

# **Indexing Case Series Articles: A Data-Driven Approach**

**Andrew Shahidehpour, MEng, PhD, Arthur W. Holt, MBA, Ang Michael Troy  
and Neil R. Smalheiser, MD, PhD**  
**University of Illinois College of Medicine, Chicago, IL 60304 USA**

## **Abstract**

*Case reports and case series comprise a significant portion of the biomedical literature, yet unlike case reports, NLM does not index case series as a Publication Type. This hurts clinicians' and researchers' ability to identify and analyze evidence from this type of study. We characterized the PubMed articles that mention "case series" in the title or abstract. We removed articles which discuss (rather than report the results of) case series studies, as well as those better indexed as other standard publication types. A random sample of these articles was evaluated by two annotators who confirmed that the great majority satisfy a formal definition of "case series". The endpoint is a corpus of case series studies that is suitable to use as a training set for automated machine learning indexing methods.*

## **Introduction**

Case reports and case series are types of observational study designs which represent a significant part of the biomedical literature. Over 2.4 million case report articles are published and indexed within PubMed, and roughly 100,000 case series are published as well. Together, case reports and case series form the lowest tier of the Evidence Hierarchy in evidence-based medicine, but despite their lowly status, they contribute in important and unique ways to medical progress [1-3]. Generally, case reports are unplanned eyewitness reports of one or a few patients, in contrast to case series which are generally planned analyses of a larger set of patients, often five or more. Whereas Case Reports is a recognized Publication Type and indexed as such in MEDLINE and PubMed, no NLM indexing of case series articles exists at all, creating a gap in the ability of researchers to identify and analyze evidence from this type of study.

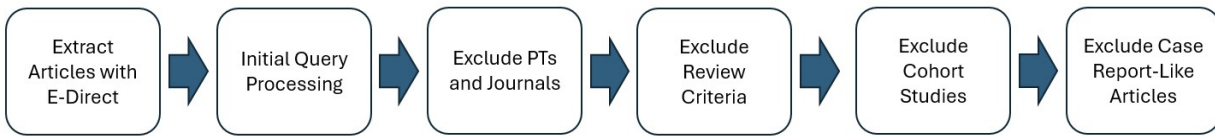
The traditional means of indexing articles involves deciding on a formal definition of "case series", and training annotators to manually evaluate each article as it is published. Alternatively, one could create a manually annotated corpus of articles that can be used as a training set for automated machine learning methods [4-6], but it requires a large amount of time and effort to train and manually annotate a large, representative gold standard corpus. As an alternative, we have pursued a data-driven approach.

First, we aimed to characterize the overall body of PubMed articles that mention the phrase "case series" in the title or in the abstract. We hypothesize that articles which employ the phrase "case series" in the title should largely be articles that authors themselves feel are case series, or that discuss one or more case series (e.g., a review of case series on a given topic, or a discussion of proper design methodology). Articles mentioning "case series" only in the abstract may also include some case series studies, though we expect that they may often discuss case series or mention them incidentally. We acknowledge that authors are not always well-informed or correct about using the term "case series" [7-9], and in fact our analyses revealed certain patterns of error in this regard. However, using author usage as a starting point seems to be a pragmatic strategy for defining "case series" studies.

Second, using these articles as a starting point, we aimed to remove the bulk of those articles which discuss (rather than report the results of) case series studies, as well as remove those better indexed as other standard publication types. The endpoint should be a corpus suitable to use as a training set for automated machine learning indexing methods.

## **Methods**

As shown in the flowchart (Figure 1), articles mentioning the phrase "case series" in title or abstract were retrieved from PubMed, processed, and progressively filtered, resulting finally in a set of articles, the vast majority of which satisfies a formal definition of a case series study.



