research studies as shown in figure 5, and described above. Although the focus of this dissertation research was on organizational level factors that influence data sharing, it became clear through this research that to fully appreciate the challenges associated with data sharing, it required a systems approach. A systems approach is a leadership and management concept that emphasizes the inter-relationships and interdependence between different factors, internal and external, at organizations to help solve complex problems.[38] This was the rationale for using the Socioecological model and Systems theory in this dissertation research to better understand the individual factors and the relationships between the factors and how they influence data sharing. As such, systems approach in the context of this study also included some exploration of individual level factors which were a key component in better understanding and defining the problem of data sharing experienced by investigators.

CULTURE

The focus of this study was not on the individual level factors, such as motivation, attitudes, beliefs, etc., however it was difficult to talk about the culture and norms at academic institutions without mentioning individual level factors that influence the culture of data sharing in research studies / institutions. The feedback from the respondents on what motivates

[38] http://www.businessdictionary.com/definition/system-approach.html

researchers to share data (*Intrinsic incentive*) was useful in understanding the connection with factors within the CULTURE construct as well as factors in other constructs.

These two factors, individual *Intrinsic incentives* and *Career concerns* were important concepts the respondents described as motivating and hindering data sharing, respectively. The lack of a reward system at the institutions to recognize or reward the researchers for sharing data impacts the individual researcher's attitude towards data sharing. Their motivation to share data is partly driven by their intrinsic values – a commitment to advancement of science and as well as their careers.

Unlike the field of epidemiology, the field of genomics has a history and culture of sharing as evidenced in the literature. This was confirmed by the respondents in this study, indicating the importance of the role of culture as a key factor in enhancing data sharing practices among researchers in different research fields. *Culture differences between fields* of genomics and of epidemiology was considered to be both a facilitator and barrier and was found to be closely related to *Career concerns.* One of the lessons learned from genomic data sharing practices as described by one staff is illustrated below.

… the realization of the investigators that the sum was greater than the parts. And that if they came together and provided access to these data and provided access to these data even before publication, that it was going to be of benefit to all of them. And most scientists don't see it that way. Most scientists definitely think that they can't do anything before their research is published because it won't go anywhere if it does. But that was something specific to the genomics data. And that happened before our policy effort. That happened with the Bermuda and Ft. Lauderdale principles (staff)

Within the CULTURE construct, the most prevalent factor across interviews were *Career concerns* and *lack of a reward system*. The analysis of the data also showed that career concerns as a barrier and changes in reward structure as an opportunity for enhancing data sharing were observed in both the interviews and document reviews. These factors in the CULTURE

construct are related to other three constructs and their associated factors, which will be discussed in more detail later in the chapter.

Intrinsic incentive as a facilitator

The concept of *Intrinsic incentive* as a facilitator for data sharing was used to describe the internal motivation for researchers to do something good, such as sharing data for the advancement of science or to increase their own knowledge in the field. This factor was mentioned by 59% of the respondents, with an even distribution between investigators and NIH staff. Although there were no major differences observed across the different types of groups interviewed, more genomicists / genetic epidemiologists compared to epidemiologists, and more experienced investigators compared to new investigators, mentioned this factor. There was no observation of *Intrinsic incentive* as a facilitator in any of the documents reviewed, indicating divergence in the data sources which can be explained by the natural ability to be able to extract rich data through direct interaction with individuals, and in this case through in-depth interviews.

While most researchers are supportive of contributing their data towards the advancement of science as an act of altruism, as evident in approximately 70% of the respondents in this study, they are often conflicted because one of their major goals is advancement in their careers (related to the factor - *Career concerns*), which is heavily dependent on the number and quality of their publications. The nature of the scientific enterprise promotes competition among researchers for grant funding and reputation among their peers (Kaye, 2009) and they often face the challenge of balancing their own personal scientific research interests with the mandate to share their data in public or controlled-access data repositories.

Career concerns as a Barrier

Individual *Career concerns* as a barrier to data sharing showed a high prevalence rate across the different groups of respondents (70%) and a comparison by the different groups of respondents revealed that this was a bigger concern expressed mostly by the epidemiologists (92%) compared to genomicists/genetic epidemiologists (45%). This difference could be attributed to the differences in culture of the field of epidemiology compared to genomics / genetics in terms of the historical culture of sharing in genomics, and the nature of the data types expected to be deposited in the repository. The complex nature of longitudinal studies that involve the collection of epidemiology variables was echoed by participants as very challenging.

Also related to the differences in culture is the fear of data misinterpretation, misuse and scooping which could potentially have adverse effects on their careers. Most of the researchers interviewed believe that if epidemiology data is accessed from a repository instead of through collaboration with the data originators, the users will not have a good understanding of the nuances around the data, and may be more likely to misinterpret and misuse the data, which could in turn potentially hurt their reputation in the scientific community, and yield false and inaccurate results which have larger implications on science.

According to one of the respondents, one of the lessons learned from genomic data sharing that could support epidemiology data sharing is the value of communication with data originators to help understand the background and nuances of the data and variables. This will help avoid misuse and misinterpretation of the data thereby increasing the quality of the analysis and also becomes less concerning for investigators on the impact on their careers.

Well, I think, like that example that I gave you of people who used our data to get the wrong answer …there's so much benefit of communication with people who actually know the data and how it got there and what the variable names mean, and which variables you can trust, which you can't, what was the study design behind it. So, to the extent that that can be encouraged, then we get a lot more good science coming out of it … But there's really so much

to be gained by the knowledge of the people who acquired the data in the first place. It just, you know, you could short circuit so many mistakes and erronea.(experienced PI, epi)

In addition, the investigators fear that their data might get scooped even before they have had a chance to analyze the data and publish the results. One might argue that if there are no specific timelines for when data should be shared (contrary to what's in the policy), it might alleviate some of these concerns. However, there's the risk that it would not only limit the discovery of new findings from the data by others, but also affect reproducibility and replication of studies. On the other hand, the evaluation of a researcher's career is strongly tied to the number and quality of peer-reviewed publications in high impact journals. Ensuring that their data is cleaned and well-annotated for external use prior to submission to the repository was mentioned as requiring a lot more time than the current data sharing policy allows, especially for epidemiological data. This presents an opportunity to re-evaluate the timelines in the data sharing policy to help motivate researchers to spend the time needed to prepare and submit good quality data in the data repository for use by the public.

The accuracy in interpretation and use of their data by external users is perceived as important to their careers / reputation and to scientific research. Of equal if not of more importance to their careers is the desire to be the first to analyze and publish on their data before others do. The reward is the academic merit and recognition received and potentially increased chances for research funding. The fear that scooping of their data might prevent them from achieving these academic goals was considered by the investigators and NIH staff in this study as a barrier to sharing data in public or controlled-access data repositories.

Scooping of data in research has been discussed widely throughout the literature as a major barrier to data sharing and includes the malicious use of data by "research parasites" not involved in the study who personally benefit at the expense of the data originators (Longo,

2016). The preference among the respondents was to be able to share data in a collaborative manner to ensure that the data is accurate and not scooped. The literature shows mixed views on whether scooping is indeed a reality in the absence any evidence, or a myth (Laine, 2017).

Interestingly, 100% of the new investigators interviewed in this dissertation research expressed concern for potential data misuse, misinterpretation and scooping of data, and how it would impact their career advancement. Their main focus is on establishing a reputation in the scientific community, achieving recognition as they engage in different types of research in an effort to shape their career trajectory. Gewin (2016) says that "one key challenge facing young scientists is how to be open without becoming scientifically vulnerable." Although none of the investigators interviewed in this study had a direct experience with data misuse or scooping, what they described was more of shared concerns with fellow colleagues, not necessarily any direct experiences in the field, or they mentioned that they weren't sure if investigators had personal experiences with being scooped. Regardless of this, it seemed to have a strong impact on their perception and reluctance to share data in a controlled-access data repository. Therefore, this should be considered important especially because it hinders broad and open sharing of data in data repositories (Laine, 2017).

Lack of a reward system as a barrier

One of the factors that was shown to be closely related to *Career concerns* was the lack of a reward system at institutions. The factor, *lack of a reward system* as a barrier to data sharing was also very prevalent and similar (approximately 80%) across the different investigator groups. Researchers tend to agree with the value of data sharing but given the time and cost involved with sharing data in a data repository, they are reluctant to go through the process. They see no clear motivation to do all this work if there is no credit or reward for sharing data.

The findings from this dissertation research corroborates findings in the literature that institutions do not have a system for rewarding researchers for data sharing. According to the respondents, it is difficult to see how sharing their data benefits them directly if they are concerned about their data being scooped and are not getting any credit or recognition from their institutions for sharing data. Bierer (2017) says that "Although it has been recognized that appropriate and meaningful incentives are essential to capitalize on the promise of data sharing and that crediting data generators is key in this effort, to date there has been no systematic implementation of a standard process and method to credit original data generators."

The recognition of data sharers needs to come from the institutions but according to a few of the respondents who sit on promotion and tenure committees at their institutions, data sharing is not part of the promotion and tenure process, which is currently the most common form of reward in academia. This is despite the fact that 68% of the respondents in this study mentioned that the culture of their institution promotes and supports collaborative research (though within the institution or departments), which requires sharing data with others in one way or the other to help answer new research questions.

The disconnect between the concepts of "data sharing" and "collaboration" was an important area that emerged from the data and was explored to further understand the role of institutional culture on data sharing practices among researchers, and the lack of a system for rewarding investigators that shared data. This finding demonstrates that the researchers involved in collaborative research consider "collaboration" to be part of "data sharing" but their institution seemed to divorce the two concepts for reasons that were not specified. This assumption was implied from the responses when some of the respondents mentioned the lack of or inadequate administrative and technical resources provided by their institutions to support data sharing. In

addition, some respondents indicated that data sharing was not a priority for investment by their institutions given other competing priorities, actual administrative and financial burden incurred, and liability concerns.

… so far, we share data – everything is for research purposes only. We share data to other faculty, for example, within institution. We also share data to graduate students for them to do their dissertation or maybe students to do thesis. That's within institution. … We are pretty open in terms of sharing our data as long as the project is approved by the research committee, by the [redacted] research committee. …. I don't think the institution disagrees with data sharing, but the institution's primary worry is liability. (new PI, epi)

With the increase in collaborative research, the sharing of data becomes even more important (Tenopir, 2011). This highlights the need for a change in institutional culture around data sharing to help improve current data sharing practices, but it may take some time. Researchers, constantly trying to balance the demands of the mandate with their own personal interests and scientific commitments expect their institutions to recognize the time and effort spent on data sharing and have it count towards their academic careers.

One of the lessons shared from genomic data sharing as described by one of the respondents is the negative impact of lack of rewards on the quality of data shared in the repository. The lack of reward or credit for sharing coupled with the burden of sharing data as well as the enforcement of policy could lead investigators to do the bare minimum or not spend the time required to ensure that the data is clean and of good quality before submission. The risk is a "data dump" that is not useful or meaningful to the secondary user.

And I can tell you, it was painful, it was thankless, and it took a tremendous amount of work. Once you have the solutions, though, you have to realize they have to be easy to use and there have to be incentives for people to actually use them and follow the rules. Because unfortunately, what we saw time and again was that people basically did the minimum that was required to adhere to the letter of the law, not the spirit of the law. So sure they'd dump their

data out there, sure they'd dump out metadata, but often it wasn't very useful. And that was sort of interesting. (experienced PI, gen/gen epi)

More NIH staff than investigators identified opportunities for addressing changes in reward structure at institutions. Most of the suggestions were around institutions considering giving credit and rewards to investigators who shared data and on the NIH end, also rewarding the data sharers. The number of publications and position of authorship are major factors considered in the criteria for promotion and tenure, which is the ultimate goal for researchers in academic institutions. Currently, the metric for success in academia is centered around the number of publications in high impact journals as well as the author's position on the publication –first and last authorship are among the highly coveted positions in research publications. Middle authorship which comes with large consortia or collaborative studies does not seem to be recognized by institutions in the evaluation process and respondents suggested this be changed in the current process.

Both investigators and staff agreed that a culture shift in how institutions reward sharing could be a good motivation and incentive for sharing data in public or controlled-access repositories. Although one of the biggest factors driving this gradual shift in culture is the enforcement of data sharing policies by NIH staff, NIH does recognize the need to credit researchers who share data and supports the idea that institutions consider data sharing in the promotion and tenure process. According to Olfson (2017), developing a common metric for data sharing (S-index), similar to the H-index used for publications could possibly be an opportunity to give credit to data originators when their data gets cited and used by others. Redefining or identifying the different types of incentives such as software and other information that could help facilitate data sharing and ensure maximum value is an approach to enhancing sharing through motivating the investigators. From the perspective of a funding agency,

opportunities for rewarding researchers who have a history of sharing data was to tie it to the grant review process and funding decisions. This would suggest that those investigators that have a strong history of funding would be rewarded for sharing their data.

The culture of the institution has a big influence in how researchers think about and conduct their research. Data sharing is no different than any other principles supported by the institutions that govern and / or advance scientific research. If the institutions focus on sharing data through internal collaborations only, which means that the data is kept within the control of the investigator and not deposited in a data repository, the researchers are likely to take on that same mind-set which limits sharing, thereby resulting in non-compliance with the NIH policy to share data broadly. A shared understanding of data sharing at academic institutions is critical to achieving the goal of broader sharing of data which would require a shift in culture, thinking and change in perspective or norm. The responsibility rests on the shoulders of the leadership to drive and implement change from the top.

POLICY

Within the POLICY construct, the a priori factors *Policy clarity* and *Enforcement of policy* as barriers and facilitators to data sharing were closely related to each other. Also related to these factors were a priori factor, *Privacy concerns* and the emergent factor *Definition of data sharing*, both as barriers to data sharing. In general, across data sources, the following factors occurred in both the interviews and document reviews: *Enforcement of policy* and *Inconsistent enforcement of policy, Privacy concerns,* Opportunities to *address privacy concerns*, and the *Definition of data sharing*.

The most prevalent factors within this POLICY construct were identified by respondents to be the *Enforcement of policy* as a facilitator for data sharing (70%), the *Lack of clarity in policy* as a barrier (89%), and *Privacy concerns* as a barrier (59%). Although not mentioned widely by the respondents, the *Clarity of policy* as a facilitator (43%) and the *Inconsistent enforcement of policy* as a barrier (46%) and were also considered to have significant implications in improving data sharing practices in NIH-funded research.

Policy clarity as a facilitator and Lack of policy clarity as a barrier

There were fewer respondents overall who considered the current NIH data sharing policies to be clear and reasonable in terms of the expectations, requirements and timelines, and more than half of the staff (58%) compared to 36% of the researchers indicated that the policy was clear. This could be some semblance of bias given the policies are developed and implemented by NIH. This is interesting because during the course of the data collection and analysis of this dissertation research, discussion with NIH colleagues at the case study site revealed that some staff thought the data sharing policies were clear but that the policy was enforced inconsistently. On the contrary, the lack of clarity of the NIH data sharing policies as well as the inconsistencies and ambiguity in interpretation was a strong sentiment reflected in the interviews as shown by the high prevalence rate overall (89%), and among researchers (84%) and staff (100%) interviewed. Both new investigators and experienced investigators mentioned similar sentiments around the lack of clarity of the data sharing policies.

To put into context these concerns, the 2003 data sharing policy states that " … NIH expects the timely release and sharing of the data to be not later than at the time of acceptance of publication of the main findings from the final dataset." The 2015 GDS policy states that "data

should be submitted once it has been cleaned … following data submission the data may be accessible only to the submitting investigators and collaborators for a period not to exceed six months … NIH will release de-identified human genomic data submitted to NIH-designated repositories no later than six months after the initial data submission begins or at the time of acceptance of the first publication, whichever comes first."[39]

The language of the policy as written is such that it could have subjective interpretations to different researchers depending on their type of study and type of data collected. Longitudinal cohort studies in particular have multiple main findings written into their research aims and because of ongoing follow up of research participants which may take several years beyond the life of the grant, investigators may not have what they consider a finite "final dataset" at the time period defined in the policy. This makes enforcement of policy challenging for staff; they understand the complexity of the study design, but must also as part of their duties enforce the policy.

Part of the challenge observed in practice during the development of this dissertation project was inconsistent interpretation of the policy among staff across NIH. . The tension between what the policy says and what the "expected" or "ideal" policy should look like was also observed in this study. The lack of consistency creates tension and confusion on both ends and impedes effective and efficient sharing of data in controlled-access data repositories.

Successful policy compliance is dependent on how clear, consistent and transparent the policies are, as written and communicated to researchers and NIH staff. The heterogeneity among data types expected to be submitted in the repository within a specific timeline makes it difficult for researchers to adhere to the policy. A couple of respondents mentioned that the

[39] https://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html

approach for data sharing should not be a "one-size fits all" approach, implying that there shouldn't be a single policy for different data types. The challenge as indicated by one staff is in the way that policies in government are developed which are intentionally meant to be broad and "*lofty*". The lack of uniform approach in policy development was recognized to be problematic by NIH and calls for consideration of other approaches to improve policy clarity and enforcement. This may be outside program's control but the hope is that the findings from this dissertation research could influence policy change; it may take a while.

Clearer guidelines and requirements could improve data sharing among researchers and communication of processes for submission and knowledge of what's required and who to go to were considered important to the respondents in this study. One of the lessons learned from genomic data sharing around communication was that funding agencies and investigators need to be clear about what variables are expected to be shared early in the research process and prior to data collection,. In addition, in terms of ensuring broad sharing and access, it was important to understand the audience and target communication with them. This may increase awareness and compliance and lead to successful submission of data in a data repository. This was also related to the availability of training, tools and materials to improve knowledge and skills for interpreting and implementing data sharing in research.

Enforcement of policy as a facilitator and Inconsistent enforcement of policy as a barrier

Similar to the issues with the lack of clarity of policy, 46% of all respondents also mentioned that the inconsistency in the enforcement of policy across NIH was not encouraging. There was the sentiment that enforcement of the policy varied across NIH institutes and centers, and between NIH extramural and intramural research programs. In addition, not all journals

require that data be shared prior to acceptance of publications; so changing this would be useful in the enforcement of data sharing policies all around.

Contradictory is the perception of some extramural investigators and staff who didn't think that intramural investigators complied with the data sharing policy, compared to the perception of intramural investigators. This also have a negative impact on the attitudes and practices of investigators toward data sharing as there's the tendency for perceived lack of fairness. The differences in opinions could be attributed to a lack of consistent and transparent approach in policy implementation across NIH institutes and centers. This is challenging to accomplish given the size of the organization and the many stakeholders involved. However, it is critical to the success of data sharing in NIH-funded research.

Successful policy compliance and implementation are dependent on how clear, consistent and transparent the implementation guidance and processes are. It will require dedicated resources in the form of administrative support such as training and education for NIH staff to effectively implement the policy. The 21st Century Cures Act of 2016 provides staff (as delegated down by the NIH Director) with the authority to enforce the data sharing policy. However, some of the staff indicated that they did not feel like they had the authority to enforce the policy. This can be attributed to their lack of comfort and confidence with the details of the policy requirements, interpretation of the policy and implementation guidance documents, and fear of potential consequences for non-compliant investigators which currently is in the form of barred funding. The investigators indicated that they were not clear on what the consequences were, whether there was a consistent application of the consequences and its effectiveness on sharing. Increased communication around enforcement strategies could be an external motivator for enhancing data sharing.

Definition of data sharing as a barrier

One of the biggest discoveries in this study is the deeper understanding gained in the differences between how researchers and staff define data sharing, and how that influenced attitudes and perception of sharing data in data repositories. The variation in definition and understanding of what data sharing means is strongly related to the *Lack of clarity of policy* among researchers. As described in chapter 4, the definition of data sharing was grouped into two main categories – collaborative model of data sharing and sharing through data repository. This goes back to the points made earlier in chapter 2 that the concepts of "data sharing" and "data" are elusive factors that influence data sharing (Borgman 2012). It means different things to different people in different fields.

Through active discussions with colleagues at EGRP, it was determined that "collaborative sharing" was not aligned with NIH's definition of "broad sharing" which is described as sharing of NIH-funded research data via public or controlled-access data repository for broad access to users or the public. This is not explicitly stated in the data sharing policy. The 2003 data sharing policy states that "Data should be made as widely and freely available as possible while safeguarding the privacy of participants and protecting confidentiality and proprietary data." (NIH Data Sharing Policy, 2003). However, the interpretation of this statement is subjective and it's not clear what this might mean to investigators or how they may interpret it. The meaning of this statement is implied to mean sharing through data repository for broad access – this is how NIH defines "broad sharing" but note that this may not be a uniform interpretation or understanding among NIH program staff; it directly affects how they try to enforce policy with their grantees. Similarly, there was varying degrees of frustration among the

investigators in the study who strongly believe that sharing data through collaboration should meet the policy requirement.

This research shows that the lack of a shared understanding of what data sharing means, is the underlying problem that cuts across all the four constructs of the study. Given the competitive culture of the scientific research enterprise, investigators are reluctant to spend their time on the effort it takes to submit their data in a data repository because of the time and cost involved, as well as the concern that it might put their data at risk for scooping and misuse, which is threatening to their careers. Both the interviews and the document reviews revealed that investigators preferred to share data through the collaborative or enclave model, which gives them more control of their data. In addition, the fear of breach of patient privacy at the individual researcher level and the institution level is perpetuated by the notion that sharing in a data repository is risky because of the perceived loss of control of the data if shared outside of the investigator's own or institution's internal repository or database.

A clear and uniform understanding of data sharing across biomedical research will impact how people think about data sharing and the processes for data sharing. Prioritization at the organizational levels could result in provision of adequate administrative, technical / technological and financial resources to support data sharing efforts at the individual and organizational levels. This is critical to improving data sharing practices among NIH-funded researchers.

Privacy concerns as a barrier

Improved clarity of the data sharing policies could help alleviate privacy concerns, perceived as a barrier to data sharing among researchers and their institutions, along with the fear of liability by the institutional leadership if adequate care is not taken to protect the privacy of human research data that's deposited in public or controlled-access data repositories. There are consequences associated with a breach of privacy of research participants' data and it's the responsibility of the institution through its Institutional Review Board (IRB), to ensure that all of their researchers adhere to the IRB requirements, and that dire measures be taken to protect the privacy of data from human subjects in their research.

The fear of potential violation of confidentiality is tainted by their belief that data that is not within their control (researchers' / institution's), i.e. deposited in a public or controlled-access repository is not safely guarded. As a result, prioritizing and taking steps within the institution to ensure that data is shared in these repositories may not be top priority for them. According to the respondents, most of the data sharing happens within institutions and departments. One of the lessons shared by one of the respondents as related to genomic data sharing was that controlled-access works; data in controlled-access data repositories are safe and secure. This was not a shared feeling among other respondents, although none of them had any direct experiences with breach in data confidentiality in their studies. NIH ensures that a certificate of confidentiality is obtained from institutions to protect the privacy of study participants enrolled in NIH-funded studies.

Some suggestions from staff and researchers were to encourage broader access through broader consent forms, and to ensure that study participants truly understand what they are consenting for prior to participation in research studies. These elucidate the level of complexity of data sharing because of the sensitivity of the data involved, and the responsibility to ensure

that the data remains confidential during research, while trying to comply with the mandate to promote broad sharing and access for the advancement of research. There is also mounting pressures on academic institutions to ensure that their researchers meet the requirements of the data sharing policies, while guaranteeing confidentiality of research participant data (Kaye, 2009).

RESOURCES

Administrative and Financial Resources as facilitators and Inadequate Administrative and Financial Resources as barriers

Within the RESOURCES construct, *administrative / technical and financial resources* were among the strongest themes reflected across data collection methods - interviews and document reviews, as facilitators and barriers to data sharing, and were very closely related to each other. Prevalence across the different types of respondents by new PI, experienced PI and staff were high and there were no significant differences observed. The analysis shows that by investigator self-identified research fields, admin/tech resources as a facilitator was mentioned more by the genomicists/genetic epidemiologists compared to the epidemiologists. This was the opposite for financial resources as a facilitator.

Inadequate admin/tech and financial resources as barriers to data sharing were mentioned by more experienced investigators and staff compared to the new investigators. This could allude to the more established researchers who have experienced challenges with data sharing since they've been doing it longer. By research field, there were more genomicists/genetic epidemiologists who mentioned both factors as barriers to data sharing. This could be due to more experience with submitting data in dbGaP given more enforcement with genomic data.

Overall, the data illustrates the importance of these factors either as barriers or facilitators to data sharing and addressing them could have a significant impact in efforts to enhance data sharing among researchers.

Throughout the literature, resources have been an integral component in the promotion of data sharing in the scientific community. The preparation of the data for submission requires personnel with the right expertise in bioinformatics and the salary to support the personnel are all key to successful data submission. Training might be required to support data sharing efforts and will require funding. The respondents in this study mentioned the importance for funding agencies and academic institutions to provide resources, such as centralized support staff at institutions with the right expertise to facilitate data sharing activities given the time and effort and administrative and financial burden on investigators.

Adequate resources are needed to conduct good quality and high impact studies and because of the competitive scientific environment, and the limited funding pool, researchers often prioritize how and when they spend their resources. "Time and money spent on documenting data for use by others are resources not spent in data collection, analysis, equipment, publication fees, conference travel, writing papers and proposals, or other research necessities." (Borgman, 2012)

Given the tight funding climate, researchers are constantly competing with their peers to acquire resources to support their research. Some have more success than others and some have multiple grants to support their research, training of students, etc. For the more junior investigators who are just getting started in their careers, they are forced to compete with the rest of their colleagues, including the established investigators to secure funding from a limited pool

of funds. The new investigators in this dissertation research study are recent recipients of their first large R01 grant, and more than two-thirds of them mentioned that having adequate administrative / technical resources was an important facilitator for sharing data, and half of them considered financial resources a facilitator for sharing.

In particular, prioritizing data sharing implies that support from organizational leadership to provide administrative and / or financial resources is needed to perform data sharing activities. NIH recognizes the resource burden associated with data sharing and encourages investigators to include funding for data sharing in their grant budget. Not all investigators were aware that data sharing policies allow investigators to include in their grant budget the cost for data sharing. This is clearly stated in the NIH GDS policy, "Any resources that may be needed to support a proposed genomic data sharing plan (e.g., preparation of data for submission) should be included in the project's budget.[40]" However, this is not clearly stated in the 2003 general data sharing policy where the statement is listed in the Frequently Asked Question section and in the Implementation Guidance for the 2003 data sharing policy, but not in the actual policy. This may not be intuitive or easy for investigators to find and speaks to the issue with clarity of policy discussed earlier. "NIH recognizes that it takes time and money to prepare data for sharing. You can request funds for data archiving and sharing as part of your grant application for collecting the data. If you have already collected the data, you may want to ask your NIH Project Officer about a competitive or administrative supplement. NIH recommends that you consider procedures and costs for data sharing during the application process rather than after the data have been collected."

[40] https://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html

In addition, the statement for requesting funding for data sharing in the grant application is not included in the final 2003 NIH data sharing policy but rather in the 2002 draft policy that was eventually updated in 2003. "The NIH will expect investigators supported by NIH funding to make their research data available to the scientific community for subsequent analyses. Consequently, the NIH will require that data sharing be addressed in grant applications (e.g., in sections related to significance, budget, and the end of the research plan) and in the review of applications. Funds for sharing or archiving data may be requested in the original grant application or as a supplement to an existing grant.[41]"

The data sharing policies ask that investigators work with their NIH program officials to obtain specific guidance on data sharing. There was concern that it would be difficult to calculate a budget estimate for data sharing to include in the application since investigators are unsure of what the projected data sharing request or needs might be. Investigators were also not thrilled about this because they mentioned that even though they include the cost for data sharing in their budget, the NIH standard programmatic cuts which are part of the funding decision process still gets applied, thereby reducing their budget even further. The availability of adequate resources is highly dependent on the support from an institution's or organization's leadership. This also includes provision of training for both researchers and staff to help them better understand the policies and processes of data sharing.

Leadership support

[41] https://grants.nih.gov/grants/guide/notice-files/NOT-OD-02-035.html

The support from leadership at funding agencies and academic institutions, under the construct RESOURCES, was considered to be one of the most important factors to bring about change, both in the conceptualization of data sharing challenges and opportunities to implement change at the organizational and individual levels. The analysis of this dissertation shows the critical role that organizational leadership plays in bridging the gap between factors that hinder and facilitate data sharing, through the provision of resources (administrative, technical, financial, training and education) needed to foster data sharing, oversight of policy implementation and prioritizing broad sharing and access in general.

For change to be made at the organizational level, it will require the support from the leadership of NIH and academic institutions to understand the facilitators, barriers and opportunities to improve data sharing. In some cases, with such a complex phenomenon as data sharing, it may require leadership support beyond the organization and extended to the systems level to further explore opportunities for change. NIH has invested in a large amount of resources and infrastructure towards the support of data sharing. This is evident in initiatives, programs and priorities developed by NIH to promote data sharing among its funded researchers. Some examples include NIH funding opportunity announcements for administrative supplements specifically targeting support for data sharing, the creation of the NIH Data Access Committee (DAC) to facilitate the request for genomic data, and more recently, the creation of the NCI Office of Data Sharing in 2018 to centrally coordinate data sharing activities across the institute.

Successful data sharing ultimately requires that data sharing efforts from policy development, compliance and implementation integrate a systems approach that includes participation of all stakeholders including NIH, academic institutions, journal editors, researchers and the public. There are opportunities as discussed in this chapter for both addressing the gaps

or challenges, and for enhancing existing efforts and processes at the organizational levels. These all require support, commitment and investment from the leadership; the change needed stems from the top down with leadership of institutions and funding agencies prioritizing data sharing.

TECHNOLOGICAL INFRASTRUCTURE

Within the TECHNOLOGICAL INFRASTRUCTURE construct, *Analytic data complexity* was the only factor observed in both the interviews and document reviews as a barrier, facilitator and opportunity. The factors *Clarity of submission/access process, Lack of expertise*, and *Sub-optimal repository capabilities* as barriers to data sharing were observed in both data sources. The most prevalent factors in this construct were *Clarity of submission/access process* as a facilitator (73%), and *Sub-optimal repository* as a barrier (73%). Although overall, the factor, *Analytic data complexity* was not as prevalent as other factors - as a facilitator (38%) or barrier (57%) - the analysis showed it is closely related to other factors within the POLICY construct as well as the RESOURCES construct, and is critical in the data sharing process.

Analytic data complexity

It is important that data submitted in the data repositories are cleaned, well-annotated with data dictionaries and in the proper format required by the repository so that users can understand the background of the data, how the variables were derived and any versioning of the datasets that will inform accurate interpretation and analysis of the data. Quality control checks on the data takes a lot of time and effort on the submission end by the investigator preparing the data, and on the receiving end by NIH staff who check the data for errors prior to accepting the data. Both the staff and investigators in this study acknowledged it was not a simple task

especially with large datasets. Their perspective is that time investigators spend on preparing the analytic data set is less time spent on doing the research.

In general, both staff and investigators in this study agreed that epidemiology data was more difficult or complicated in terms of submission. However, a divergent perspective from this was from one experienced genomicist/genetic epidemiologist who mentioned that both genomic data and epidemiology data have their issues, and one is not necessarily easier than the other because it depends on which type of genotype data file is submitted and for phenotype data there's a risk that cohort data in repositories could easily be outdated and inaccurate.

The extra effort required to prepare the data for external use compared to internal use was considered burdensome for investigators. One could argue that investigators should do the work upfront, that is, invest the time and effort in the data cleaning and documentation early in the data collection phase. This way when it's time to submit the analytic data in the repository, it might not be as difficult or challenging. The problem is that some of the longitudinal cohort studies collected data prior to the release of the data sharing policy and may have to go back to format the data for submission in the repository if it's tied to their current research / cohort data or reconsent their study participants.

One factor noted by the respondents as being affected by the ambiguity in the policy is the analytic results / metadata that is expected to be submitted to the data repository. Investigators expressed concern with the lack of clarity around the types of data to be submitted and the timeline in the policy for data submission. In addition, the respondents, except for one, mentioned that unlike genomic data, phenotypic data or data generated from longitudinal epidemiology studies are more challenging to prepare for submission in data repositories, in

terms of data format and documentation of the variables. This is also related to the quality of the metadata in the data repositories and how reliable they are for the secondary user.

The issue of not fully trusting the data in the repository also came up and negatively influenced respondents' perception of data sharing. This is the interpersonal level aspect of the systems model that was not the focus on this research. The fear that other users may not understand the nuances around the data if accessed directly from a data repository, without collaboration or consultation with the data originator, was an important concept that was related to *Career concerns*. Misinterpretation of the data by secondary users may have a negative impact on researcher's career; they are unable to publish on their data if someone has already analyzed and published on the same data, and publication of incorrect analyses may also taint their reputation among their peers in the scientific community.

Clarity of submission/access process, and Expertise as facilitators and barriers

The factor, *Clarity of submission/access process* was expressed by majority of the investigators and the researcher as a critical component to successful data sharing. The data sharing process is a complicated one that involves many people and many steps for data submission and access in a data repository. Without clear and transparent systems, processes and tools to facilitate data submission and access it becomes difficult to comply with the policy.

A consistent comment by experienced and new investigators was the need for clearer processes for how more junior investigators could access the data in the repositories and who they could go to get support or advice from for their studies. This was related to the points observed in the interviews and the document reviews that it is not easy to know what types of

studies are in dbGaP since the submission of data is done in a piecemeal way, project by project, which doesn't allow investigators to easily identify new or different research questions to ask.

The availability of resources to facilitate efficient and clear processes for investigators should be a priority. However, it is important to recognize the potential for information overload that comes from NIH programs, and to identify ways to streamline the processes and make it easier for the investigators. Adequate *expertise* and knowledge are required to properly navigate these processes, including the awareness of key resources needed to facilitate data sharing in data repositories.

Repository capabilities as facilitators and Sub-optimal repositories as barriers

For data generated from NIH-funded research to be shared broadly as encouraged by NIH, the researchers are required to deposit their data in a public or controlled-access data repository and others must be able to access the data for reuse. This requires well functional, operational and efficient repositories, with the capacity to effectively handle the different data types and sizes.

There were more genomicists/genetic epidemiologists (82%) who mentioned this factor as a facilitator compared to 43% of the epidemiologists, and slightly more new investigators (67%) compared with experienced investigators (58%). The differences in respondents could be attributed to the investigators experiences with data submission / access in a data repository. The dbGaP is the main controlled-access data repository that NIH supported investigators are required to submit data to. This database was originally designed for genotype data although it also accepts phenotype data associated with genomic / genetic information. The genomicists / genetic epidemiologists with genotype data probably have more experience with dbGaP, hence

the difference in response between epidemiologist. The new investigators who are new at data sharing consider clarity of processes a key facilitator in data submission and access in dbGaP.

The investment by NIH in optimal data repository is critical in ensuring that data sharing happens in an efficient and timely manner. Delays in the process have huge impacts on scientific progress. Therefore, having for example, the resources to support the data sharing infrastructure, such as streamlining and automating the processes for data transfer agreements, is something that could be supported by the organizational leadership or funding agencies. There are several models of data repositories where best practices or lessons learned can be gleaned to enhance dbGaP e.g. automating the processes as much as possible and having adequate staffing support at NCBI to support investigators.

Emergent factor: Data use, cost and value - Opportunity

Data use, cost and value was a factor that emerged from the data collection and analysis as an opportunity to improve data sharing. Despite the overall low prevalence across respondents (24%), 16% of the investigators and 42% of staff mentioned this factor as an important opportunity for improving data sharing in NIH supported research. There were more references of this factor attributed to new investigators and staff, and epidemiologists compared to the genomicists/genetic epidemiologists. The respondents in this study unanimously agreed that data sharing was beneficial for answering new research questions, for replication and reproducibility of findings, and to advance the pace of scientific discovery. However, the use of data shared through a data repository was not readily clear, despite the understanding that it is critical for biomedical research (Corpas et al, 2018).

It is important that there's a common understanding of the value of data shared in data repositories so that there's an increased appreciation for the amount of investment in resources

by funding agencies, investigators and their institutions to support this type of data sharing (Coady et al, 2017). The lack of knowledge of what data in the repositories had been used for in terms of their contribution in scientific discovery was consistent among investigators and staff. Although studies have been published that describe the utilization of genotype data in dbGaP (Paltoo et al, 2014) and clinical trial data in the National Heart, Lung, and Blood Institute data repository, BioLINCC (Coady et al, 2017), the respondents didn't seem to be aware of these publication, which indicates the need for increased awareness in resources generated by NIH. According to these publications, both repositories have demonstrated the value of data reuse in facilitating new research questions on cancer, mental health and cardiovascular disease as well as "demonstrating that a small set of genes contributes to a range of psychiatric disorders, including schizophrenia, bipolar disorder and autism." (Paltoo et al, 2014)

The concept of cost-benefit analysis of data sharing in repositories illustrates the need for evaluation of the cost compared to the benefit or value of sharing data, especially given the cost of data sharing and the huge investment by NIH. It was not clear from the respondents that such an analysis had been done and presents a great opportunity to help with understanding whether the costs of sharing data in data repositories outweigh the benefit in terms of the type of science done, as well as to inform how investigators could use the data in the repositories for research. The analysis could potentially be an opportunity for a more compelling argument for investigators to share data and to help funding agencies with making more informed decisions on priority setting around data sharing and investment of resources to support data sharing activities.

At the end of the day I think researchers do respond to compelling arguments that things that are good for the science are things that you know we should do. And so I think one of the other things that is potentially useful for like sort of the entire community to think about is use cases for data sharing that can really show in a very compelling way that data sharing actually

is something worth doing and you know maybe these different approaches of data sharing are worth doing. … You know, scientists tend to do things that they think are good for science or for solving whatever it is that they're trying to solve, you know, whether it's finding the next treatment for a disease or overcoming some scientific or technical barrier. And so if you can make a compelling case about why data sharing facilitates that, I tend to think you would get more buy in from the research community. Or even I think, you know, particularly if it's something that is advocated by very well respected scientists in whatever their respected fields is. Scientists tend to listen to other scientists. (staff)

Finally, related to this factor was the suggestion for an evidence-based qualitative research that clearly shows a shared understanding of the concepts of data sharing and informed consent among research participants. This was suggested as another opportunity to mitigate concerns around patient or participant privacy or violation of confidentiality. Investigators mentioned privacy concerns as a barrier to data sharing because they felt that data in a data repository was less secure than if were within the control of their own personal or institutional repositories, enclaves or servers. NIH continues to ensure that its servers and repositories have the highest level of security to protect data in the repositories.

Extreme measures are evident in the creation of new laws to protect confidentiality of research participants. With the increase in large global research collaborations and international consortia research, it is important to ensure that confidentiality of study participant data is protected. New international laws such as the European Union General Data Protection Regulations (GDPR) enacted in May 2018 were created to strengthen privacy rights of research participants in the European Economic Area (EEA), superseding the Health Insurance Portability and Accountability Act (HIPAA). This has implications for data sharing in NIH research involving European collaborators, and could potentially lead to delays, impede access and use of data collected from participants who are in the EEA.