**Figure 1.** Flowchart of case series articles processing and exclusions.

### Preprocessing and feature extraction

The Arrowsmith biomedical tokenizer ([https://arrowsmith.psych.uic.edu/arrowsmith\\_uic/download/tokenizer.txt](https://arrowsmith.psych.uic.edu/arrowsmith_uic/download/tokenizer.txt)) was used to lowercase and word tokenize the title and abstract text. To process the article title and abstract text and MeSH terms, the Arrowsmith 1400 word stoplist ([https://arrowsmith.psych.uic.edu/arrowsmith\\_uic/data/stopwords\\_1400](https://arrowsmith.psych.uic.edu/arrowsmith_uic/data/stopwords_1400)) was implemented, except that the following words were excluded from the stoplist: “case”, “one”, “two”, “three”, “four”, “five.” “Case” was excluded due to the importance of “case series” and “case reports”, and the number words were excluded to allow for later sample size extraction (i.e., number of cases or patients described in the study). The following highly frequent stop words were removed from the MeSH terms: “Humans”, “Female”, “Male”, “Animals”, “Adult”, “Middle Aged”, “Aged”, “Adolescent, Child”, “Rats”, “Mice”, “Time Factors”, “Treatment, Outcome”, “Child, Preschool”, “United States”, “Aged, 80 and over”, “Pregnancy”, “Risk Factors”, “Infant.” Unigrams, bigrams, and trigrams were extracted from the processed article titles and abstracts. N-grams appearing in less than 10 articles or more than 70% of articles were removed. For the remaining n-grams, their document frequency as a percentage of the total articles in each set was calculated. A similarly normalized document frequency was calculated for each MeSH term, publication type, and journal.

Bigrams matching the pattern “X case(s)” or “X patient(s)” where X is a number were extracted and written number words were converted into numerals (e.g., “twenty” into “20”). “COVID-19”, “case series”, “case report”, and “case control” mentions were converted to “[removed COVID-19 mention]”, “[removed case series]”, “[removed case report]”, and “[removed case control]” to avoid erroneously extracting these references as sample sizes. The bigrams were then grouped into ranges of 1-5, 6-10, 11-19, and  $\geq 20$  in order to better capture the number of patients or cases discussed by articles. 100 articles were manually evaluated to ensure each article was sorted into the correct category.

### Multi-Tagger model for assigning publication types

Many PubMed articles are not indexed for publication types by NLM. In addition to employing the NLM indexing terms on each article, if present, we also inferred their assignment to one or more of 50 publication types and study designs as predicted by a probabilistic machine learning-based model developed by our group, Multi-Tagger [4], which employs metadata features such as title, abstract, journal, MeSH terms, and number of authors.

Multi-Tagger scores were used to differentiate “case report-like” vs. “non-case report-like” articles (Figure 2). The case report probability score assigned to each article in the retrieved [title] and [abstract] sets was assigned “case report-like” if either it is indexed as a case report by NLM or predicted by Multi-Tagger using an optimized F1 decision threshold of 0.491 (Figure 2). If neither NLM nor Multi-Tagger predict the article to be a case report, it is deemed non-case report-like.

### Definition of case series article for the purpose of evaluating the proposed training set

Several formal definitions of “case series” were examined, including published definitions [7-9] as well as those proposed by Cochrane [[https://community.cochrane.org/sites/default/files/uploads/inline-files/Definitions-Study-Characteristics\\_Cochrane-COVID-19-Study-Register\\_0.pdf](https://community.cochrane.org/sites/default/files/uploads/inline-files/Definitions-Study-Characteristics_Cochrane-COVID-19-Study-Register_0.pdf)], Wikipedia [[https://en.wikipedia.org/wiki/Case\\_series](https://en.wikipedia.org/wiki/Case_series)], and Sage Research Methods [<https://methods.sagepub.com/encyc/edvol/encyc-of-epidemiology/chpt/case-reports-case-series>]. We settled on a working definition of case series that, while not official, was employed for objectively evaluating and characterizing the articles in the proposed training set:

*A case series is a descriptive study that follows a group of patients who have a similar presenting history, diagnosis, clinical presentation and progression, or prognosis in individual patients, or who are undergoing the same procedure, or share an adverse event, over a certain period of time.*

Several clarifying notes were added for annotators:

- A case series is usually a planned study and generally does not consist of incidental observations.
- A case series is always a group of patients, often 4 or more, but sometimes a large number (e.g., >50).
- A case series usually has no control group (except for self-controlled studies where the same patient is their own control).
- A case series can satisfy more than one publication type or study design at the same time, e.g. a case series could also incorporate a review, or could also be described as a case report, a cohort study, or even other types of studies within the same article.
- A case series must be observational and cannot be an interventional study. The patients may have undergone a treatment or procedure, and the authors can comment on the efficacy of the treatment or procedure. This may be assessed by e.g., a retrospective chart review. However, if patients are actively recruited for the study, this is an uncontrolled clinical trial instead. The only exception would be if a single article reported a case series study AND a clinical trial study.
- A case series differs from a cohort study insofar as a cohort follows and contrasts TWO different groups of subjects whereas a case series follows one group. However, sometimes a case series compares two subgroups, in which case the distinction between a case series and a cohort becomes difficult to decide.