B. Limitations

One of the limitations of this study is that the sample size was small and therefore the views of the respondents – investigators and NIH staff – may not be accurately reflective of and generalizable to all investigators supported by NIH research and NIH staff, or the entire scientific community. This includes new investigators, experienced investigators, epidemiologists and genomicists / genetic epidemiologists whose research are supported by NIH. The findings are also not generalizable to all NIH institutes and centers, other public health agencies or organizations and institutions that support biomedical research. However, the findings of this study are transferable to other similar organizations or federal agencies or similar research groups outside of NIH who conduct and / or support similar types of research and deal with challenges of sharing data. The findings are transferable to data sharing practices in epidemiology and other federal and non-federal research groups.

The selection of participants interviewed for this dissertation research ended up being a convenience within purposeful sampling. At the beginning of the study, specific types of investigators and NIH staff were purposely recruited into this study. However, given the large number of participants who expressed interest in participating in the study, convenience sampling was used such that participants were selected on a first come first serve basis in order to manage the number of respondents. Extreme care was taken to maintain balance between new investigators, more experienced investigators, epidemiologists and genomicists / genetic epidemiologists, and the different types of NIH staff that were included in the study.

This project was developed in collaboration with stakeholders at NCI where professional relationships with most of the respondents were established prior to the onset of the research, and could have resulted in respondent bias and study investigator bias. However, the advantage of

being an "insider" researcher on this dissertation research study was that this study was responsive to challenges observed in practice. Relationships with EGRP/NCI stakeholders were key to the interpretation of the data and findings from this study would benefit the organization as it grapples with the best ways to improve sharing of data generated from NIH-funded research.

Despite the systematic approach employed in this study to distinguish between facilitators, barriers or opportunities, there was some overlap observed because of the way the respondents framed their comments and the interpretation by the Study Investigator. Specific questions in the interviews were asked about facilitators and barriers to data sharing. The opportunities for improving data sharing were deduced based on the Study Investigator's knowledge of whether the ideas described were new (opportunities) or existing or currently in place (facilitators). Decisions were made in a systematic and consistent manner to apply the best fitting codes as accurately as possible, guided by the previously established definitions, and informed by expertise of the Study Investigator an "insider" researcher.

In addition, discussions with a seconder coder after co-coding a proportion of the interviews was helpful in refining the definitions of the codes. The co-occurrence analysis also helped uncover relationships between these different factors and showed that they were not mutually exclusive, i.e. not totally separate categories. The limitation with the overlapping codes is that it creates a need for further discussion with and presentation to other stakeholders to validate the findings. The design of this study and analysis built in ways to check the interpretation and vet the findings through collaboration with stakeholders.

C. Implications and Recommendations for Practice and Leadership in Public Health

There are several implications of this research and its findings for improving the sharing of data generated from research studies across the scientific community and the general public, and promoting broader data access. The NIH is the largest funder of biomedical research in the world and as such has a responsibility for ensuring that the data generated from NIH-funded research is made as widely available as possible while protecting the privacy of the research participants or patients. To help maximize the investment of tax payer dollars in the funding of scientific research, it is imperative that the NIH and other federal agencies develop clear and coordinated policies that will promote broad access to scientific data generated from federally-funded research studies (Holdren, 2013).

The benefit of data sharing, also referred to across the literature as "open data" has been proven to be invaluable in the discovery and advancement of science. However, the findings of this study, corroborated by the literature, showed that investigators are reluctant to share data, especially in public or controlled-access data repositories. The complexity of data sharing in scientific research is well described in the literature and in this study and is attributed to the many components of the data sharing process, the different key players involved at the NIH and at academic institutions, and the differences in perspectives.

Findings from this study were invaluable in highlighting critical factors that facilitate or hinder data sharing, as well as opportunities for maximizing the value of the data generated from NIH-funded research. The lessons learned from genomic data sharing practices, many of which were discussed as facilitators, and barriers to data sharing and opportunities to enhance data sharing, were informative in understanding key factors that influence policy development and implementation at the organizational level e.g. communication / clarity of policy and processes,

and leadership support. As a result, some recommendations were developed from this case study and describe opportunities for how NIH and other federal agencies can improve data sharing among its funded investigators (Appendix H).

CULTURE

Since the Fort Lauderdale agreement, efforts have been made by NIH, as the largest funder of biomedical research, to promote broad sharing of data generated from research studies through the development of data sharing policies. While there are specific policies focused on sharing of genomic data (GWAS and GDS policies), there is none specific for epidemiology data, although it is expected that sharing of epidemiology data along with other data types are covered under the general NIH 2003 data sharing policy. The lack of specificity of this policy has been challenging in the enforcement / implementation of data sharing, particularly for the field of epidemiology which, unlike genomic data sharing, does not have a culture for sharing data.

This dissertation research showed that institutions lack a reward system that would incentivize its researchers to share data. Majority of the respondents in this case study mentioned that getting credit or recognition was a motivation for them to share data, but that this was not existent in current culture of their institutions. Not addressing this concern could have serious implications in the quality of data that is deposited in the repository, especially if there is no efficient mechanism in place for quality control and given the limited resources, time and effort in place for such activities.

The NCBI resources are set up to check the quality of the data submitted but due to lack of adequate administrative and technical support at the NIH end, there are delays in the data

sharing process through dbGaP. Investigators are pressured to comply with the policy may do the bare minimum to prepare their data before submitting to the repository, just so they can focus their efforts and resources on their research and at the same time meet the policy requirements. Therefore, changing the current reward structure at institutions and creating a system at academic institutions and at NIH to reward investigators who share data is important.

Some of the recommendations to change the reward structure at institutions are: 1) including data sharing as part of the promotion and tenure criteria at academic institutions; 2) considering a new metric for data sharing, the S-Index, in the evaluation of investigators careers, analogous to the H-index for publications; 3) NIH to reward investigators who have a track record for sharing; 4) NIH to set aside a percentage of funding in grant awards specifically for data sharing; and 5) NIH and academic institutions to consider collaborations as part of "broad" data sharing for large collaborative or consortia research studies, and recognize middle authorship from collaborative studies as equally important in the evaluation of academic research careers and funding opportunities.

POLICY

This case study revealed one of the most crucial factors underlying the issues with data sharing in NIH-supported research; the differences in the definition of data sharing. The concept of data sharing was confirmed in this study to be an elusive concept with varied definition by investigators and NIH in the study, as well as NIH staff at this case study practice site. In the policy, data sharing as defined by NIH implies sharing of data through submission in dbGaP or any of the NIH-supported data repositories. This is considered broad data sharing because the data is not controlled at the investigator or institution level, but rather with appropriate controls and approved access by NIH, the data is openly available to anyone for secondary use.

According to the investigators, and their institutions, the collaborative model of data sharing is considered to be the ideal, which is in contrast to the NIH expectations.

A lack of shared understanding of data sharing, if not addressed, will lead to ongoing issues with policy compliance and data sharing in general. Knowledge gained from the findings from this study illustrates the importance of clarity of policy and the need for engagement or buy-in from stakeholders to enact change. There's potential for this to impact how government policies are developed and implemented and the communication of processes for effective implementation.

Some of the recommendations pertinent to the lack of clarity of policy that was developed as a result of findings from this study are: 1) all stakeholders including NIH and non-NIH leaders should come together to clarify and define what "data sharing" means so there's uniformity in the language used in the policy, and implementation guidelines – this also requires engagement or buy-in from stakeholders; 2) NIH to reassess the policy timeline and expectations based on study an data types, with the understanding that there is not a one-size fits all policy for different data types; 3) NIH to identify more effective strategies beyond the RFI mechanism for soliciting feedback from the community on the development of policy, e.g. increased early and frequent engagement and communication with targeted key stakeholders, clear policies and processes, and training.

The enforcement of data sharing policies was observed in the interviews and document reviews as challenging. According to the data sharing policies, NIH project officers or program directors are charged with ensuring that their grantees comply with the NIH data sharing policies. Enforcement of the data sharing policies is primarily done by NIH staff. The findings

of this study revealed that there's a lack of consistency in the implementation of data sharing policies across different institutes and centers at NIH. In addition, not all journals are enforcing data sharing, which adds to the challenge. Changing the way that policies are enforced may influence the researcher's attitude and perspective on data sharing. Consistent and uniform enforcement of policy at the NIH and academic institutions, and among journals, will send the message to investigators that data sharing is a priority and is valued by their institutions.

Some recommendations for addressing this issue of enforcement are: 1) to bring on academic institutions as enforcers of data sharing policy, which will require buy-in from the institutions e.g. the NCI director could talk with cancer center directors and request that they support implementation / enforcement of NIH data policy among their researchers; 2) make the processes for implementation clear, consistent and transparent.

One of the concerns that was raised by the respondents was around privacy of participant data in research and consent. The NIH data sharing policies state the sharing of research data should be consistent with what's in the informed consent forms. This is evident in the NIH GDS policy protocol that requires a Certificate of Confidentiality be signed by institutions submitting the data to NIH, and the Data Use Limitation that indicates the scope and limitations of data use as described in the consent form. These are all efforts to ensure the highest standard or level of protection of participant data confidentiality in research. Despite this, investigators were still apprehensive about data submission in a controlled-access data repository such as dbGaP. This could be an opportunity for increasing awareness and NIH efforts through training and communication. One recommendation is for NIH to conduct an evidence-based research to assess participants' understanding of consent and data sharing. This may help investigators feel

more comfortable sharing data if they knowthat their study participants are supportive of their data being shared.

Data sharing in large-scale collaborative studies can be challenging especially with the different types of consent forms from different collaborating institutions. To help navigate the consent process across studies, the respondents suggested developing broad consent forms, standardized if possible, to make it easier to share data across multiple studies, with clear mechanisms in place to ensure protection of patient privacy.

RESOURCES

Prioritization of data sharing by funding agencies and academic institutions is critical to improving data sharing among investigators. The findings of this case study showed that investigators have limited resources in their grants to do research and properly prepare and document data for submission in a repository for others to use. They rely on support from funding agencies through grant awards which are very competitive, especially in this funding climate, as well as on their institutions to provide administrative or technical support for data sharing activities. The sentiment among the investigators was that institutions have other priorities that compete with data sharing, which ends up being a lower priority given the cost of data sharing and the fear of potential liability on the institution in the event of a breach in confidentiality of participant data.

Increased institutional or leadership support and investment in data sharing through the provision of administrative and financial resources e.g. funding to investigators, is a recommendation from this case study for a few reasons. It could lead to more positive attitude and increased compliance due to the reduced burden on investigators, and credit or recognition

given for sharing as described earlier. It could also be helpful to support ongoing submission of data in a data repository from grants that might have ended but still generating data. This data could be useful to answer new research questions that may not otherwise be answered without that data.

Although it is stated in the GDS policy that data sharing costs are allowed in the grant budget, not many investigators seemed to be aware of this. This specifically addresses the cost for data documentation, formatting, etc., to make it easier for sharing, especially phenotypic data in repositories. Therefore, it is recommended that there's increased awareness and communication about this, as well as how to budget for unanticipated data requests. Other suggestions were related to institutions providing a central support system with dedicated staff to support data sharing activities at the institutions; and for NIH to invest more in epidemiology data sharing which currently is more evident in genomic data sharing.

Training was an important factor in this study for improving data sharing in public or controlled-access databases. There are policies, processes and systems involved in data sharing that are not always clear, as indicated by the findings from this study. Therefore, having the right training will increase knowledge, skills, expertise, understanding and confidence around data sharing policy requirements, expectations, and processes at institutions and at NIH. This will facilitate and enhance the sharing of data generated from research studies among investigators.

Some of the recommendations related to training are: 1) incorporate data sharing as part of the curriculum at institutions e.g. training on reproducibility and metadata standards, as an early investment in investigator's careers to help shape their thinking around data sharing; 2)

hold workshops on data sharing for investigators, perhaps focusing on early career investigators who may not be familiar or experienced with data sharing; 3) ensure broad and bi-directional communication and training on data sharing processes, systems and tool e.g. dbGaP, targeting novice users; and 4) developing training materials, tools on new and existing resources, and materials to support data sharing activities by investigators and NIH staff.

TECHNOLOGICAL INFRASTRUCTURE

It was quite evident from the interviews and document reviews that the sub-optimal data repository capabilities were a barrier to data sharing. The FAIR Guiding Principles, a framework established and accepted by the scientific community, was developed to ensure that data is Findable, Accessible, Interoperable, and Reusable (Wilkinson et al, 2016). A big part of this is in the design of the repository and the requirements for the analytic data expected to be submitted to the repository. The technological infrastructure of data sharing is a huge investment by the NIH given the number of studies funded by the many institutes and centers of NIH. Robust, efficient, effective and adequate data repository is required for successful data submission and access. It requires increased investment in resources by NIH to modernize and enhance the existing repositories. Addressing this will facilitate data submission, minimize delays in quality control checks by NCBI staff who manage dbGaP, and improve quality and functionality of the repository.

One of the recommendations related to this is for NIH to increase its investment in engineering and technical support of dbGaP through provision of administrative/technical support and financial resources. In addition, automating and standardizing as much of the data

sharing process as possible will increase efficiency. The design of the repository should be such that makes it easy to find out what's in dbGaP; changing the way that data is submitted, currently in piecemeal, is important. Comparing other models of existing data repositories could provide ideas and opportunities for how to improve dbGaP. Archiving and analyzing data through the NCI Genomic Data Commons and the cloud are currently being explored for genomic and clinical data, but until these are fully functional, the challenges with dbGaP submission of data, especially epidemiology data continue.

The lack of clarity in the data submission / access process was one of the most recurring themes in the data – interviews and document reviews. There are many steps and people involved in the process prior to submission of data. The institution's IRB is set up to check and confirm that caution has been taken to share patient data in a data repository, and this requires a lot of administrative work by the investigators and the institutions. The various steps, processes, requirements tend to be overwhelming and require training or education to increase knowledge and build skills, and develop expertise with using data repository for data submission and access. Equally at the NIH end, the processes for data submission and access may not be as intuitive or clear as indicated by findings of this study.

It is important to address the issues of lack of clarity in processes so that all key stakeholders including investigators and funders are on the same page. The recommendation to identify mechanisms or approaches for increasing clarity of processes and access could be accomplished through: 1) training by the experts at NIH / NCBI staff, including education of what resources are available to support data sharing; 2) streamlined process for data transfer agreements; and 3) automation and standardization of the process as much as possible.

DATA USE, COST-BENEFIT AND VALUE

This was an emergent factor from the interviews and highlights the need to better understand the cost of investment in data sharing compared to the value or benefit of data sharing submitted to a data repository. Doing a cost-benefit analysis, as recommended by a few respondents in this study, may help reshape the perspectives and priorities of investigators, institutions and NIH in particular, leading to change.

The different analysis recommended for consideration by the NIH to help enhance data sharing are: 1) cost-benefit analysis of data sharing; 2) increase awareness of research publications that have already analyzed the use of data in data repositories (dbGaP and BioLINCC), as well as determine the efficiency of existing resources such as dbGaP. Additional recommendations for NIH to consider based on findings from this research are to conduct a focus group with a group of targeted investigators and staff, to help explore what aspects of the policy and submission / access process are not clear, as well as how to best address the issues related to the complexity of analytic data.

VI. CONCLUSION

The diversity in composition of the two groups of key informants (NIH staff and NIH-funded researchers) added to the richness of the data and perspectives on issues around data sharing in NIH-funded research. The depth of knowledge gained would not have been possible if the method of data collection was primarily done through surveys, even with open ended questions. The data would not have been as rich and the nuances and subtle differences around the data and examples provided would not have been otherwise captured.

A comparison across data sources revealed that in this study there were more factors coded as barriers across interviews and document reviews compared to facilitators or opportunities. There were 29% of the factors coded as facilitators in this study, 36% of factors coded as barriers, and 34% of factors coded as opportunities. Convergence of these data sources was considered when at least one factor was present in both the interview data and at least one of the documents reviewed. The major difference in the data sources was that CULTURE construct, specifically the factors *Intrinsic incentive* and *Differences in culture of research fields* as facilitators, were only evident in the interviews and not the document reviews.

The contribution of the findings of this research to the existing body of knowledge on data sharing is invaluable because of the unique perspectives supported by recurring themes in existing reports / previous studies, which helped increase the internal validity of this study. The expectation prior to the onset of this dissertation project was that the NIH GDS policy was a good model for data sharing policies and that the experiences of researchers whose focus is on genomics would have different experiences than epidemiologist given the culture or history of data sharing in the field of genomics. The findings from this research showed that there were no major differences between the groups of investigators, and that the genomicists/genetic

epidemiologists and the epidemiologist shared similar sentiments in terms of what factors they considered as facilitating or hindering data sharing. The lessons learned from genomic data sharing that could be applied to enhance epidemiology data sharing, as well as the opportunities garnered from the data would be helpful with informing future directions in the development and implementation of data sharing policies. The insights provided by the new investigators on factors that facilitate or hinder data sharing as well as the opportunities, were corroborated by the experienced investigators and NIH staff who participated in this study.

The findings in this dissertation research led to the generation of knowledge around factors and the relationships between those factors that facilitate and hinder data sharing, as well as opportunities to apply knowledge gained and lessons learned from genomic data sharing to epidemiology data sharing. Developing a culture of data sharing for epidemiology studies will take time and will require buy-in from stakeholders to make the culture shift. Understanding the key factors that influence data sharing at the organizational level, and the relationships between these factors will help with the improving data sharing overall among researchers generating data from NIH-funded research.

The knowledge generated from this research was used in the development of a set of recommendations that hopefully will inform changes in policy, development of new policies, and enhanced implementation strategies. It is hoped that these will result in broader and enhanced data sharing in NIH-funded research, which will advance scientific progress and ultimately provide public health and clinical benefit.

CITED LITERATURE

Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JP. Public availability of published research data in high-impact journals. *PloS one*. 2011;6(9):e24357.

Arias JJ, Pham-Kanter G, Campbell EG. The growth and gaps of genetic data sharing policies in the united states. *Journal of Law and the Biosciences*. 2015;2(1):56-68.

Ascoli GA, Maraver P, Nanda S, Polavaram S, Armañanzas R. Win-win data sharing in neuroscience. *Nature Methods*. 2017;14(2):112-116.

Arzberger P, Schroeder P, Beaulieu A, et al. Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*. 2004;3:135-152.

Bierer BE, Crosas M, Pierce HH. Data authorship as an incentive to data sharing. *N Engl J Med*. 2017;376(17):1684-1687. doi: 10.1056/NEJMsb1616595 [doi].

Borgman CL. The conundrum of sharing research data. *J Am Soc Inf Sci Technol*. 2012;63(6):1059-1078.

Bruna EM. Scientific journals can advance tropical biology and conservation by requiring data archiving. *Biotropica*. 2010;42(4):399-401.

Butlin R. Data archiving. *Heredity*. 2011;106(5):709. doi:10.1038/hdy.2010.43.

Coady SA, Mensah GA, Wagner EL, Goldfarb ME, Hitchcock DM, Giffen CA. Use of the national heart, lung, and blood institute data repository. *N Engl J Med*. 2017;376(19):1849-1858.

Corpas M, Kovalevskaya NV, McMurray A, Nielsen FG. A FAIR guide for data providers to maximise sharing of human genomic data. *PLoS computational biology*. 2018;14(3):e1005873.

Creswell, JW. (2014). Research Design. Qualitative, quantitative and mixed methods approaches. Los Angeles, London, New Delhi, Singapore, Washington, DC: Sage Publications, Inc.

Enke N, Thessen A, Bach K, Bendix J, Seeger B, Gemeinholzer B. The user's view on biodiversity data sharing—Investigating facts of acceptance and requirements to realize a sustainable use of research data—. *Ecological Informatics*. 2012;11:25-33.

Fox CW, Irschick DJ, Knapp AK, Thompson K, Baker L, Meyer J. Functional ecology: Moving forward into a new era of publishing. *Funct Ecol*. 2014;28(2):291-292.

Frasier M. Perspective: Data sharing for discovery. *Nature*. 2016;538(7626):S4-S4.

Friesike DS, Fecher B, Hebing M, Linek S. Reputation instead of obligation: Forging new policies to motivate academic data sharing. *Impact of Social Sciences Blog*. 2015.

Gardner D, Toga AW, Ascoli GA, et al. Towards effective and rewarding data sharing. *Neuroinformatics*. 2003;1(3):289-295.

Harris, R. (2017). Rigor Mortis: How sloppy science creates worthless cures, crushes hope, and wastes billions. New York, NY: Basic Books.

Holdren J. Memorandum for the heads of executive departments and agencies: Increasing access to the results of federally funded scientific research. office of science and technology policy. 2013. *Data sharing: empty archives*.

Hudson KL, Collins FS. The 21st century cures Act—A view from the NIH. *N Engl J Med*. 2016.

Ioannidis JP, Khoury MJ. Assessing value in biomedical research: The PQRST of appraisal and reward. *JAMA*. 2014;312(5):483-484.

Kaye J, Heeney C, Hawkins N, De Vries J, Boddington P. Data sharing in genomics—re-shaping scientific practice. *Nature Reviews Genetics*. 2009;10(5):331-335.

Kim Y, Stanton JM. Institutional and individual influences on scientists' data sharing practices. *Journal of Computational Science Education*. 2012;3(1):47-56.

Kim Y, Adler M. Social scientists' data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories. *Int J Inf Manage*. 2015;35(4):408-418.

Kim Y, Stanton JM. Institutional and individual factors affecting scientists' data- sharing behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology*. 2016;67(4):776-799.

Knoppers BM. Framework for responsible sharing of genomic and health-related data. *The HUGO journal*. 2014;8(1):3.

Koers H. How do we make it easy and rewarding for researchers to share their data? A publisher's perspective. *J Clin Epidemiol*. 2016;70:261.

Laine H. Afraid of scooping – case study on researcher strategies against fear of scooping in the context of open science. *Data Science Journal*. 2017;16(29):1-14.

Longo DL, Drazen JM. Data sharing. *N Engl J Med*. 2016;374(3):276-277. http://dx.doi.org/10.1056/NEJMe1516564. doi: 10.1056/NEJMe1516564.

Marquardt MJ, Leonard HS, Freedman AM, Hill CC. *Action learning for developing leaders and organizations: Principles, strategies, and cases*. American Psychological Association; 2009

Maxwell JA. *Qualitative research design: An interactive approach.* Vol 41. Sage publications; 2012.

Midgley G. Systemic intervention for public health. *Am J Public Health*. 2006;96(3):466-472.

Mueller-Langer F, Andreoli Versbach P. Open access to research data: Strategic delay and the ambiguous welfare effects of mandatory data disclosure. 2014.

National Institutes of Health. *Plan for increasing access to scientific publications and digital scientific data from NIH funded scientific research*. 2015.

National Library of Medicine (NLM), (n.d.) NIH Data Sharing Policies. Retrieved from https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_policies.html

Nelson B. Empty archives: Most researchers agree that open access to data is the scientific ideal, so what is stopping it happening? Bryn Nelson investigates why many researchers choose not to share. *Nature*. 2009;461(7261):160-164.

NIH Data Science: Open Science Symposium: How open data and open science are transforming biomedical research, December 1, 2016. Retrieved from https://datascience.nih.gov/OpenDataScienceSymposiumCal

NIH Genomic Data Sharing (GDS), (n.d.) Retrieved from https://gds.nih.gov/

NIH Guide Notice (2002). NIH Announces Draft Statement on Sharing Research Data. Retrieved from http://grants.nih.gov/grants/guide/notice-files/NOT-OD-02-035.html

NIH Guide Notice (2003). Final NIH Statement on Sharing Research Data. Retrieved from http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

NIH Office of Extramural Research (OER). NIH Data Sharing Policy (n.d.). Retrieved from http://grants.nih.gov/grants/policy/data_sharing/

NIH Office of the Director (NIH OD). (n.d.) Research, Funding and Coordination. Retrieved from https://www.nih.gov/institutes-nih/nih-office-director/research-funding-coordination

NIH Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research. February 2015. Retrieved from https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf

NIH Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS), (n.d.) Retrieved from https://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html

Olfson M, Wall MM, Blanco C. Incentivizing data sharing and collaboration in medical research—the s-index. *JAMA psychiatry*. 2017;74(1):5-6.

Paltoo DN, Rodriguez LL, Feolo M, et al. Data Use under the NIH GWAS Data Sharing Policy and Future Directions. *Nature genetics*. 2014;46(9):934-938. doi:10.1038/ng.3062.

Parker M, Bull S. Ethics in collaborative global health research networks. *Clinical Ethics*. 2009;4(4):165-168.

Patton MQ. *Qualitative evaluation and research methods.* SAGE Publications, inc; 2015.

Pham-Kanter G, Zinner DE, Campbell EG. Codifying collegiality: Recent developments in data sharing policy in the life sciences. *PLoS One*. 2014;9(9):e108451.

Pisani E, AbouZahr C. Sharing health data: Good intentions are not enough. *Bull World Health Organ*. 2010;88(6):462-466.

Pisani E, Whitworth J, Zaba B, Abou-Zahr C. Time for fair trade in research data. *The Lancet*. 2010;375(9716):703-705.

Piwowar HA, Becich MJ, Bilofsky H, Crowley RS. Towards a data sharing culture: Recommendations for leadership from academic health centers. *PLoS medicine*. 2008;5(9):e183.

Poline J, Breeze JL, Ghosh SS, et al. Data sharing in neuroimaging research. *Frontiers in neuroinformatics*. 2012;6:9.

Rolland, B. Blog: Data Sharing and Reuse: Expand our concept of collaborations. Posted March 29, 2016. Retrieved from https://www.teamsciencetoolkit.cancer.gov/public/ExpertBlog.aspx?tid=4

Rowhani-Farid A, Allen M, Barnett AG. What incentives increase data sharing in health and medical research? A systematic review. *Research Integrity and Peer Review* [Incentives]. 2017;2(4):1-2-10.

Sane, J., and Edelstein, M. (2015). Overcoming barriers to data sharing in public health: a global perspective. London: Chatham House, the Royal Institute of International Affairs.

Sane J, Edelstein M. Overcoming barriers to data sharing in public health. *A global perspective*. 2015.

Savage CJ, Vickers AJ. Empirical study of data sharing by authors publishing in PLoS journals. *PloS one*. 2009;4(9):e7078.

Stanley B, Stanley M. Data sharing: The primary researcher's perspective. *Law Hum Behav*. 1988;12(2):173.

Stokols D, Misra S, Moser RP, Hall KL, Taylor BK. The ecology of team science: Understanding contextual influences on transdisciplinary collaboration. *Am J Prev Med*. 2008;35(2):S96-S115.

Sturges, P., Bamkin, M., Anders, J. H., Hubbard, B., Hussain, A., & Heeley, M. (2015). Research data sharing: Developing a stakeholder- driven model for journal policies. *Journal of the Association for Information Science and Technology*, *66*(12), 2445-2455.

Taichman DB, Backus J, Baethge C, et al. Sharing clinical trial data: A proposal from the international committee of medical journal EditorsSha ring clinical trial data. *Ann Intern Med*. 2016;164(7):505-506.

Taichman DB, Backus J, Baethge C, et al. Sharing clinical trial data: A proposal from the international committee of medical journal editors. *CMAJ*. 2016;188(2):91-92. doi: 10.1503/cmaj.151465 [doi].

Tenopir C, Allard S, Douglass K, et al. Data sharing by scientists: Practices and perceptions. *PloS one*. 2011;6(6):e21101.

Tenopir C, Dalton ED, Allard S, et al. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS One*. 2015;10(8):e0134826.

Vasilevsky NA, Minnier J, Haendel MA, Champieux RE. Reproducible and reusable research: Are journal data sharing policies meeting the mark? *PeerJ*. 2017;5:e3208.

Wallis JC, Rolando E, Borgman CL. If we share data, will anyone use them? data sharing and reuse in the long tail of science and technology. *PloS one*. 2013;8(7):e67332.

The White House. "Precision Medicine Initiative." January 30, 2015. Retrieved from https://www.whitehouse.gov/precision-medicine.

The White House, Office of the Press Secretary. "FACT SHEET: Investing in the National Cancer Moonshot." February 1, 2016. Retrieved from https://www.whitehouse.gov/the-press-office/2016/02/01/fact-sheet-investing-national-cancer-moonshot.

Wikipedia (n.d.). Bermuda Principles. Retrieved from https://en.wikipedia.org/wiki/Bermuda_Principles

Wikipedia (n.d.). Data Sharing. Retrieved from https://en.wikipedia.org/wiki/Data_sharing

Wikipedia (n.d.). Fort Lauderdale Agreement. Retrieved from https://en.wikipedia.org/wiki/Fort_Lauderdale_Agreement

Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific data*. 2016;3:160018.

Yin RK. *Case study research: Design and methods.* Sage publications; 2009.

Yozwiak NL, Schaffner SF, Sabeti PC. Make outbreak research open access. *Nature*. 2015;518(7540):477.

Zinner DE, Pham-Kanter G, Campbell EG. The changing nature of scientific sharing and withholding in academic life sciences research: Trends from national surveys in 2000 and 2013. *Acad Med*. 2016;91(3):433-440. doi: 10.1097/ACM.0000000000001028 [doi].

APPENDICES

Appendix A: UIC IRB letter of exemption

Exemption Granted

April 11, 2018

Nonye Harvey
Community Health Sciences
13110 Brewers Tavern Terrace
Clarksburg, MD 20871
Phone: (703) 508-2297

RE: Research Protocol # 2018-0324
"Data Sharing in Biomedical Research: A case study of data sharing practices and opportunities in NIH-funded research"
Sponsor(s): None

Dear Ms. Harvey:

Your Claim of Exemption was reviewed on April 11, 2018 and it was determined that your research protocol meets the criteria for exemption as defined in the U. S. Department of Health and Human Services Regulations for the Protection of Human Subjects [45 CFR 46.101(b)]. You may now begin your research.

Exemption Period: **April 11, 2018 – April 11, 2-21**
Lead Performance Site: NIH/NCI
Other Site(s): UIC
Subject Population: Adult (18+ years) NIH staff and NIH funded researchers only
Number of Subjects: 172
The specific exemption category under 45 CFR 46.101(b) is:

2 Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless: (i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.

Please note the Review History of this submission:

Receipt Date	Submission Type	Review Process	Review Date	Review Action
03/14/2018	Initial Review	Exempt	03/22/2018	Modifications Required
03/27/2018	Response To Modifications	Exempt	04/11/2018	Approved

Appendix A (continued)

You are reminded that investigators whose research involving human subjects is determined to be exempt from the federal regulations for the protection of human subjects still have responsibilities for the ethical conduct of the research under state law and UIC policy. Please be aware of the following UIC policies and responsibilities for investigators:

1. Amendments You are responsible for reporting any amendments to your research protocol that may affect the determination of the exemption and may result in your research no longer being eligible for the exemption that has been granted.

2. Record Keeping You are responsible for maintaining a copy all research related records in a secure location in the event future verification is necessary, at a minimum these documents include: the research protocol, the claim of exemption application, all questionnaires, survey instruments, interview questions and/or data collection instruments associated with this research protocol, recruiting or advertising materials, any consent forms or information sheets given to subjects, or any other pertinent documents.

3. Final Report When you have completed work on your research protocol, you should submit a final report to the Office for Protection of Research Subjects (OPRS).

4. Information for Human Subjects UIC Policy requires investigators to provide information about the research to subjects and to obtain their permission prior to their participating in the research. The information about the research should be presented to subjects as detailed in the research protocol, application and supporting documents.

Please be sure to use your research protocol number (2018-0324) on any documents or correspondence with the UIC IRB concerning your research protocol.

We wish you the best as you conduct your research. If you have any questions or need further help, please contact the OPRS office at (312) 996-1711 or me at (312) 355-5438. Please send any correspondence about this protocol to OPRS via OPRS Live.

Sincerely,

Tina S. Johnson, MA
IRB Coordinator, IRB # 7
Office for the Protection of Research Subjects

cc: Jesus Ramirez-Valles, Community Health Sciences, M/C 923
Kristina Risley, Community Health Sciences, M/C 923

Appendix B: NIH IRB letter of exemption

From: OHSRP Determinations
Sent: Thursday, March 01, 2018 11:49 AM
To: Harvey, Chinonye (NIH/NCI) [E] <harveyn@mail.nih.gov>
Cc: Grant, Nicole (NIH/NCI) [E] <grantn@mail.nih.gov>
Subject: OHSRP Determination '18-NCI-00479' - Excluded from IRB Review

Date: 3/1/2018
SI Name: Harvey, Chinonye (NCI)
OHSRP ID#: 18-NCI-00479
Project Title: Data Sharing in Biomedical Research: A Case Study of Data Sharing Practices and Opportunities in NIH-funded Research

The activity listed above is Excluded from IRB Review per 45 CFR 46 and NIH policy for the use of interview procedures.

This research is exempt because it will involve the use of interview procedures; and if the information obtained will be recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects, the disclosure of the human subjects' responses outside the research will not reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.

You may proceed.

Please retain this documentation and the attached PDF copy of your submission, as you would other research records. You may also access the PDF by copying and pasting the following link into your browser: https://ohsr.od.nih.gov/determinations/GeneratePDFReport.php?recordID=2011. Additionally, retain any supporting documentation such as de-identification or Honest Broker agreements, proof of which must be provided to OHSRP upon request.

Changes (amendments) to the research must be submitted to OHSRP for review prior to initiation, as changes may affect this determination. To request amendments, please return to the OHSRP determination website to amend this project with changes to this research activity, OHSRP ID#: 18-NCI-00479 - https://ohsr.od.nih.gov/determinations/Start.php.

If you have any questions or need further assistance, please feel free to contact us.

Sincerely,

Office of Human Subjects Research Protections (OHSRP) National Institutes of Health 301-402-3444-Office ohsrp_determinations@mail.nih.gov OHSRP website: https://ohsr.od.nih.gov/nih/index.php (NIH Login required)

Appendix C: Measurement Table

Initial measurement table

Main Research Question 1: How do organizational / institutional level factors facilitate or hinder the sharing of research data in public or controlled-access databases or data repositories?
Sub research question 1a: What are the organizational / institutional level factors that facilitate or hinder the sharing of research data in public or controlled-access databases or data repositories?
Sub research question 1b: How do these organizational / institutional factors facilitate or hinder the sharing of research data in public or controlled-access databases or data repositories?

Construct(s)	**Factors**	**Measures**	**Data Sources**	**Analysis Plan and Triangulation**
Technological Infrastructure	Funding	Financial and physical resources needed to develop, manage and sustain the infrastructure	• In-depth interviews (researcher and staff) • Document reviews	Thematic coding and Pattern Matching - identify and code key terms / phrases using a-priori coding and notation of emergent themes; - collate themes and codes in a matrix Review and compare key themes, patterns and use Atlas.Ti to explore relationships among the codes, for concordance or discordance. Summarize and interpret findings based on patterns. Triangulation of data related to each construct/factor across data sources at the following levels: - researcher interviews vs. NIH staff interviews
	Expertise	Technical expertise to develop, manage and use the infrastructure	• In-depth interviews (researcher and staff) • Document reviews	
	Access	Access to data in data repositories / databases	• In-depth interviews (researcher and staff) • Document reviews	
Regulatory Policy and Law	Existing policies	Description of existing policies	• In-depth interviews (researcher and staff) • Document reviews	
	Clarity of policies	Communication	• In-depth interviews (researcher and staff) • Document	

			reviews	(based on responses to similar questions asked) - researcher & NIH staff interviews vs. document reviews Member check-in with selection of a sub-set of interviewees to review responses for accuracy and analysis purposes.
	Enforcement	Enforcement of policy by organizational authorities	• In-depth interviews (researcher and staff) • Document reviews	
	Compliance	Individual researcher's perspective on compliance	• In-depth interviews (researcher and staff) • Document reviews	
Institutional and Organizational practices	Incentives and Rewards	Description of existing policy for incentives and rewards for individual researchers to share data	• In-depth interviews (researcher and staff) • Document reviews	
	Academic institutional norms	Culture, beliefs, practices of institution and researchers. Existing policies and practices related to academic Promotion/Tenure (Formal and informal practices)	• In-depth interviews (researcher and staff)	
Main Research Question 2: **What are opportunities for improving / enhancing the sharing of federally funded research data in public or controlled-access databases or data repositories?**				
Construct (s)	**Factors**	**Measures**	**Data Sources**	**Analysis Plan and Triangulation**
Support	Needed administrative support	Identified and perceived needs requiring administrative support such as personnel / staffing needs.	• In-depth interviews (researcher and staff) • Document reviews	Thematic coding and Pattern Matching - identify and code key terms / phrases using a-priori coding and notation of emergent themes; - collate themes and codes in a matrix
	Needed financial support	Identified and perceived needs requiring financial support	• In-depth interviews (researcher and staff)	

			• Document reviews	Review and compare key themes, patterns and use Atlas.Ti to explore relationships among the codes, for concordance or discordance. Summarize and interpret findings based on patterns. Triangulation of data related to each construct/factor across data sources at the following levels: - researcher interviews vs. NIH staff interviews (based on responses to related questions asked) - researcher & NIH staff interviews vs. document reviews Member check-in with selection of a sub-set of interviewees to review responses for accuracy and analysis purposes.
	Leadership support needed	Identified and perceived needs requiring support from organizational / institutional leadership	• In-depth interviews (researcher and staff) • Document reviews	
	Training needed	Self-efficacy	• In-depth interviews (researcher and staff) • Document reviews	
Academic institutional / organizational culture	Needed changes in institutional / organizational culture to facilitate data sharing	Incentives, policies and practices around systems Practice: Perceptions about what's acceptable	• In-depth interviews (researcher and staff) • Document reviews	
Technological Infrastructure	Needed Funding	Financial and physical resources needed to develop, manage and sustain the infrastructure	• In-depth interviews (researcher and staff) • Document reviews	
	Expertise needed	Technical expertise to develop, manage and use the infrastructure	• In-depth interviews (researcher and staff) • Document reviews	
	Needed Access and Submission	Access to data in data repositories / databases	• In-depth interviews (researcher and staff) • Document reviews	
Main Research Question 3: How can what has been learned from genomic data sharing be transferred to epidemiological data sharing? Sub research question 3a: What has been learned from genomic data sharing that could support epidemiological data sharing? Sub research question 3b: In what ways can these lessons learned support epidemiological data sharing?				