## Results

### Retrieving the initial sets of “Case Series” articles and performing initial exclusions

Two article sets, referred to as the [title] set and [abstract] set, were gathered using the PubMed E-Utilities tool to query PubMed and extract articles. The first query gathered articles mentioning the phrase “case series” in the title, while the second focused on articles with “case series” in the abstract (but not the title). Since PubMed does not support searching by abstract alone, the [abstract] set was extracted by retrieving articles with “case series” in the title or abstract and then discarding the articles with empty abstracts or “case series” in the title. Both sets included articles published 01/01/1987 - 12/31/2023 and written in English. Each article was downloaded in XML format and the PMID, article title, author last names, abstract text, publication year, publication types, MeSH terms, journal title, journal ISO abbreviation, and page numbers were extracted. Articles indexed according to the following NLM Publication Types and MeSH study designs either in PubMed or Multi-Tagger were immediately excluded for both [title] and [abstract] sets as being inherently inconsistent with being a case series study: "Published Erratum", "Retraction of Publication", "Retracted Publication", "Duplicate Publication", "Bibliography", "Portrait", "Legal Case", "Lecture", "Congress", "Pictorial Work", "Newspaper Article", "Book Illustrations", "Webcast", "Video Audio Media", "Electronic Supplementary Materials", "Comment", "Editorial", "Case Control Studies", "Clinical Trial", "Controlled Clinical Trial", "Randomized Controlled Trial", "Clinical Trial, Phase I", "Clinical Trial, Phase II", "Clinical Trial, Phase III", "Clinical Trial, Phase IV", "Clinical Trial Protocol", "Pragmatic Clinical Trial", "Clinical Trial, Veterinary", and "Randomized Controlled Trial, Veterinary."

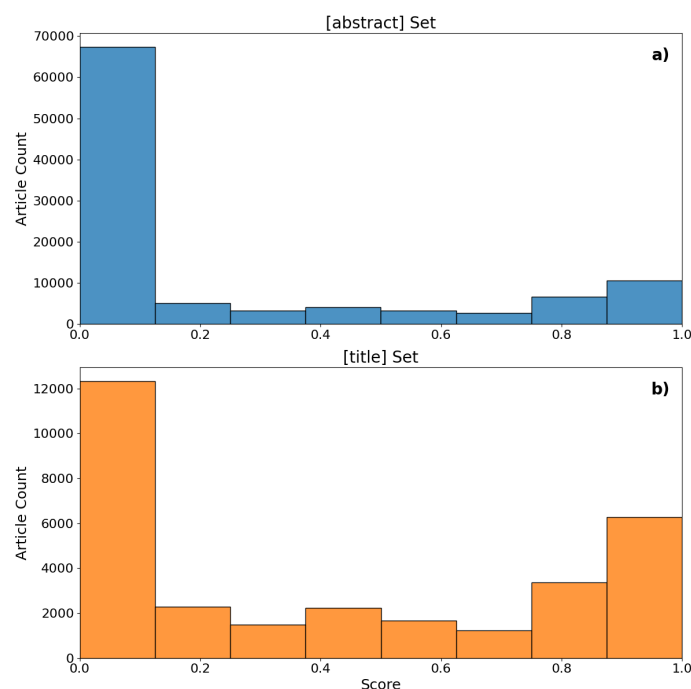
We also excluded articles published in the following methodological or statistical journals, as likely to be discussing methodology rather than presenting results of a case series study: "Ann Epidemiol", "Bioinformatics", "Biom J", "Biostatistics", "BMC Genet", "BMC Med Genomics", "BMC Med Res Methodol", "BMC Syst Biol", "Br J Math Stat Psychol", "Bull Math Biol", "Clin Trials", "Comput Methods Programs Biomed", "Contemp Clin Trials", "Control Clin Trials", "Epidemiology", "Eur J Epidemiol", "Genet Sel Evol", "IEEE ACM Trans Comput Biol Bioinform", "IEEE Trans Pattern Anal Mach Intell", "IMA J Math Appl Med Biol", "Int J Health Geogr", "Int J Methods Psychiatr Res", "J Anim Breed Genet", "J Biopharm Stat", "J Clin Epidemiol", "J Pharmacokinet Pharmacodyn", "J R Soc Interface", "Lifetime Data Anal", "Math Biosci", "Math Med Biol", "Mol Syst Biol", "Neural Netw", "Pharm Stat", "Phys Rev E Stat Nonlin Soft Matter Phys", "PLoS Comput Biol", "Psychol Methods", "Stat Appl Genet Mol Biol", "Stat Med", "Stat Methods Med Res", and "Theor Biol Med Model".