Construct (s)	**Factors**	**Measures**	**Data Sources**	**Analysis Plan and Triangulation**
Technological Infrastructure	Funding	Financial and physical resources needed to develop, manage and sustain the infrastructure	• In-depth interviews (researcher and staff) • Document reviews	Thematic coding and Pattern Matching - identify and code key terms / phrases using a-priori coding and notation of emergent themes; - collate themes and codes in a matrix Review and compare key themes, patterns and use Atlas.Ti to explore relationships among the codes, for concordance or discordance. Summarize and interpret findings based on patterns. Triangulation of data related to each construct/factor across data sources at the following levels: - researcher interviews vs. NIH staff interviews (based on responses to similar questions asked) - researcher & NIH staff interviews vs. document reviews
	Expertise	Technical expertise to develop, manage and use the infrastructure	• In-depth interviews (researcher and staff) • Document reviews	
	Access	Access to data in data repositories / databases	• In-depth interviews (researcher and staff) • Document reviews	
Regulatory Policy and Law	Existing policies	Description of existing policies	• In-depth interviews (researcher and staff) • Document reviews	
	Clarity of policies	Communication	• In-depth interviews (researcher and staff) • Document reviews	
	Enforcement	Enforcement of policy by organizational authorities	• In-depth interviews (researcher and staff) • Document reviews	
	Compliance	Individual researcher's perspective on compliance	• In-depth interviews (researcher and	

			staff) • Document reviews	Member check-in with selection of a sub-set of interviewees to review responses for accuracy and analysis purposes.
Institutional and Organizational practices	Incentives and Rewards	Description of existing policy for incentives and rewards for individual researchers to share data	• In-depth interviews (researcher and staff) • Document reviews	
	Academic institutional norms	Culture, beliefs, practices of institution and researchers. Existing policies and practices related to academic Promotion/Tenure (Formal and informal practices)	• In-depth interviews (researcher and staff)	
Support	Needed administrative support	Identified and perceived needs requiring administrative support such as personnel / staffing needs.	• In-depth interviews (researcher and staff) • Document reviews	Thematic coding and Pattern Matching - identify and code key terms / phrases using a-priori coding and notation of emergent themes; - collate themes and codes in a matrix Review and compare key themes, patterns and use Atlas.Ti to explore relationships among the codes, for concordance or discordance. Summarize and interpret findings based on patterns. Triangulation of data related to each construct/factor
	Needed financial support	Identified and perceived needs requiring financial support	• In-depth interviews (researcher and staff) • Document reviews	
	Leadership support needed	Identified and perceived needs requiring support from organizational / institutional leadership	• In-depth interviews (researcher and staff) • Document reviews	
	Training needed	Self-efficacy	• In-depth interviews (researcher and staff) Document reviews	
Academic institutional / organizational	Needed changes in institutional /	Incentives, policies and practices around systems	• In-depth interviews	

culture	organizational culture to facilitate data sharing	Practice: Perceptions about what's acceptable	(researcher and staff) • Document reviews	across data sources at the following levels: - researcher interviews vs. NIH staff interviews (based on responses to related questions asked) - researcher & NIH staff interviews vs. document reviews Member check-in with selection of a sub-set of interviewees to review responses for accuracy and analysis purposes.

Appendix C (continued)

Revised measurement table

Main Research Question 1: How do organizational / institutional level factors facilitate or hinder the sharing of research data in public or controlled-access databases or data repositories?
Sub research question 1a: What are the organizational / institutional level factors that facilitate the sharing of research data in public or controlled-access databases or data repositories?

Sub research question 1b: How do these organizational / institutional factors facilitate the sharing of research data in public or controlled-access databases or data repositories?

Sub research question 1c: What are the organizational / institutional level factors that hinder the sharing of research data in public or controlled-access databases or data repositories?

Sub research question 1d: How do these organizational / institutional factors hinder the sharing of research data in public or controlled-access databases or data repositories?

Construct(s)	**Factors**	**Measures**	**Data Sources**	**Analysis Plan and Triangulation**
Institutional / Organizational Culture and Practices	Intrinsic incentives	Internal motivation to share for the advancement of science and personal benefit	In-depth interviews (researcher and staff) Document reviews	Thematic coding and Pattern Matching - identify and code key terms / phrases using a-priori coding and notation of emergent themes; - collate themes and codes. Review and compare key themes, patterns and use NVivo to explore relationships among the codes, for concordance or discordance. Summarize and
	Institutional reward system for data sharing	Existing system at institutions for rewarding researchers who share data including promotion and tenure	In-depth interviews (researcher and staff) Document reviews	
	Individual career concerns	Concerns related to scooping, misinterpretation and misuse of data	In-depth interviews (researcher and staff) Document reviews	

	Culture differences in research fields	Differences in the culture, practices and perception of data sharing in different fields	In-depth interviews (researcher and staff) Document reviews	interpret findings based on patterns. Triangulation of data related to each construct/factor across data sources at the following levels: - researcher interviews vs. NIH staff interviews (based on responses to similar questions asked) - researcher & NIH staff interviews vs. document reviews Member check-in with selection of a sub-set of interviewees to review responses for accuracy and analysis purposes
Regulatory Policy and Law	Privacy concerns	Concerns in current policies related to participant confidentiality and consent.	In-depth interviews (researcher and staff) Document reviews	
	Clarity of policies	Clarity or vagueness and communication around policy requirements	In-depth interviews (researcher and staff) Document reviews	
	Enforcement	Enforcement of policy by institutional / organizational officials at NIH	In-depth interviews (researcher and staff) Document reviews	
	Definition of data sharing	Perception and understanding of the definition of the term 'data sharing'.	In-depth interviews (researcher and staff) Document reviews	
Resources	Administrative / Technical resources	Administrative / technical support such as personnel, staffing needs required to support data sharing	In-depth interviews (researcher and staff) Document reviews	Thematic coding and Pattern Matching - identify and code key terms / phrases using a-priori coding and notation of

	Financial resources	Financial cost and resources required to support data sharing	In-depth interviews (researcher and staff) Document reviews	emergent themes; - collate themes and codes in a matrix Review and compare key themes, patterns and use NVivo to explore relationships among the codes, for concordance or discordance. Summarize and interpret findings based on patterns. Triangulation of data related to each construct/factor across data sources at the following levels: - researcher interviews vs. NIH staff interviews (based on responses to similar questions asked) - researcher & NIH staff interviews vs. document reviews Member check-in
	Leadership support	Support of leadership through provision of resources and guidance and oversight	In-depth interviews (researcher and staff) Document reviews	
	Training	Training for researchers and staff to enhance knowledge	In-depth interviews (researcher and staff) Document reviews	
Technological Infrastructure	Repository capabilities	Capacity, effectives, adequacy and efficiency of data repositories	In-depth interviews (researchers and staff) Document reviews	
	Expertise	Technical expertise to use, and manage submission/access processes for repository	In-depth interviews (researchers and staff) Document reviews	
	Clarity of submission / access process	Existing processes / guidelines/ tools in place to assist with data submission /access	In-depth interviews (researchers and staff) Document reviews	

	Analytic data complexity	Requirements/ standardization of data format, documentation and preparation for submission in repository	In-depth interviews (researchers and staff) Document reviews	with selection of a sub-set of interviewees to review responses for accuracy and analysis purposes.
Main Research Question 2: What are opportunities for improving / enhancing the sharing of federally funded research data in public or controlled-access databases or data repositories?				
Construct (s)	**Factors**	**Measures**	**Data Sources**	**Analysis Plan and Triangulation**
Institutional / Organizational Culture and Practices	Reward structure changes needed	Incentives, policies or efforts at institutions to reward, recognize or credit data sharers	In-depth interviews (researcher and staff) Document reviews	Thematic coding and Pattern Matching - identify and code key terms / phrases using a-priori coding and notation of emergent themes; - collate themes and codes in a matrix Review and compare key themes, patterns and use NVivo to explore relationships among the codes, for concordance or discordance. Summarize and interpret findings based on patterns. Triangulation of data related to each
	Culture shift in research fields	Shift in thinking and perception around sharing data in different fields of research	In-depth interviews (researcher and staff) Document reviews	
Regulatory policy and law	Addressing privacy concerns	Potential changes and opportunities to address privacy concerns through policy	In-depth interviews (researcher and staff) Document reviews	
	Clarity of policy needed	Changes to clarify policy requirements and expectations	In-depth interviews (researcher and staff) Document reviews	
	Change needed in enforcement	uniform, consistent, clear implementation strategies and enforcement mechanisms	In-depth interviews (researcher and staff) Document reviews	

Resources/ Support	Addressing administrative needs	Identified and perceived needs requiring administrative support such as personnel / staffing needs.	In-depth interviews (researcher and staff) Document reviews	construct/factor across data sources at the following levels: - researcher interviews vs. NIH staff interviews (based on responses to related questions asked) - researcher & NIH staff interviews vs. document reviews Member check with selection of a sub-set of interviewees to review responses for accuracy and analysis purposes.
	Addressing financial needs	Identified and perceived needs requiring financial support	In-depth interviews (researcher and staff) Document reviews	
	Leadership support needed	Identified and perceived needs requiring support from organizational / institutional leadership	In-depth interviews (researcher and staff) Document reviews	
	Training needed	Identified training needs and opportunities to build skills	In-depth interviews (researcher and staff) Document reviews	
Technological Infrastructure	Addressing repository capabilities	Identified repository needs related to capacity, effectives, adequacy and efficiency	In-depth interviews (researchers and staff) Document reviews	
	Expertise needed	Identified need to enhance technical expertise to use, and manage submission / access processes	In-depth interviews (researchers and staff) Document reviews	
	Clarity of submission /access process needed	Identified needs related to improving clarity of existing processes / guidelines/ tools in place to assist with data submission	In-depth interviews (researchers and staff) Document reviews	

		/access		
	Addressing analytic data complexity	Identified needs related to data format, documentation and preparation	In-depth interviews (researchers and staff) Document reviews	
OTHER- Data Use, Cost and Value	Addressing data use, cost and value / benefit	Identified need for cost-benefit analysis of data, value/ use of data, and risk assessment of data shared in data repositories	In-depth interviews (researchers and staff)	
Main Research Question 3: How can what has been learned from genomic data sharing be transferred to epidemiological data sharing? **Sub research question 3a:** What has been learned from genomic data sharing that could support epidemiological data sharing? **Sub research question 3b:** In what ways can these lessons learned support epidemiological data sharing?				
Construct (s)	**Factors**	**Measures**	**Data Sources**	**Analysis Plan and Triangulation**
Institutional / Organizational Culture and Practices	Intrinsic incentives	Internal motivation to share for the advancement of science and personal benefit	In-depth interviews (researcher and staff) Document reviews	Thematic coding and Pattern Matching - identify and code key terms / phrases using a-priori coding and notation of emergent themes; - collate themes and codes in a matrix
	Institutional reward system for data sharing	Existing system at institutions for rewarding researchers who share data including promotion and tenure	In-depth interviews (researcher and staff) Document reviews	

	Individual career concerns	Concerns related to scooping, misinterpretation and misuse of data	In-depth interviews (researcher and staff) Document reviews	Review and compare key themes, patterns and use NVivo to explore relationships among the codes, for concordance or discordance. Summarize and interpret findings based on patterns. Triangulation of data related to each construct/factor across data sources at the following levels: - researcher interviews vs. NIH staff interviews (based on responses to similar questions asked) - researcher & NIH staff interviews vs. document reviews Member check-in with selection of a sub-set of interviewees to review responses for accuracy and analysis purposes.
	Culture differences in research fields	Differences in the culture, practices and perception of data sharing in different fields	In-depth interviews (researcher and staff) Document reviews	
Regulatory Policy and Law	Privacy concerns	Concerns in current policies related to participant confidentiality and consent.	In-depth interviews (researcher and staff) Document reviews	
	Clarity of policies	Clarity or vagueness and communication around policy requirements	In-depth interviews (researcher and staff) Document reviews	
	Enforcement	Enforcement of policy by institutional / organizational officials at NIH	In-depth interviews (researcher and staff) Document reviews	
Resources	Administrative / Technical resources	Administrative / technical support such as personnel, staffing needs required to support data sharing	In-depth interviews (researcher and staff) Document reviews	

	Financial resources	Financial cost and resources required to support data sharing	In-depth interviews (researcher and staff) Document reviews	
	Leadership support	Support of leadership through provision of resources and guidance and oversight	In-depth interviews (researcher and staff) Document reviews	
	Training	Training for researchers and staff to enhance knowledge	In-depth interviews (researcher and staff) Document reviews	
Technological Infrastructure	Repository capabilities	Capacity, effectives, adequacy and efficiency of data repositories	In-depth interviews (researcher and staff) Document reviews	
	Expertise	Technical expertise to use, and manage submission/access processes for repository	In-depth interviews (researcher and staff) Document reviews	
	Clarity of submission / access process	Existing processes / guidelines/ tools in place to assist with data submission /access	In-depth interviews (researcher and staff)	

			Document reviews	
	Analytic data complexity	Requirements/ standardization of data format, documentation and preparation for submission in repository	In-depth interviews (researcher and staff) Document reviews	

Appendix D: Interview guides

For NIH-funded researchers

Introduction

Thank you for taking the time to speak with me today about my research study. My name is Nonye Harvey and I am in the Epidemiology and Genomics Research Program at the National Cancer Institute. I am currently pursuing a Doctor of Public Health degree in Leadership in Public Health at the University of Illinois Chicago School of Public Health.

My thesis is on data sharing in biomedical research and I am interested in hearing about your experiences with the **sharing of de-identified genomic and epidemiological data**, and [organizational / institutional level] **factors you perceive as facilitating or hindering the sharing of these data** in public or controlled-access databases.

There are no right or wrong answers to the interview questions. Your participation is voluntary and please be assured that all responses will be de-identified and kept confidential. Is it okay if I make an audio record of this discussion to complement my notes and ensure that I capture all points made and accurately represent your views?

The recording will be transcribed and the information you give will only be used for this research study.

This interview will consist of 29 questions with some follow-ups for certain questions. This should last about 45 minutes. Before we begin, do you have any questions for me?

A. Opening Questions

1. Please describe your current position at your institution.
    a. Probe: What is your official job title?
    b. Probe: How long have you been at your current institution?
    c. Probe: How many years have you been doing research at your current institution or the institution where you were the longest?

2. What is your main research area of focus?
    a. Probe: Is your work primarily focused on Epidemiology or Genomics/Genetics or a combination of both or other?

B. Defining and Characterizing Data Sharing

Since the focus of this study is on the sharing of data generated from NIH-funded research, I would like to hear about your experiences with data sharing.

3. Can you please tell me what sharing research data means to you?
    a. Probe: How would you define data sharing?

4. Can you please describe your role and experience with data sharing at your institution?
    a. Probe: Can you describe your experience with access to and submission of [genomic and / or epidemiological] data in NIH data repositories?

5. Can you please describe your understanding of policies and laws that exist around data sharing?
    a. Probe: Can you tell me what others may be saying about NIH data sharing policies that is working or that needs to be changed?

6. In your opinion, when do you think data should be shared?

7. Can you please describe your perception of enforcement process of the data sharing policies by NIH to get researchers to comply?
    a. Probe: What are some concerns you may have with enforcement of the policies?
    b. Probe: What do you perceive as hindering compliance?
    c. Probe: What do you perceive as facilitating compliance?

Next, I will give you a copy of a statement of how NIH describes data sharing and a description of NIH data sharing policies and we will discuss when you are ready.

[THE FOLLOWING STATEMENTS WILL BE GIVEN TO THE RESPONDENTS IN ADVANCE]

NIH describes data sharing to include the submission and access of de-identified data in public or controlled-access databases or data repositories such as the NIH database of Genotypes and Phenotypes (dbGaP) and NHLBI's Biologic Specimen and Data Repositories Information Coordinating Center (BioLINCC) data repository**.**

The 2003 NIH Data Sharing policy states that all investigators with grants of "$500K or more in direct cost in a single year will be expected to address data sharing in their grant application.[42]" This policy states that NIH expects the timely release and sharing of the data to be "**no later than the time of acceptance of publication of the main** findings from the final dataset. [41]"

[42] https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

The 2015 NIH Genomic Data Sharing policy states that all NIH-funded investigators regardless of funding level are encouraged to share broadly large-scale human or non-human genomic data. Data should "be **submitted once it has been cleaned**, i.e. the analytical dataset is finalized. Following data submission, the **data may be accessible** only to the submitting investigators and collaborators for a period **not to exceed six months**.[43]" This policy also states that NIH will **release de-identified** human genomic data submitted to NIH-designated repositories **no later than six months after the initial data submission begins or at the time of acceptance of the first publication**, whichever comes first.

8. Can you please describe your overall perception of the NIH policies and approach to data sharing?

9. What do you think about the timelines?
    a. Probe: What are some challenges with them?

10. Can you give me some examples of the types of data you would say is the most valuable or critical to share?
    a. Probe: What types of data do you think that NIH should have researchers share?

C. Experience with Submitting and Accessing Data in Data Repositories

I would like to ask you a few questions about your [and other researchers'] experiences with data sharing i.e. submitting data and accessing data in public or controlled-access NIH data repositories such as dbGaP, BioLINCC or any other data repositories.

Data Submission

11. In your opinion, can you tell me to what extent researchers such as yourself are sharing or not sharing their [genomic and / or epidemiological] data in public or controlled-access data repositories?
    a. Probe: What do you think are the perceived benefits of sharing research data?
    b. Probe: For the researchers you know, have any of them shared data in public or controlled-access databases?
    c. Probe: For those that are sharing, what was the impetus or incentive for sharing their data?
    d. Probe: For those that are sharing, can you tell me what type of repository they submit data to?

[43] https://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html

e. Probe: What type of data is shared / submitted? Was it genomic or epidemiological data or other?
f. Probe: How easy is it for them to submit their data in the repository?
g. Probe: If they are NOT sharing, what are they saying is preventing them from sharing?

12. Please describe some of the factors that are currently in place that are helping facilitate data sharing.
    a. Probe: What makes data sharing easy or facilitates the sharing of data?
    b. Probe: What are some of the challenges with that?

13. Please describe what you perceive as essential for successful data submission in a data repository.
    a. Probe: What do you think could happen to improve data sharing?
    b. Probe: What are some challenges you foresee with that?
    c. Probe: What aspects of the technological infrastructure (such as data preparation processing, management, expertise, funding support, access) may have helped researchers successfully submit [genomic and / or epidemiological] data in data repositories?

14. Please describe some other ways you have shared their data **outside of a data repository**?
    a. Probe: If direct investigator-to-investigator sharing, have you worked with this individual before or this type of data before?
    b. Probe: What are some challenges with that?

Data Access

15. Can you please explain how you (or other researchers) have accessed data other than your own, from a data repository?
    a. Probe: Can you please tell me what type of repository was accessed?
    b. Probe: What was the type of data accessed? Was it genomic or epidemiological data or other?
    c. Probe: How easy was it to access data from the repository?
    d. Probe: How easy was it to use the data?
    e. Probe: What challenges did you experience?
    f. Probe: What has prevented you from accessing other data?

16. Please describe some of the factors that are currently in place that are facilitating access to shared data?
    a. Probe: What makes data access easy or facilitates the access of data?
    b. Probe: What are some challenges with that?

17. Please describe some of the factors you perceive as essential for successful data access.
    a. Probe: What are some ways to increase access to shared data?
    b. Probe: What do you think could happen to improve data access?
    c. Probe: What are some challenges you foresee with that?

d. Probe: What aspects of the technological infrastructure may have helped you (or other researchers) successfully access [genomic and / or epidemiological] data in data repositories?

18. Please describe criteria you would use for maintaining datasets in data repositories?
    a. Probe: How long should data be maintained in data repositories?
    b. Probe: What timeline would you consider?
    c. Probe: Would you limit the number of requests?
    d. Probe: Would you consider how the data will be used?

D. Organizational / Institutional Factors to Facilitate or Hinder Data Sharing

The next few questions will focus on your perception of **organizational / institutional level factors** that may facilitate or hinder the sharing of research data in public or controlled-access databases.

19. Please tell me what you know about your institution's norms, i.e. culture, beliefs and practices on data sharing?
    a. Probe: Please describe the culture of data sharing at your institution?
    b. Probe: How do your colleagues support data sharing?
    c. Probe: Please describe **existing policies at your institution** that address data sharing? Will you be willing to share the language / statement about data sharing that your institution and IRB use (i.e. regarding how data may be shared and with whom)?
    d. Probe: How is data sharing considered in the **promotion and tenure** process?
    e. Probe: What changes will be the most important to help with the processes?

20. When I say 'institution', in regard to data sharing beliefs and practices, what definition or definitions do you think is relevant?
    a. Probe: Your program, your department or division, your college or school, your university? Does more than one level have to be considered?
    b. Please describe what you know about norms, i.e. culture, beliefs and practices on data sharing that may exist at different organization / institution levels.

21. Please tell me about somethings academic institutions could do to **facilitate sharing** of federal research data?
    a. Probe: What resources and support do you feel researchers may need that academic institutions could provide – e.g. personnel?
    b. Probe: What is your understanding of the use of rewards and incentives to encourage data sharing at your institution? What types of rewards exist?
    c. Probe: What changes will be the most important to help with the processes?

22. Please tell me about somethings NIH could do to **facilitate data sharing** among NIH-funded researchers?
    a. Probe: What resources and support do you not already have that NIH or your Institution may provide?

23. What is NIH doing well that helps facilitate data sharing?

[The next few questions ask about lessons learned from genomic data sharing and your thoughts around those]

24. Based on your experience, please describe what you may have learned from genomic data sharing practices that may be applied to epidemiological data sharing practices?
    a. Probe: How can some of these lessons learned support epidemiological data sharing?
    b. Probe: What things present opportunities for enhancing epidemiological data sharing?
    c. Probe: Are there any aspects of the technological infrastructure (e.g. funding, access, expertise) you've learned that could support epidemiological data sharing?
    d. Probe: What aspects of regulatory policies / laws you've learned could support epidemiological data sharing?
    e. Probe: Please describe what you may have learned around institutional / organizational culture in genomic data sharing that could support epidemiological data sharing?
    f. Probe: Please describe what types of support you may have learned from genomic data sharing that may support / enhance epidemiological data sharing.

25. What are some things from genomic data sharing that might not be applicable to epidemiological data sharing?
    a. Probe: Please describe what modifications are necessary to make them applicable to epidemiological data sharing.

26. What additional suggestions do you have for improving data sharing practices among NIH funded researchers?

27. Please describe any other additional factors not already mentioned that you perceive as **facilitating the sharing** of [genomic and / or epidemiological] research data among NIH funded researchers in data repositories.
    a. Probe: Can you give some examples?

28. Please describe any other additional factors not already mentioned that you perceive as **barriers or challenges** with sharing data that you have observed?
    a. Probe: How were the challenges overcome or resolved?
    b. Probe: Can you give an example of how that has impacted their research / research career?

29. Is there anything else you want to share that I have not asked about?

Closing Remarks

We have reached the end of our discussion today.

Thank you for your time. It has been a pleasure hearing about your experiences and thoughts around data sharing practices and potential opportunities to enhance data sharing in NIH funded research. As part of my analysis and to help with validity of my study, I will be confirming themes from discussions with study participants after the interviews have been completed. Will you be okay with me contacting you again? If you have any questions about this study, you can contact me at my email address – charve7@uic.edu.

For NIH Staff

Introduction

Thank you for taking the time to speak with me today about my research study. My name is Nonye Harvey and I am in the Epidemiology and Genomics Research Program in the Division of Cancer Control and Population Sciences at NCI. I am currently pursuing a Doctor of Public Health degree in Leadership in Public Health at the University of Illinois Chicago School of Public Health.

My thesis is on data sharing in biomedical research and I am interested in hearing about your experiences with the **sharing of de-identified genomic and epidemiological data**, and [organizational / institutional level] **factors you perceive as facilitating or hindering the sharing of these data** in public or controlled-access databases.

There are no right or wrong answers to the interview questions. Your participation is voluntary and please be assured that all responses will be de-identified and kept confidential. Is it okay if I make an audio record of this discussion to complement my notes to ensure that I capture all points made and accurately represent your views?

The recording will be transcribed and the information you give will only be used for this research study.

This interview will consist of 28 questions with some follow-ups for certain questions. This should last about 45 minutes. Before we begin, do you have any questions for me?

A-1. Opening Questions

1. Please describe your current position at NIH.
   a. Probe: How long have you been in your current position?
   b. Probe: How long have you been at NIH?
   c. Probe: What is your official job title? [OR I'd like to quickly confirm your job title is…]

2. Is your work primarily focused on Epidemiology or Genomics/Genetics or a combination of both or other?

B-1. Defining and Characterizing Data Sharing and Data Sharing Policies

Since the focus of this study is on the sharing of data generated from NIH-funded research, I would like to hear about your experiences with data sharing.

3. Can you please tell me what sharing research data means to you?
   a. Probe: How would you define data sharing?

4. Can you please describe your role and experience with data sharing at NIH?
   a. Probe: Can you describe your experience with facilitating access to and submission of [genomic and / or epidemiological] data in NIH data repositories?

5. Can you please describe your understanding of policies and laws that exist around data sharing?
   a. Probe: Can you tell me what others may be saying about NIH data sharing policies that is working or that needs to be changed?

6. In your opinion, when do you think data should be shared?

7. Can you please describe your experiences with **enforcing the data sharing policies**?
   a. Probe: Do you feel that you have the authority or capability for enforcement?
   b. Probe: What makes is easy to enforce data sharing policies?
   c. Probe: What makes it difficult to enforce data sharing policies?

Next, I will give you a copy of a statement of how NIH describes data sharing and a description of NIH data sharing policies and we will discuss when you are ready.

[THE FOLLOWING STATEMENTS WILL BE GIVEN TO THE RESPONDENT]

NIH describes data sharing to include the submission and access of de-identified data in public or controlled-access databases or data repositories such as the NIH database of Genotypes and Phenotypes (dbGaP) and NHLBI's Biologic Specimen and Data Repositories Information Coordinating Center (BioLINCC) data repository**.**

The 2003 NIH Data Sharing policy states that all investigators with grants of "$500K or more in direct cost in a single year will be expected to address data sharing in their grant application.[44]" This policy states that NIH expects the timely release and sharing of the data to be **no later than the time of acceptance of publication of the main** findings from the final dataset.

The 2015 NIH Genomic Data Sharing policy states that all NIH funded investigators regardless of funding level are encouraged to share broadly large-scale human or non-human genomic data. Data should be "**submitted once it has been cleaned**, i.e. the analytical dataset is finalized. Following data submission, the **data may be accessible** only to the submitting investigators and collaborators for a period **not to exceed six months**.[45]" This policy also states that NIH will **release de-identified** human genomic data submitted to NIH-designated repositories "**no later than six months after the initial data submission begins or at the time of acceptance of the first publication**, whichever comes first.[44]"

8. Can you please describe your overall perception of the NIH policies and approach to data sharing?

9. What do you think about the timelines?
    a. Probe: What are some challenges with them?

10. Can you give me some examples of the types of data you would say is the most valuable or critical to share?
    a. Probe: What types of data do you think that NIH should have researchers share?

C-1. Experience with Investigators Submitting and Accessing Data in Data Repositories

I would like to ask you a few questions about your knowledge and experience with researchers submitting data and accessing data in public or controlled-access NIH data repositories such as dbGaP, BioLINCC or any other data repositories.

<u>Data Submission</u>

11. In your opinion, can you tell me to what extent researchers are sharing or not sharing their [genomic and / or epidemiological] data in public or controlled-access data repositories?
    a. Probe: What do you think are the perceived benefits of sharing research data?
    b. Probe: For the researchers you know, have any of them shared data in public or controlled-access databases?
    c. Probe: For those that are sharing, what was the impetus or incentive for sharing this data?

[44] https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html
[45] https://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html

d. Probe: For those that are sharing, can you tell me what type of repository they submit data to?
e. Probe: What type of data is shared / submitted? Was it genomic or epidemiological data or other?
f. Probe: How easy is it for them to submit their data in the repository?
g. Probe: If they are NOT sharing, what are they saying is preventing them from sharing?

12. Please describe some of the factors that are currently in place that are helping facilitate data sharing.
    a. Probe: What makes data sharing easy or facilitates the sharing of data?
    b. Probe: What are some of the challenges with that?

13. Please describe what you perceive as essential for successful data submission in a data repository.
    a. Probe: What do you think could happen to improve data sharing?
    b. Probe: What are some challenges you foresee with that?
    c. Probe: What aspects of the technological infrastructure (such as data preparation processing, management, expertise, funding support, access) may have helped researchers successfully submit [genomic and / or epidemiological] data in data repositories?

14. Please describe some other ways researchers have shared their data **outside of a data repository**?
    a. Probe: If direct investigator-to-investigator sharing, have you worked with this individual before or this type of data before?
    b. Probe: What are some challenges with that?

Data Access

15. Can you please explain your understanding of how researchers access data other their own from a data repository?
    a. Probe: Can you please tell me what type of repository they have accessed?
    b. Probe: What was the type of data accessed? Was it genomic or epidemiological data or other?
    c. Probe: How easy was it for them to access data from the repository?
    d. Probe: How easy was it to use the data?
    e. Probe: What challenges did they experience?
    f. Probe: What are researchers saying has prevented them from accessing other data?

16. Please describe some of the factors that are currently in place that facilitate access to shared data?
    a. Probe: What makes data access easy or facilitates the access of data?
    b. Probe: What are some challenges with that?

17. Please describe some of the factors you perceive as essential for successful data access.
    a. Probe: What are some ways to increase access to shared data?
    b. Probe: What do you think could happen to improve data sharing?
    c. Probe: What are some challenges you foresee with that?
    d. Probe: What aspects of the technological infrastructure may have helped researchers successfully access [genomic and / or epidemiological] data in data repositories?

18. lease describe criteria you would use for maintaining datasets in data repositories?
    a. Probe: How long should data be maintained in data repositories?
    b. Probe: What timeline would you consider?
    c. Probe: Would you limit the number of requests?
    d. Probe: Would you consider how the data will be used?

D-1. Organizational / Institutional Factors that Facilitate or Hinder Data Sharing

The next few questions will focus on your perception of **organizational / institutional level factors** that facilitate or hinder the sharing of research data in public or controlled-access databases.

19. Please tell me about somethings academic institutions could do to **facilitate sharing** of federal research data?
    a. Probe: What resources and support do you feel researchers may need that academic institutions could provide – e.g. personnel?
    b. Probe: How is data sharing considered in the University's promotion and tenure process?
    c. Probe: What is your understanding of the use of rewards and incentives to encourage data sharing among researchers? What types of rewards exist?
    d. Probe: What changes will be the most important to help with the processes?

20. Please tell me about somethings NIH could do to **facilitate data sharing** among NIH-funded researchers?
    a. Probe: What resources and support do you not already have as NIH staff that NIH may provide?

21. What is NIH doing well that helps facilitate data sharing?

22. Can you give me an example of how you have implemented data sharing policies at NIH?
    a. Probe: What were some challenges with doing this?

b. Probe: How were they handled?
c. Probe: What changes will be most important to help us enhance the process?

[The next few questions ask about lessons learned from genomic data sharing and your thoughts around those]

23. Based on your experience, please describe what you may have learned from genomic data sharing practices that may be applied to epidemiological data sharing practices?
    a. Probe: How can some of these lessons learned support epidemiological data sharing?
    b. Probe: What things present the best opportunities for enhancing epidemiological data sharing?
    c. Probe: Are there any aspects of the technological infrastructure (e.g. funding, access, expertise) you've learned that could assist epidemiological data sharing?
    d. Probe: What aspects of regulatory policies / laws you've learned could support epidemiological data sharing?
    e. Probe: Please describe what you may have learned around institutional / organizational culture in genomic data sharing that could support epidemiological data sharing?
    f. Probe: Please describe what types of support you may have learned from genomic data sharing that may support / enhance epidemiological data sharing.

24. What are some things from genomic data sharing that might not be applicable to epidemiological data sharing?
    a. Probe: Please describe what modifications are necessary to make them applicable to epidemiological data sharing.

25. What additional suggestions do you have for improving data sharing practices among NIH funded researchers?

26. Please describe any other additional factors not already mentioned that you perceive as **facilitating the sharing** of [genomic and / or epidemiological] research data among NIH funded researchers in data repositories.
    a. Probe: Can you give some examples?

27. Please describe any other additional factors not already mentioned that you perceive as **barriers or challenges** with sharing data that you have observed?
    a. Probe: How were the challenges overcome or resolved?
    b. Probe: Can you give an example of how that has impacted their research / research career?

28. Is there anything else you want to share that I have not asked about?

Closing Remarks

We have reached the end of our discussion today.

Thank you for your time. It has been a pleasure hearing about your experiences and thoughts around data sharing practices and potential opportunities to enhance data sharing in NIH funded research. As part of my analysis and to help with validity of my study, I will be confirming themes from discussions with study participants after the interviews have been completed. Will you be okay with me contacting you again? If you have any questions about this study, you can contact me at my email address – charve7@uic.edu.

Appendix E: Codebook

Constructs and Factors	Codes	Definition of Codes	A priori or Emergent
Facilitator	Fac	Factor already existing that is working and should be continued	A priori, Emergent
Barrier	Bar	Factor currently in the way that may likely impact ability to accomplish goal	A priori, Emergent
Opportunity	Opp	New factor or idea that must be put in place or addressed to achieve goal of improving data sharing	A priori, Emergent
CONSTRUCT: Culture and Practices			
Culture and Practices	Culture	Culture, practices and norms of institutions and researchers around data sharing	A priori
Intrinsic incentives	Intrinsic incentive-Fac	Reference to internal motivation to share data for advancement of science or personal benefit	A priori
Career concerns	Career concerns-Bar	Reference to concerns or perceived threats to researchers' careers related to data misuse, misinterpretation, scooping and negative criticism	A priori
Lack of a reward system	Reward system-Bar	Reference to the lack of a system at institutions to reward or credit researchers for sharing data, including references to promotion and	A priori
Reward structure changes needed	Reward structure changes-Opp	Reference to opportunities to change reward structure at institutions to reward researchers for data sharing	A priori
Culture differences in research fields	Research fields culture-Fac or Bar	Reference to differences in culture of different fields (genomic vs epidemiology), which may facilitate or hinder data sharing	A priori

Culture shift in research fields	Research fields culture shift-Opp	Reference to a shift in culture, thinking and perception around data sharing in different fields	A priori
CONSTRUCT: Regulatory Policy and Law			
Regulatory Policy and Law	Policy	Reference to data sharing policies, laws governing human subject research and participant privacy	A priori
Clarity of policy	Clarity of policy-Fac	Reference to how clear data sharing policies are related to timelines, guidelines and expectations	A priori
Lack of clarity of policy	Clarity of policy-Bar	Reference to vagueness or ambiguity in data sharing policies related to timelines, guidelines and expectations	A priori
Clarity of policy needed	Clarity of policy needed-Opp	Reference to ideas for improving the clarity of data sharing policies	A priori
Enforcement of policy	Enforcement-Fac	Reference to enforcement of data sharing policies by NIH staff	A priori
Inconsistent enforcement of policy	Enforcement-Bar	Reference to the inconsistent enforcement or implementation of data sharing policies by NIH staff across NIH institutes and centers	A priori
Change needed in enforcement	Change needed in enforcement - Opp	Reference to opportunities to make changes to existing strategies and mechanisms for enforcing data sharing policies	A priori
Privacy concerns	Privacy concerns-Bar	Reference to concerns about participant data confidentiality, consent and data sharing in research studies	A priori
Addressing privacy concerns	Addressing privacy concerns-Opp	Reference to opportunities for alleviating privacy concerns related to sharing of human research data	A priori

Definition of data sharing	Definition of data sharing-BarE	Reference to differences in meaning or respondent's definition of data sharing	Emergent
CONSTRUCT: Resources			
Resources	Resources	Reference to resources to support data sharing activities by investigators or NIH staff	A priori
Administrative / Technical resources	Admin/Tech resources-Fac or Bar	Reference to adequate or insufficient administrative support related to staff time and effort, and technical support for data sharing processes	A priori
Administrative / technical resources needed	Admin/tech resources needed-Opp	Reference to opportunities for improving administrative/technical support for researchers and staff for data sharing efforts	A priori
Financial resources-Fac/Bar	Financial resources-Fac or Bar	Reference to the cost, funding or financial resources required to support data sharing as adequate or insufficient related to personnel, technological support	A priori
Financial resources needed	Financial resources needed-Opp	Reference to opportunities for improving financial resources / funding to enhance data sharing	A priori
Leadership support	Leadership support- Fac or Bar	Reference to support (or lack of support) rom NIH and institutions leadership through the provision of resources, priority and oversight to enhance data sharing	A priori
Leadership support needed	Leadership support needed-Opp	Reference to opportunities for increased leadership support and priority for data sharing	A priori
Training	Training-Fac or Bar	Reference to training (or insufficient training) available for researchers and staff to enhance knowledge, skills for effective data submission, access and management	A priori

Training needed	Training needed-Opp	Reference to opportunities for training to enhance understanding of data sharing processes and resources	A priori
CONSTRUCT: Technological Infrastructure			
Technological Infrastructure	TechInfrastr.	Reference to technology, support related to data management and preparation, and repository capabilities	A priori
Analytic data complexity	Analytic data complexity-Fac or Bar	Reference to analytic data format, standardization, and metadata documentation / annotation of datasets, as facilitators or barriers to data sharing.	A priori
Addressing analytic data complexity	Addressing analytic data complexity-Opp	Reference to opportunities to simplify process and requirements for analytic data / metadata expected to be submitted in a data repository	A priori
Clarity of submission/access process	Clarity of submission / access process-Fac or Bar	Reference to clarity (or lack of clarity) of technical / administrative processes related to submission and access of data in a data repository	A priori
Clarity of submission/access process needed	Clarity of submission/ access process-Opp	Reference to opportunities to improve clarity in the submission and access process of data in a data repository	A priori
Expertise	Expertise-Fac or Bar	Reference to NIH staff and researchers' expertise (or lack of expertise) with and knowledge of processes and available resources for data submission and access data	A priori
Expertise needed	Expertise needed-Opp	Reference to opportunities to improve expertise needed for successful data sharing processes	A priori
Repository capabilities	Repository capabilities-Fac or Bar	Reference to adequate capacity, effectiveness, efficiency of repository or sub-optimal capability of repository	A priori

Addressing repository capabilities	Addressing repository capabilities-Opp	Reference to opportunities to improve the effectiveness, adequacy and efficiency of data repository	A priori
OTHER – Emergent factor			
Addressing data use, cost-benefit and value	Addressing data use, cost-benefit and value-Opp	Reference to opportunities for NIH to conduct cost/risk-benefit analysis to understand use or value of data shared via data repository, and participant's understanding of consent for sharing data	Emergent

Appendix F: Co-Occurrence Tables

TABLE I: INTRINSIC INCENTIVE-FAC AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Intrinsic incentive-Facilitator	
Enforcement of policy-Fac	3
Financial resources-Fac	3
Lack of a reward system-Bar	2