Since case series studies often include a review of the literature, articles indexed as reviews were not removed initially. However, at this point, articles mentioning both terms “systematic review” and “reports” in the title were removed from the [title] set, as these were comprised of systematic reviews covering case reports and case series.

### “Case series” [title] set

The initial set of 33,720 articles mentioning “case series” in the title was trimmed by removing those without abstracts, with incompatible publication types, those published in methodological and statistical journals, and those with “reports” and “systematic reviews” in the title, leaving 28,829 articles in the initial set to be characterized further.

A striking finding is that more than half of these articles are indexed as case reports by NLM or predicted by the Multi-Tagger model. In fact, in the [title] set, the distribution of Multi-Tagger predictive scores for the Case Reports publication type was clearly bimodal (Figure 2). Of note, 8 of the top 20 journals publishing case report-like articles explicitly had “case reports” in the title of the journal (vs. only 1 of the top 20 in the other subset).



**Figure 2.** Distribution of Multi-Tagger predictive scores for the case report publication type for both the initial a) [abstract] and b) [title] sets. The optimized F1 score decision threshold was 0.491 assigning articles with a higher score as case report-like and articles with a lower score as non-case report-like.

Case report-like (i.e., indexed or predicted to be Case Reports) and non-case-report like (i.e., all others) subsets were characterized further by extracting statements of the form “X cases” and “X patients” mentioned in the title or abstract separately, where X is a number written either as a numeral or as a word. As shown in Table 1, in the [title] set, 44.7% of the case report-like articles mentioned 5 or fewer cases or patients in the abstract, vs. 12.9% of the non-case report like subset. Conversely, only 13.2% of the case report-like articles mentioned 11 or more cases or patients in the abstract, vs. 42.3% of the non-case report-like subset (Table 1).

Several previous analyses have pointed out that many case series articles satisfy study design criteria of descriptive and/or population-based cohort studies [7-9]. However, we found that only 5.0% of the non-case report-like articles were indexed by NLM or predicted by Multi-Tagger as Cohort Studies, and only 2.51% were indexed as Clinical Study. Other clinical publication types were equally or more uncommon (<1%), e.g., Evaluation Study, Predictive

Value of Tests, Cross-Sectional Studies, and Longitudinal Studies. This argues that the non-case report-like articles comprise a unique type of biomedical literature, not well covered by any existing indexing terms.

	Case Report-Like		Non-Case Report-Like	
	[title] Set	[abstract] Set	[title] Set	[abstract] Set
1-5 case(s)	3,162 (20.57%)	1,243 (11.93%)	436 (3.45%)	659 (1.72%)
1-5 patient(s)	3,705 (24.10%)	2,841 (27.27%)	1,189 (9.42%)	2,579 (6.71%)
6-10 case(s)	670 (4.36%)	268 (2.57%)	335 (2.65%)	440 (1.15%)
6-10 patient(s)	1,521 (9.89%)	1,119 (10.74%)	1,565 (12.40%)	2,975 (7.74%)
11-19 case(s)	253 (1.65%)	100 (0.96%)	272 (2.16%)	422 (1.10%)
11-19 patient(s)	599 (3.90%)	433 (4.16%)	1,470 (11.65%)	3,683 (9.59%)
>=20 case(s)	402 (2.61%)	161 (1.55%)	594 (4.71%)	1,543 (4.02%)
>=20 patient(s)	779 (5.07%)	568 (5.45%)	3005 (23.81%)	13,138 (34.20%)
None	4,283 (27.86%)	3,686 (35.38%)	3,755 (29.75%)	12,976 (33.78%)

**Table 1.** Abstract mentions of “X cases” and “X patients” in the [title] and [abstract] sets, separated into case report-like and non-case report-like subsets.

#### “Case series” [abstract] set

The set of articles mentioning “case series” in the title or abstract (106,033 articles) was trimmed by removing articles with “case series” in the title, incompatible publication types, and those published in methodological and statistical journals, leaving 66,861 articles in the set to be characterized further. About 10% of the articles were case report-like (Figure 2). Similar to the [title] set, almost half of the non-case report-like articles in the [abstract] set mentioned >10 cases or patients (Table 1).