TABLE II**:** CULTURE DIFFERENCES IN RESEARCH FIELDS-FAC AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Culture differences in research fields-Facilitator	
Enforcement of policy-Fac	1
Reward structure changes needed-Opp	1
Definition of data sharing-BarE	1

TABLE III: CLARITY OF POLICY-FAC AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Clarity of policy-Facilitator	
Lack of clarity of policy-Bar	15
Clarity of submission/access process-Fac	10
Lack of expertise-Bar	1
Expertise-Fac	6
Training-Fac	11

TABLE IV**:** ENFORCEMENT OF POLICY-FAC AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Enforcement of policy-Facilitator	
Leadership-Fac	7
Inadequate Admin/Tech resources-Bar	2
Inadequate Financial resources-Bar	3
Change needed in enforcement approach-Opp	9
Lack of clarity of policy-Bar	9
Clarity of policy-Fac	3
Inconsistent enforcement of policy-Bar	13
Intrinsic incentive-Fac	3
Repository capabilities-Fac	3

TABLE V: ADMINISTRATIVE/TECHNICAL RESOURCES-FAC AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Administrative/technical resources-Fac	
Financial resources-Fac	17
Clarity of submission / access process-Fac	17
Repository capabilities-Fac	31
Expertise-Fac	10
Admin/Tech resources needed-Opp	5
Admin/Tech resources-Bar	9
Leadership support-Fac	5
Training-needed-Opp	3
Training-Fac	2
Analytic data complexity-Fac	3
Clarity of submission/access process needed-Opp	5
Expertise needed-Opp	4
Lack of expertise-Bar	4
Addressing Repository capabilities-Opp	6
Sub-optimal Repository capabilities-Bar	4

TABLE VI**:** FINANCIAL RESOURCES-FAC AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Financial resources-Facilitator	
Admin /Tech resources-Fac	17
Financial resources needed-Opp	4
Inadequate Financial resources-Bar	3
Leadership support-Fac	12
Clarity of submission / access process-Fac	7
Repository capabilities-Fac	9

TABLE VII: LEADERSHIP SUPPORT-FAC AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Leadership support-Facilitator	
Admin /Tech resources-Fac	5
Enforcement of policy-Fac	7
Financial resources-Fac	12
Clarity of submission / access process-Fac	4
Repository capabilities-Fac	5

TABLE VIII: TRAINING-FAC AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Training-Facilitator	
Clarity of policy-Fac	11
Clarity of submission / access process-Fac	19
Expertise-Fac	13
Expertise needed-Opp	3
Leadership support-Fac	2

TABLE IX**:** ANALYTIC DATA COMPLEXITY-FAC AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Analytic data complexity-Facilitator	
Clarity of submission/access process-Fac	5
Expertise-Fac	4
Career concerns-Bar	2
Leadership support-Fac	1
Training needed-Opp	2
Addressing Analytic data complexity-Opp	3
Repository capabilities-Fac	2

TABLE X**:** CLARITY OF SUBMISSION/ACCESS PROCESS-FAC AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Clarity of submission/access process-Facilitator	
Expertise-Fac	30
Training-Fac	19
Repository capabilities-Fac	16
Clarity of policy-Fac	10
Admin/tech resources-Fac	17
Financial resources-Fac	7
Leadership support-Fac	4
Training needed-Opp	6
Lack of clarity of submission/access process-Bar	9
Lack of expertise-Bar	8

TABLE XI**:** EXPERTISE-FAC AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Expertise-Facilitator	
Clarity of submission/access process-Fac	30
Training-Fac	13
Admin/tech resources-Fac	10
Clarity of policy-Fac	6
Training needed-Opp	3
Expertise needed-Opp	3

TABLE XII**:** REPOSITORY CAPABILITIES-FAC AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Repository capabilities-Facilitator	
Admin/tech resources-Fac	31
Clarity of submission/access process-Fac	16
Financial resources-Fac	9
Leadership support-Fac	5
Training-Fac	3
Analytic data complexity-Bar	3
Analytic data complexity-Fac	2
Enforcement of policy-Fac	3
Expertise-Fac	3
Addressing repository capabilities-Opp	4

TABLE XIII: CAREER CONCERNS-BAR AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Career concerns-Barrier	
Culture differences in research fields-Bar	3
Lack of reward system-Bar	3
Lack of clarity of policy-Bar	5
Analytic data complexity-Bar	6
Clarity of submission/access process needed-Opp	3
Lack of expertise-Bar	3

TABLE XIV: CULTURE DIFFERENCES IN RESEARCH FIELDS-BAR AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Culture differences in research fields-Barrier

Career concerns-Bar	3
Culture shift in research fields-Opp	1
Lack of clarity of policy-Bar	2
Inconsistent enforcement of policy-Bar	1
Enforcement of policy-Fac	1
Inadequate Admin/tech resources-Bar	1
Inadequate Financial resources-Bar	2
Analytic data complexity-Bar	2
Clarity of submission/access process-Fac	1
Lack of expertise-Bar	1

TABLE XV: LACK OF REWARD SYSTEM-BAR AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Lack of reward system-Barrier	
Leadership support-Bar	6
Reward structure changes needed-Opp	4
Career Concerns-Bar	3
Inconsistent enforcement of policy-Bar	3
Intrinsic incentive-Fac	2
Culture shift in research fields-Opp	2
Inadequate Admin/Tech resources-Bar	2

TABLE XVI: LACK OF CLARITY OF POLICY-BAR AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Lack of clarity of policy-Barrier	
Career concerns-Bar	5
Clarity of policy needed-Opp	9
Clarity of policy-Fac	15
Definition of data sharing-BarE	2
Change needed in enforcement approach-Opp	3
Inconsistent enforcement of policy-Bar	30
Enforcement of policy-Fac	9
Privacy concerns-Bar	7
Inadequate Admin/tech resources-Bar	8
Inadequate Financial resources-Bar	3

TABLE XVII**:** INCONSISTENT ENFORCEMENT OF POLICY-BAR AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Inconsistent enforcement of policy-Barrier	
Lack of reward system-Bar	3
Enforcement of policy-Fac	13
Privacy concerns-Bar	2
Inadequate Admin/tech resources-Bar	3
Inadequate Financial resources-Bar	2
Leadership support-Bar	4
Lack of expertise-Bar	3
Sub-optimal repository capabilities-Bar	3
Clarity of policy-Fac	5

TABLE XVIII**:** PRIVACY CONCERNS-BAR AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Privacy concerns-Barrier	
Clarity of policy needed-Opp	3
Clarity of policy-Fac	7
Analytic data complexity-Bar	4
Lack of clarity of submission / access process-Bar	5

TABLE XIX**:** DATA SHARING DEFINITION-BAR (EMERGENT) AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Definition of data sharing-Barrier (Emergent)	
Career concerns-Bar	6
Lack of clarity of policy-Bar	2
Leadership support-Bar	6
Leadership support-Fac	5
Analytic data complexity-Bar	4

TABLE XX: INADEQUATE ADMINISTRATIVE/TECHNICAL RESOURCES-BAR AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Inadequate Administrative/Technical Resources-Barrier	
Lack of clarity of policy-Bar	8
Inadequate Financial resources-Bar	31
Analytic data complexity-Bar	14
Lack of clarity of submission / access process-Bar	17
Lack of expertise-Bar	16

Addressing repository capabilities-Opp	6
Sub-optimal repository capabilities-Bar	13

TABLE XXI**:** INADEQUATE FINANCIAL RESOURCES-BAR AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Inadequate Financial resources-Barrier	
Inadequate Admin/Tech resources-Bar	31
Admin/tech resources-Fac	3
Financial resources needed-Opp	3
Clarity of submission/access process-Fac	2
Sub-optimal repository capabilities-Bar	5

TABLE XXII**:** LEADERSHIP SUPPORT-BAR AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Leadership support-Barrier	
Lack of reward system-Bar	6
Definition of data sharing-BarE	6
Inconsistent enforcement of policy-Bar	4
Inadequate Admin/Tech resources-Bar	2
Leadership support needed-Opp	2
Analytic data complexity-Bar	3

TABLE XXIII**:** TRAINING-BAR AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Training-Barrier	
Clarity of policy-Fac	2
Training-fac	5
Lack of clarity of submission / access process-Bar	2
Clarity of submission/access process-Fac	5
Lack of expertise-Bar	3
Expertise-Fac	5

TABLE XXIV**:** ANALYTIC DATA COMPLEXITY-BAR AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Analytic data complexity-Barrier	
Career concerns-Bar	2

Leadership support-Fac	1
Training needed-Opp	2
Addressing Analytic data complexity-Opp	3
Clarity of submission/access process-Fac	5
Expertise-Fac	4
Repository capabilities-Fac	2

TABLE XXV: LACK OF CLARITY OF SUBMISSION/ACCESS PROCESS-BAR AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Lack of clarity of submission/access process-Barrier	
Lack of clarity of policy-Bar	3
Privacy concerns-Bar	5
Inadequate Admin/Tech resources-Bar	17
Analytic data complexity-Bar	11
Clarity of submission/access process-Fac	9
Lack of expertise-Bar	29
Sub-optimal repository capabilities-Bar	12

TABLE XXVI: LACK OF EXPERTISE-BAR AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Lack of expertise-Barrier	
Inadequate Admin/Tech resources-Bar	16
Analytic data complexity-Bar	12
Career concerns-Bar	3
Inconsistent enforcement of policy-Bar	3
Lack of clarity of submission / access process-Bar	29
Sub-optimal repository capabilities-Bar	9
Expertise-Fac	7
Addressing repository capabilities-Opp	5

TABLE XXVII: SUB-OPTIMAL REPOSITORY CAPABILITIES-BAR AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Sub-optimal Repository Capabilities-Barrier	
Inadequate Admin/Tech resources-Bar	13
Analytic data complexity-Bar	10
Lack of clarity of submission / access process-Bar	12
Lack of expertise-Bar	9
Addressing repository capabilities-Opp	5

Inadequate Financial resources-Bar	5

TABLE XXVIII: CULTURE SHIFT IN RESEARCH FIELDS-OPP - SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Culture shift in research fields-Opportunity	
Career concerns-Bar	1
Reward structure changes needed-Opp	4
Lack of a reward system-Bar	2
Clarity of policy needed-Opp	2
Changes in enforcement approach needed-Opp	2
Enforcement of policy-Fac	2
Training needed-Opp	2
Analytic data complexity-Fac	1

TABLE XXIX: REWARD STRUCTURE CHANGES NEEDED-OPP- SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Reward structure changes needed-Opportunity	
Culture shift in research fields-Opp	4
Financial resources needed-Opp	4
Leadership support-Fac	2
Leadership support needed-Opp	3

TABLE XXX: CLARITY OF POLICY NEEDED-OPP AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Clarity of policy needed-Opportunity	
Lack of clarity of policy-Bar	9
Clarity of policy-Fac	12
Change needed in enforcement approach-Opp	7
Addressing privacy concerns-Opp	7
Privacy concerns-Bar	3
Training needed-Opp	3
Clarity of submission/access process needed-Opp	3

TABLE XXXI: CHANGE NEEDED IN ENFORCEMENT APPROACH- OPP- SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Change needed in enforcement approach-Opportunity	
Clarity of policy needed-Opp	7

Enforcement of policy-Fac	9
Admin/tech resources needed-Opp	4
Leadership support-Fac	2
Training needed-Opp	2
Addressing repository capabilities-Opp	3

TABLE XXXII: ADDRESSING PRIVACY CONCERNS-OPP AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Addressing privacy concerns-Opportunity	
Clarity of policy needed-Opp	7
Privacy concerns-Bar	1
Clarity of submission/access process needed-Opp	2

TABLE XXXIII: ADMINISTRATIVE/TECHNICAL RESOURCES NEEDED-OPP AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Administrative/Technical Resources Needed-Opportunity	
Inadequate Admin/Tech resources-Bar	4
Change needed in enforcement approach-Opp	4
Financial resources needed-Opp	5
Leadership support needed-Opp	4
Training needed-Opp	5
Clarity of submission/access process needed-Opp	5
Clarity of submission/access process-Fac	1
Expertise needed-Opp	4
Addressing repository capabilities-Opp	9
Repository capabilities-Fac	4

TABLE XXXIV: FINANCIAL RESOURCES NEEDED-OPP AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Financial Resources Needed-Opportunity	
Reward structure changes needed-Opp	4
Admin/tech resources needed-Opp	5
Inadequate Financial resources-Bar	3
Financial-Fac	4
Leadership support-Fac	1
Leadership support needed-Opp	5
Training needed-Opp	5

Analytic data complexity-Bar	4

TABLE XXXV: LEADERSHIP SUPPORT NEEDED-OPP AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Leadership Support Needed-Opportunity	
Reward structure changes needed-Opp	3
Changes in enforcement approach needed-Opp	5
Admin/tech resources needed-Opp	4
Admin/tech resources-Fac	1
Financial resources needed-Opp	5
Financial resources-Fac	1
Leadership support-Bar	2
Leadership support-Fac	2

TABLE XXXVI: TRAINING NEEDED-OPP- SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Training Needed-Opportunity	
Clarity of submission/access process needed-Opp	15
Culture shift in research fields-Opp	2
Financial resources needed-Opp	5
Clarity of submission/access process-Fac	6
Expertise needed-Opp	6
Admin/tech resources needed-Opp	5

TABLE XXXVII: ADDRESSING ANALYTIC DATA COMPLEXITY-OPP- SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Addressing Analytical Data Complexity-Opportunity	
Admin/tech resources needed-Opp	1
Financial resources needed-Opp	1
Training needed-Opp	1
Analytic data complexity-Fac	3
Clarity of submission/access process needed-Opp	3
Addressing repository capabilities-Opp	2

TABLE XXXVIII: CLARITY NEEDED FOR SUBMISSION / ACCESS PROCESS-OPP AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Clarity Needed for Submission / Access Process-Opportunity

Training needed-Opp	15
Expertise needed-Opp	8
Addressing repository capabilities-Opp	4
Admin/tech resources needed-Opp	5
Inadequate Admin/Tech resources-Bar	4

TABLE XXXIX: EXPERTISE NEEDED-OPP AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Expertise needed-Opportunity	
Training needed-Opp	6
Training-Fac	3
Clarity of submission/access process needed-Opp	8
Clarity of submission/access process-Fac	3
Lack of expertise-Bar	2
Expertise-Fac	3
Repository capabilities-Fac	2

TABLE XL: ADDRESSING REPOSITORY CAPABILITIES-OPP AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Addressing Repository Capabilities-Opportunity	
Admin/tech resources needed-Opp	9
Inadequate Admin/Tech resources-Bar	6
Admin/tech resources-Fac	6
Clarity of submission/access process needed-Opp	4
Lack of expertise-Bar	5

TABLE XLI: ADDRESSING DATA USE, COST-BENEFIT AND VALUE – OPP AND SELECT CO-OCCURRING FACTORS WITH CODING REFERENCE COUNTS

Addressing Data Use, Cost-Benefit and Value-Opportunity	
Intrinsic incentive-Fac	1
Training needed-Opp	3
Inadequate Financial resources-Bar	1

Appendix G: Comparison of Findings Between Interviews and Document Reviews

By Facilitators, Barriers and Opportunities

Facilitators	Factors	*In-depth Interviews (PIs and staff)*	*RFI - dbGaP Processes*	*RFI - Data Sharing Strategies*	*EGRP Report #1 Jan. 2016*	*EGRP Report #2 Jul. 2016*	*EGRP Report #3 Apr. 2017*
Culture	Intrinsic incentives - Facilitator	X					
	Culture differences in research fields – Facilitator	X					
Policy	Clarity of policy – Facilitator	X					
	Enforcement of policy – Facilitator	X		X			
Resources	Admin/Tech resources – Facilitator	X	X		X		
	Financial resources – Facilitator	X		X			
	Leadership support – Facilitator	X		X			
	Training – Facilitator	X					
Technological Infrastructure	Analytic data complexity – Facilitator	X		X			
	Clarity of Submission/Access process – Facilitator	X					
	Expertise – Facilitator	X					
	Repository capabilities – Facilitator	X					

Barriers	Factors	*In-depth Interviews (PIs and staff)*	*RFI - dbGaP Processes*	*RFI - Data Sharing Strategies*	*EGRP Report #1 Jan. 2016*	*EGRP Report #2 Jul. 2016*	*EGRP Report #3 Apr. 2017*
Culture	Career concerns – Barrier	X		X		X	X
	Culture differences in research fields – Barrier	X					
	Lack of a reward system - Barrier	X					
Policy	Clarity of policy – Barrier	X					
	Inconsistent enforcement of policy –Barrier	X		X			
	Privacy concerns - Barrier	X		X		X	X
	Definition of data sharing– Barrier (Emergent)	X				X	
Resources	Admin/ tech resources –Barrier	X	X	X	X	X	
	Inadequate Financial resources –Barrier	X		X	X	X	
	Leadership support –Barrier	X					
	Training –Barrier	X					
Technological Infrastructure	Analytic data complexity –Barrier	X	X	X	X	X	
	Clarity of Submission/Access process –Barrier	X	X			X	
	Lack of expertise – Barrier	X		X		X	
	Repository capabilities – Barrier	X	X			X	X

Opportunities	Factors	*In-depth Interviews (PIs and staff)*	*RFI - dbGaP Processes*	*RFI - Data Sharing Strategies*	*EGRP Report #1 Jan. 2016*	*EGRP Report #2 Jul. 2016*	*EGRP Report #3 Apr. 2017*
Culture	Culture shift in research fields	X					
	Reward structure changes needed	X		X			
Policy	Addressing lack of clarity in policy	X					
	Change needed in enforcement	X					
	Addressing privacy concerns	X	X	X			
Resources	Addressing admin/ tech needs	X		X			
	Addressing financial needs	X		X			
	Leadership support needed	X					
	Training needed	X	X	X			
Technological Infrastructure	Addressing Analytic data complexity	X	X	X			X
	Addressing lack of Clarity in submission/access process	X	X	X			X
	Expertise needed	X					
	Addressing repository capabilities	X	X	X			X
Other	Addressing data use, cost and value (Emergent)	X					

Comparison of Findings Between Interviews and Document Reviews

By Construct

Factors	**In-depth Interviews (investigation and staff)**	**RFI Comments on dbGaP Processes**	**RFI Comments on Data Sharing Strategies**	**EGRP Report #1 Jan. 2016**	**EGRP Report #2 Jul. 2016**	**EGRP Report #3 April 2017**
CULTURE						
Intrinsic incentives - Facilitator	X					
Career concerns – Barrier	X		X		X	X
Lack of a reward system - Barrier	X					
Reward structure changes needed	X		X			
Culture differences in research fields – Facilitator, Barrier	X					
Culture shift in research fields	X					
POLICY						
Clarity of policy – Facilitator, Barrier	X					
Addressing lack of clarity in policy	X					
Definition of data sharing differences – Barrier (Emergent)	X				X	
Enforcement of policy – Facilitator, Barrier	X		X			
Change needed in enforcement	X					
Privacy concerns - Barrier	X		X		X	X
Addressing privacy concerns	X	X	X			
RESOURCES						
Administrative and technical resources – Facilitator, Barrier	X	X	X – Barrier	X	X – Barrier	
Addressing	X		X			

administrative and technical needs						
Financial resources – Facilitator, Barrier	X		X	X - Barrier	X – Barrier	
Addressing financial needs	X		X			
Leadership support – Facilitator, Barrier	X		X – Facilitator			
Leadership support needed	X					
Training – Facilitator, Barrier	X					
Training needed	X	X	X			
TECHNOLOGICAL INFRASTRUCTURE						
Analytic data complexity – Facilitator, Barrier	X	X - Barrier	X	X – Barrier	X – Barrier	
Addressing Analytic data complexity	X	X	X			X
Clarity of Submission/Access process – Facilitator, Barrier	X	X - Barrier			X – Barrier	
Addressing lack of Clarity in submission/access process	X	X	X			X
Expertise – Facilitator, Barrier	X		X – Barrier		X – Barrier	
Expertise needed	X					
Repository capabilities – Facilitator, Barrier	X	X - Barrier			X – Barrier	X - Barrier
Addressing repository capabilities	X	X	X			X
Addressing data use, cost and value (Emergent)	X					