#### Manual evaluation of the provisional training set

At this point, we contemplated combining the non-case report-like [title] and [abstract] sets as a representative training corpus for indexing case series studies. However, it was unclear whether the great majority of these articles are “true” case series studies: that is, whether they would satisfy formal definitions of “case series”. Although there is no official, entirely accepted or consistent definition of “case series” [8], we examined several of the available definitions and created our own consensus definition to guide manual annotation (see Methods). We randomly chose 50 articles from the non-case report-like “case series” [title] subset and 50 from the “case series” [abstract] articles. We also added 10 randomly chosen case report-like articles, to learn whether these would also satisfy definitions of case series. These 110 articles were shuffled and presented to two annotators, blind to article assignment, who independently read the title and abstract (and full-text if necessary), scoring whether it satisfied our formal working definition of “case series”, noting the type of design and any unusual features, and extracting the total number of patients studied. Differences were reconciled by discussion.

Of the 50 “case series” [title] articles, 88% were judged to satisfy the working definition. Three prominent subtypes of case series were detected: One subtype followed a group of patients who shared a given diagnosis or condition; one followed a group of patients who were subjected to a particular type of surgery or other treatment or intervention; and a third carried out anatomical or technical measurements on samples derived from subjects. Most studies were retrospective, ranging from three to thousands of patients. A few studies, which had both prospective design and active recruitment of patients, were regarded as uncontrolled clinical trials and scored outside the definition of a case series study. A few articles compared subjects to control groups, which also excluded them from our definition. Note that of the 10 case report-like articles evaluated, 9 did satisfy the definition of “case series”. Of the two articles that were predicted to be cohort studies according to the Multi-Tagger model, both satisfied the criteria to be called case series as well. Thus, the scope of case series studies vs. case reports is not entirely distinct, and a single article may show more than one type of design.

Among the 50 “case series” [abstract] articles, the same range of study heterogeneity was observed as in the [title] set. However, 15 of the articles were systematic reviews or other reviews (not reporting the results of an individual case series study) that had not been adequately removed by previous rules. Apart from these reviews, 94.2% of the [abstract] articles were judged to satisfy the working definition of case series. We then tested an additional exclusion rule, to remove any articles from the [abstract] set that were indexed as Review or Systematic Review according to NLM or Multi-Tagger. Since this rule removed 14 of the 15 reviews from the test sample but none of the other articles, the rule was thus implemented across the entire [abstract] set.

After the evaluation of the 110 articles, we implemented the review exclusion rule in the [abstract] set (as just discussed), further removed any articles indexed as Cohort Studies by NLM or predicted by Multi-Tagger in the combined set, and finally removed the case-report like articles. The final proposed training corpus thus consisted of 12,621 [title] articles and 38,415 [abstract] articles, or a total combined of 51,036 case series studies.

## Discussion and Conclusions

We have characterized the set of biomedical articles that mention the phrase “case series” in the title or abstract, in order to understand what types of studies are regarded as case series by the authors themselves, as well as to identify articles that are not case series studies – either because a) they discuss methodology or review other case series studies (rather than present results of an individual study), or b) because they are better indexed as other types of studies, particularly case reports, cohort studies, review articles, or uncontrolled clinical trials. Our goal has been to remove the latter articles as far as possible, leaving a curated set of case series articles that can be utilized as a training corpus for automated machine learning indexing methods.

As shown in Table 2, only about half of the articles that mention “case series” in the title or abstract are actually typical case series studies; most of the remainder are better described as case reports, reviews, or discussions of methodology.

	[title] set	[abstract] set	Sum	percentage
Methodology	46	128	174	0.2%
Review	90	15593	15683	15.1%
Case Report	15374	10419	25793	24.9%
Case Series	12621	38415	51036	49.2%
Other	5589	5381	10970	10.6%
Sum	33720	69936	103656	100%

**Table 2.** Estimates of article types as a percentage of all retrieved articles mentioning “case series” in the title or in the abstract, based on the exclusion procedures described in this paper.

A major finding is that almost half of articles that mention the phrase “case series” in title strongly resemble case reports, both because they are explicitly indexed as such by NLM (and/or predicted as such by our publication type model Multi-Tagger [4]), and because they predominantly deal with a much smaller number of patients than do the non-case report-like case series articles. We removed case report-like articles from our final case series training corpus, with the rationale that they are better indexed as Case Reports[Publication Type]. However, it should be noted that they generally satisfied our formal working definition of case series, indicating that case series and case reports are not entirely exclusive concepts [10].