Appendix H: Recommendations

Changing the Reward Structure at Institutions - CULTURE	
1. Academic institutions to change their reward structure and include data sharing as part of the promotion and tenure criteria. This way researchers who share data get the credit and recognition for sharing as part of their academic career evaluation	Academic institutions
2. Integrate use of a new metric for data sharing, the S-Index, in the evaluation of investigators careers, analogous to the H-index for publications	Academic institutions and NIH
3. NIH to reward investigators who have a track record for sharing	NIH
4. NIH to set aside a percentage of funding in grant awards specifically for data sharing	NIH
5. NIH and academic institutions to consider collaborations as part of "broad" data sharing for large collaborative or consortia research studies, and recognize middle authorship from collaborative studies as equally important in the evaluation of academic research careers and funding opportunities	Academic institutions and NIH
Improving the Clarity of Existing Data Sharing Policies - POLICY	
6. NIH to clarify in the policy the definition of data sharing so there's a shared understanding of what it means o Requires engagement with NIH and non-NIH leaders	NIH
7. NIH to reassess the policy timeline and expectations based on study an data types, with the understanding that there is not a one-size fits all approach to data sharing in scientific research	NIH
8. NIH to identify more effective strategies beyond the RFI mechanism for soliciting feedback from the community on the development of policy. Considerations: o increased early and frequent engagement and communication with targeted key stakeholders for clearer policies and processes, and training o conduct focus groups with targeted stakeholders to explore areas to improve clarity in policy	NIH
Improving Strategies and Clarity in Enforcement of Data Sharing Policies - POLICY	
9. Academic institutions to act as enforcers of policy, alongside NIH staff	Academic institutions and NIH

Considerations: o NCI director to meet with cancer center directors and propose joint enforcement of NIH data sharing policy among their investigators	
10. NIH to make the processes for implementation clear, consistent, uniform and transparent across all institutes and centers at NIH o conduct focus groups with targeted stakeholders to explore areas to improve clarity in implementation processes	NIH
Addressing Concerns with Privacy, Patient Consent and Data Sharing - POLICY	
11. NIH to conduct an evidence-based research to assess participants' understanding of consent and data sharing	
Prioritizing the Investment and Communication of Resources – RESOURCES	
12. Leadership at academic institutions should increase support and investment in resources to support data sharing activities conducted by their researchers Considerations: o Institutions to provide central support system with dedicated staff to support data sharing activities at the institutions	Academic institutions
13. NIH to invest more resources in supporting epidemiology data sharing, similar to genomic data sharing	NIH
14. NIH to increase awareness and communication about provisions in the existing policy that allows researchers to include data sharing costs in their grant budget Considerations: o Making it more explicit and easy to find the statements in the policies may require some revision to the way the policies are written	NIH
15. NIH to provide training for investigators and staff to increase understanding of policy requirements, implementation processes and systems, skills and expertise needed to effectively support data sharing Considerations: o Institutions to incorporate data sharing as part of the curriculum at institutions e.g. training on reproducibility and metadata standards o NIH to hold workshops on data sharing for investigators, perhaps focusing on early career investigators who may not be familiar or experienced with data sharing o NIH to ensure broad and bi-directional communication	Academic institutions and NIH

and training on data sharing processes, systems and tool e.g. dbGaP, targeting novice users o NIH to develop training materials, tools on new and existing resources, and materials to support data sharing activities by investigators and NIH staff.	
Increasing Investment in Technological Infrastructure to Support Data Sharing – TECHNOLOGICAL INFRASTRUCTURE	
16. NIH to increase investment in engineering and technical support of dbGaP Considerations: o Automating, standardizing and streamlining the data sharing process o Improving functionality and interoperability of dbGaP; changing the way the data is submitted	NIH
Improving the clarity of data submission / access process – TECHNOLOGICAL INFRASTRUCTURE	
17. NIH to increase the clarity in the data submission / access process, ensuring that the various steps and processes are clear and transparent Considerations: o Training by NCBI / NIH staff on the processes, including education of what resources are available to support data sharing at NIH o Conduct focus groups with targeted stakeholders to explore areas to improve clarity in data submission/access process o Develop a streamlined process for data transfer agreements for large collaborative research projects	NIH
Increasing Understanding on the Use of Data, the Cost and Value of Sharing Data via Data Repositories - OVERARCHING ISSUE	
18. NIH to do a cost-benefit analysis of data sharing in data repositories	NIH
19. NIH to increase awareness of research publications that have already analyzed the use of data in data repositories (e.g. dbGaP and BioLINCC)	NIH

VITA

CHINONYE E. HARVEY, M.P.H., Dr.P.H. candidate
13110 Brewers Tavern Terrace, Clarksburg, MD 20871
Phone: (703) 508-2297
Email: harveyn@mail.nih.gov

EDUCATION

Dr.P.H. (candidate), Leadership in Public Health. Expected May 2019.
University of Illinois Chicago, Chicago, IL

M.P.H., Prevention and Community Health - Maternal and Child Health and International Health, May 2001
George Washington University School of Public Health and Health Services, Washington, DC,

B.S., Medical Technology, May 1998
Michigan State University, College of Natural Sciences, East Lansing, MI,

M.T. (ASCP), Certified Medical Technologist, Detroit Medical Center-University Laboratories, Detroit, MI, June 1999

PROFESSIONAL EXPERIENCE

National Cancer Institute (NCI)
Division of Cancer Control and Population Sciences (DCCPS)
Epidemiology and Genomics Research Program (EGRP)
Public Health Advisor 07/2008 – Present

- Planning and budget lead with delegated authority to guide and make budgetary decisions, and perform high level complex administrative actions for EGRP; HHS Contracting Officer's Technical Representative for scientific and technical support contracts
- Executive Director (since 2005) of NCI's Cohort Consortium, a diverse and international consortium of 60 large cancer epidemiology cohorts; Lead for the consortium's strategic planning and implementation process, and revision of consortium Bylaws
- Direct strategic and scientific planning activities – lead facilitator for the EGRP strategic planning and implementation process; Manage overall program planning, operations, reorganization / realignment activities, working closely with senior program and division leadership to direct and guide program goals; Support analyses of grants and funding opportunities, and synthesize critical data, to inform the development of new program initiatives and methodologies; Supervise and mentor permanent and temporary staff
- Convene stakeholder meetings to foster cross-study collaborations, transdisciplinary research and scientific direction of program; Collaborate across NCI, other NIH institutes, HHS agencies and non-federal organizations on cross-cutting initiatives

National Cancer Institute
DCCPS – EGRP and the Office of Cancer Survivorship
Program Analyst 12/2004 – 7/2008

- Branch analyst, Program planning and budget lead. Managed large extramural research grant portfolios; conducted portfolio analyses on grants, program funding and scientific initiatives; analyzed proposals and recommended funding actions; Implemented communication strategies to promote research priorities and evaluate progress; Synthesized, disseminated and monitored program objectives for compliance with NIH administrative regulations and grant policies; Supervised staff, provided technical guidance on complex funding processes, and managed program operations and administrative actions such as recruitment and hiring of personnel; Led administrative process, in coordination with leadership, for the reorganization of EGRP
- Served as Executive Director of the NCI Cohort Consortium, and Adjunct Investigator with the NCI Division of Cancer Epidemiology and Genetics, while co-leading the NCI Cohort Consortium's Vitamin D Pooling Project

The George Washington University School of Public Health and Health Services
Department of Environmental and Occupational Health, Washington, DC
Research Associate 10/2000 – 12/2004

- Project Manager for the *Mid-Atlantic Center for Children's Health and the Environment*, a Pediatric Environmental Health Specialty Unit (PEHSU) for EPA Region 3 (DC, MD, VA, DE, PA). Managed half million-dollar worth of grants funded by US EPA, Agency for Toxic Substances and Disease Registry and Association of Occupational and Environmental Clinics to conduct health education and research activities on children's environmental health. Coordinated research activities on grants focused on environmental and occupational health grants, as well as on inflammatory breast cancer research
- Supervised the Lead Safe DC project, a primary prevention program; developed and implemented recruitment strategies to educate and test lead levels in newborns in Washington, DC. Facilitated focus groups, conducted health surveys and analyzed results of a project on Pesticide Education and Training for health care providers.
- Small group instructor for graduate level environmental health course at George Washington University, School of Public Health and Health Services; Administered mandatory environmental and occupational health training sessions to DC Department of Health staff as part of first responder training on Weapons of Mass Destruction and Bioterrorism following September 2001 terrorist attacks

JSI Research & Training Institute, Inc., Rosslyn, VA
Graduate Student International Health and Development Intern 01/2001 – 05/2001

- Assisted with management and overall coordination of grant review activities and funding decisions for the Reproductive Health for Refugees (RHR) project; Led collaboration of internal and external stakeholders in the grant proposal application and review process that

resulted in timely award of small grants; Fostered communication among partner organizations and working groups to raise global awareness, secure resources and formulate strategic interventions for sexual and gender-based violence prevention in different parts of the world; Applied knowledge and experience gained from this internship to write my graduate school special project (Master's thesis – *Reproductive Health in Refugee Settings*)

Howard University Hospital
Department of Geriatrics, Washington, DC
Research Assistant 10/1999-10/2000

- Managed an NIH-funded multi-center clinical trial on Alzheimer's disease prevention; Recruited and enrolled elderly African American women into trial and completed pre-screening and administration of clinical trial drugs in the second phase; Administered and analyzed Neuropsychological tests on study participants; Developed educational and recruitment materials and presented at multiple health outreach events; Developed and managed database of study participants and related research project activities; Assisted with writing and editing grant proposals for funding, scientific reports and manuscripts

PUBLICATIONS

Swerdlow AJ, **Harvey CE**, Milne RL, Pottinger, CA, et al. The National Cancer Institute Cohort Consortium: an international pooling collaboration of 58 cohorts from 20 countries. *Cancer Epidemiol Biomarkers Prev.* 2018 July 17 pii: cebp.0182.2018

Kennedy AE, Khoury MJ, Ioannidis JP, Brotzman M, Miller A, Lane C, Lai GY, Rogers SD, **Harvey C**, Elena JW, Seminara D. The cancer epidemiology descriptive cohort database: A tool to support population-based interdisciplinary research. *Cancer Epidemiol Biomarkers Prev*. 2016;25(10):1392-1401. doi: 1055-9965.EPI-16-0412 [pii].

Sonderman SS, Bethea TN, Kitahara CM, Patel AV, **Harvey C**, Knutsen SF, et al. Multiple Myeloma Mortality in relation to obesity among African Americans. JNCI 2016; 108(10)

Cohen SS, Park Y, Signorello LB, Patel AV, Boggs DA, Kolonel LN, Kitahara CM, Knutsen SF, Gillanders E, Monroe KR, Berrington de Gonzalez A, Bethea TN, Black A, Fraser G, Gapstur S, Hartge P, Matthews CE, Park S, Purdue MP, Singh P, **Harvey C**, Blot WJ, Palmer JR. (2014) A Pooled Analysis of Body Mass Index and Mortality among African Americans. PLOS ONE 9(11): e111980.

Bethea TN, Kitahara CM, Sonderman J, Patel AV, **Harvey C**, Knutsen SF, et al. A Pooled analysis of body mass index and pancreatic cancer in African Americans. Cancer Epidemiol Biomarkers Prev 2014; 23(10): 2119-25.

Breast Cancer and the Environment, Prioritizing Preventions: *A report of the Interagency Breast Cancer and Environmental Research Coordinating Committee (2013)*, submitted to Secretary, U.S. Department of Health and Human Services

Muin J. Khoury, Andrew N. Freedman, Elizabeth M. Gillanders, **Chinonye E. Harvey**, Christie M. Kaefer, Britt C. Reid. Frontiers in Cancer Epidemiology: A Challenge to the Research Community from the Epidemiology and Genomics Research Program at the National Cancer Institute. Cancer Epidemiol Biomarkers Prev; 1–3. 2012 AACR

McGuinn LA, Ghazarian AA, Ellison GL, **Harvey CE**, Kaefer CM, Reid BC. Cancer and Environment: Definitions and Misconceptions. Environ Res. 2012; 112: 230-234

National Cancer Institute Investment in Pancreatic Cancer Research, Action Plan: Fiscal Year 2011 Action Plan, September 2010. http://www.cancer.gov/researchandfunding/priorities

National Cancer Institute, Pancreatic Cancer: A Summary of NCI's Portfolio and Highlights of Recent Research Progress, September 2010.
http://www.cancer.gov/cancertopics/types/pancreatic

Gallicchio L, Helzlsouer KJ, Chow WH, Freedman DM, Hankinson SE, Hartge P, Hartmuller V, **Harvey C**, Hayes RB, Horst RL, Koenig KL, Kolonel LN, Laden F, McCullough ML, Parisi D, Purdue MP, Shu XO, Snyder K, Stolzenberg-Solomon RZ, Tworoger SS, Varanasi A, Virtamo J, Wilkens LR, Xiang YB, Yu K, Zeleniuch-Jacquotte A, Zheng W, Abnet C, Albanes D, Bertrand K, Weinstein SJ. Circulating 25-Hydroxyvitamin D and the Risk of Rarer Cancers: Design and Methods of the Cohort Consortium Vitamin D Pooling Project of Rarer Cancers. Am J Epidemiol. 2010 Jul 1; 172(1): 10-20

Stolzenberg-Solomon RZ, Jacobs EJ, Arslan AA, Qi D, Patel AV, Helzlsouer KJ, Weinstein SJ, McCullough ML, Purdue MP, Shu XO, Snyder K, Virtamo J, Wilkins LR, Yu K, Zeleniuch-Jacquotte A, Zheng W, Albanes D, Cai Q, **Harvey C**, Hayes R, Clipp S, Horst RL, Irish L, Koenig K, Le Marchand L, Kolonel LN. Circulating 25-Hydroxyvitamin D and Risk of Pancreatic Cancer: Cohort Consortium Vitamin D Pooling Project of Rarer Cancers. Am J Epidemiol. 2010 Jul 1; 172(1): 81-93

Zheng W, Danforth KN, Tworoger SS, Goodman MT, Arslan AA, Patel AV, McCullough ML, Weinstein SJ, Kolonel LN, Purdue MP, Shu XO, Snyder K, Steplowski E, Visvanathan K, Yu K, Zeleniuch-Jacquotte A, Gao YT, Hankinson SE, **Harvey C**, Hayes RB, Henderson BE, Horst RL, Helzlsouer KJ. Circulating 25-Hydroxyvitamin D and Risk of Epithelial Ovarian Cancer: Cohort Consortium Vitamin D Pooling Project of Rarer Cancers. Am J Epidemiol. 2010 Jul 1; 172(1): 70-80

McCullough ML, Weinstein SJ, Freedman DM, Helzlsouer K, Flanders WD, Koenig K, Kolonel L, Laden F, Le Marchand L, Purdue M, Snyder K, Stevens VL, Stolzenberg-Solomon R, Virtamo J, Yang G, Yu K, Zheng W, Albanes D, Ashby J, Bertrand K, Cai H, Chen Y, Gallicchio L, Giovannucci E, Jacobs EJ, Hankinson SE, Hartge P, Hartmuller V, **Harvey C**, Hayes RB, Horst RL, Shu XO. Correlates of Circulating 25-Hydroxyvitamin D: Cohort

Consortium Vitamin D Pooling Project of Rarer Cancers. Am J Epidemiol. 2010 Jul 1; 172(1): 21-35

Helzlsouer KJ; VDPP Steering Committee. Overview of the Cohort Consortium Vitamin D Pooling Project of Rarer Cancers. Am J Epidemiol. 2010 Jul 1;172(1):36-46

Harvey CE, Lynch SM, Rogers SD, Winn DM. The National Cancer Institute (NCI) Consortium of Cohorts [abstract]. In: Proceedings of the 99th Annual Meeting of the American Association of Cancer Research; 2008 April 12-16; San Diego, CA. Philadelphia (PA): AACR; 2008. p 335. Abstract nr 3865

Paulson, J.A.; **Harvey, C.E**. Animal Safety. In Safe and Healthy School Environments. Frumkin, H., Geller, R.J., Rubin, I.L., Nodvin, J., Eds.; Oxford University Press: New York, 2006; Chapter 7

Balbus, John M, **Harvey, Chinonye E**, McCurdy Leyla. Educational Needs Assessment for Pediatric Health Care Providers on Pesticide Toxicity. Journal of Agromedicine, 2006, 11(1):27-38

Harvey Chinonye E, Guidotti Tee L. The Role of Pediatric Environmental Health Specialty Units (PEHSU) in Asthma Management through Parent and Practitioner Education. Journal of Children's Health, 2004, 2(1):3-9

PROFESSIONAL TRAINING

- HHS Contracting Officer's Representative (COR) Level II Certification, 2005 - present
- HHS - Appropriations Law training, June 2011 to present
- NIH - Simplified Acquisitions Training - annual purchase card refresher training; Funding Agreement System; Green Purchase training (2005 – present)
- NCI – NCI Leadership Education Action Program (2015)
- NCI – The Art of Crucial Conversations Training (2015)
- NIH - Supervisory Essentials Training, July 2012
- NIH - Holding Employees Accountable Training, June 2012
- NCI - The Empowered Supervisor Program, February 2012
- NCI - Executive Coaching and Leadership Program, January - June 2012
- NCI - Coaching Skills for Managers and Supervisors for NCI, February 2011
- NIH - Leadership Skills for Non-Supervisors Training, May 2010
- The Human Element, 3-Day Radical Collaboration Course, October 2009
- NIH - Grants Management training 2005
- The University of Michigan School of Public Health Department of Epidemiology, Certificate of Participation, Graduate Summer Session in Epidemiology, 2008
- Johns Hopkins Bloomberg School of Public Health, Graduate Summer Institute of Epidemiology and Biostatistics - Genetic Epidemiology and Genome-Wide Association Studies course, July 2008

PROFESSIONAL AND ACADEMIC HONORS AND AWARDS

- Golden Key International Honour Society member (2016 – present)
- NIH Director's Merit Award for working across Institutes and Centers and scientific disciplines providing sustained leadership, scientific direction, and programmatic management of the Breast Cancer and the Environment Research Program, September 2014
- NIH Director's Merit Award for leadership on the NCI Cohort Consortium Secretariat - Federal members, November 2013
- NIH Director's Award for outstanding initiative, cooperation, synergy and productivity in support of management and scientific mission of EGRP, DCCPS, NCI, 2008
- NCI's DCCPS-DCEG Award for outstanding service to the NCI Cohort Consortium Secretariat, 2006-2007
- NCI's DCCPS-DCEG Award for leadership on the NCI Cohort Consortium Vitamin D Pooling Project, 2009
- Academic Merit Award, The George Washington University School, August 1999
- Honorary Award Recognition with biography published in the 21st Annual Edition of The National Dean's List, 1997-1998, Honoring America's Outstanding College Students
- Alpha Epsilon Delta, Pre-medical Honors Society, Michigan Gamma Chapter, 1996-1998

THIS PAGE INTENTIONALLY LEFT BLANK