Another surprising finding is that very few articles that mention the phrase “case series” are indexed as Cohort Studies by NLM, nor predicted as such by Multi-Tagger. Several prior analyses have pointed out that many, perhaps most case series also share aspects of design with descriptive and/or population-based cohort studies [7-9], wherein a single group of subjects are followed over time, either after a particular diagnosis was made or after a particular intervention was carried out. However, even if case series do share general features with descriptive and population-based cohort studies, these stand in contrast to classical analytic cohort studies which compare two groups or subsets of patients [<https://www.ncbi.nlm.nih.gov/mesh/?term=cohort+studies>].

A limitation of our study is that we did not attempt to ascertain how many case series studies exist in the literature that do not mention the phrase “case series” at all. We suspect that most authors publishing case series articles will mention that phrase explicitly, but this is a potential source of bias that might potentially cause the training corpus to underestimate the true number and heterogeneity of case series articles. Conversely, the fact that all the articles in our proposed training corpus all mention “case series” could potentially cause an overestimation of its importance as a feature for machine learning. This effect can be dealt with in at least two ways: First, articles that mention this phrase but were excluded from the training corpus can serve as negative examples and provide an estimate of the a priori probability that the phrase “case series” is predictive of a case series article. Second, one can potentially perform down weighting of the phrase as a feature during training of the machine learning model.

In conclusion, our analysis indicates that no existing publication type or study design indexing term captures typical case series studies, supporting our effort to create a training corpus and create a new specific “case series” indexing term. The training corpus has been deposited in the UIC INDIGO data repository (<https://indigo.uic.edu/https://doi.org/10.25417/uic.28593611.v1>). In the future, we plan to investigate whether case series articles are amenable to indexing using PubmedBERT-based transformer models which simultaneously assign predictive scores for case series as well as > 60 Publication types and study designs [5]. If so, the model scores will be evaluated and disseminated publicly for the use of the biomedical community.

## References

1. Murad MH, Sultan S, Haffar S, Bazerbachi F. Methodological quality and synthesis of case series and case reports. *BMJ Evid Based Med*. 2018 Apr;23(2):60-63. doi: 10.1136/bmjebm-2017-110853.
2. Smith EG, Patel KM. The Role of Case Series and Case Reports in Evidence-Based Medicine. *J Clin Psychopharmacol*. 2024 Mar-Apr 01;44(2):81-85. doi: 10.1097/JCP.0000000000001826.
3. Preskorn SH, Armstrong AG. Can the Publication of Case Series or Case Reports Lead to a Change in Clinical Practice? *J Psychiatr Pract*. 2023 Mar 1;29(2):137-141. doi: 10.1097/PRA.0000000000000701.
4. Cohen AM, Schneider J, Fu Y, McDonagh MS, Das P, Holt AW, Smalheiser NR. Fifty ways to tag your pubtypes: Multi-tagger, a set of probabilistic publication type and study design taggers to support biomedical indexing and evidence-based medicine. *medRxiv*. 2021 Jul 16:2021-07. doi: 10.1101/2021.07.13.21260468.
5. Menke JD, Kilicoglu H, Smalheiser NR. Publication Type Tagging using Transformer Models and Multi-Label Classification. *medRxiv* 2025 doi: 10.1101/2025.03.06.25323516 (Proc AMIA 2024, in press).
6. Wallace BC, Noel-Storr A, Marshall IJ, Cohen AM, Smalheiser NR, Thomas J. Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach. *J Am Med Inform Assoc*. 2017 Nov 1;24(6):1165-1168. doi: 10.1093/jamia/ocx053.
7. Sargeant JM, O'Connor AM, Cullen JN, Makielski KM, Jones-Bitton A. What's in a Name? The Incorrect Use of Case Series as a Study Design Label in Studies Involving Dogs and Cats. *J Vet Intern Med*. 2017 Jul;31(4):1035-1042. doi: 10.1111/jvim.14741.
8. Esene IN, Ngu J, El Zoghby M, et al. Case series and descriptive cohort studies in neurosurgery: The confusion and solution. *Childs Nerv Syst* 2014;30:1321–1332.
9. Dekkers OM, Egger M, Altman DG, et al. Distinguishing case series from cohort studies. *Ann Intern Med* 2012;156:37–40.
10. Abu-Zidan FM, Abbas AK, Hefny AF. Clinical "case series": a concept analysis. *Afr Health Sci*. 2012 Dec;12(4):557-62