Comparison of Changes in the Number of Included Patients Between Interventional Trials and Observational Studies Published from 1995 to 2014 in Three Leading Journals

Arnaud DEZETREE\textsuperscript{a}, Emmanuel CHAZARD\textsuperscript{a}, Daniel R. SCHLEGEL\textsuperscript{b}, Sylvester SAKILAY\textsuperscript{c}, Peter L. ELKIN\textsuperscript{c}, Grégoire FICHEUR\textsuperscript{a,1}

\textsuperscript{a}Department of Public Health, CHU Lille, CERIM EA 2694, Lille University, Lille, France

\textsuperscript{b}Department of Computer Science, SUNY Oswego, Oswego, New York, USA

\textsuperscript{c}Department of Biomedical Informatics, State University of New York at Buffalo, Buffalo, New York, USA

Abstract

Introduction: Since the late 1990s, research and administrative institutions have been developing health data warehouses and increasingly reusing claims data. The impact of these changes is not yet completely quantified. Our objective was to compare the change in the number of patients included per study between observational and interventional studies over a 20-year period starting in 1995.

Materials and methods: We extracted all abstracts from studies published in three leading medical journals over the period 1995–2014 (18,107 studies). Then, we divided our study into two steps. First, we constructed an SVM-based predictive model to categorize each abstract into “observational”, “interventional” or “other” studies. In a second step, we built an algorithm based on regular expressions to automatically extract the number of included patients.

Results: During the investigated period, the median number of enrolled patients per study increased for interventional studies, from 282 in 1995–1999 to 629 in 2010–2014. In the same time, the median number of patients increased more for observational studies, from 368 in 1995–1999 to 2078 in 2010–2014.

Discussion: The routine storage of an increasing amount of data (from data warehouses or claims data) has had an impact in recent years on the number of patients included in observational studies. The recent development of “randomized registry trials” combining, on the one hand, an intervention and, on the other hand, the identification of the outcome through data reuse, may also have an impact, over the next decade, on the number of patients included in randomized clinical trials.

This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).

\textsuperscript{1}Corresponding Author, Grégoire Ficheur, Public Health Department, Lille University Hospital, 2 avenue Oscar Lambret, 59037 Lille Cedex, FRANCE; gregoire.ficheur@univ-lille.fr.
Keywords
Data reuse; data warehouse; claims data; support vector machine

1. Introduction

Since the late 1990s, research and administrative institutions have been developing health data warehouses and increasingly reusing data from claims databases [1]. The availability of such data is associated with an increase in the number of included patients, and thus in the statistical power of observational studies. To the best of our knowledge, no work has studied the impact of increased data availability on biomedical research.

Moreover, several automated methods have been developed to conduct systematic reviews [2] considering a large number of published studies. Thus, machine learning can help categorize and extract data from text in order to improve these reviews [3–4]: such tools can optimize their practice, cornerstone of the evidence-based medicine.

Our objective was to compare the change in the number of patients included per study between observational and interventional studies over a 20 year period (1995–2014), by analyzing the publications of three leading medical research journals.

2. Material and Methods

2.1. Material

All the abstracts from the New England Journal of Medicine, the Lancet, and the Journal of the American Medical Association were extracted from the MEDLINE database. The extraction covered the period from January 1995 to December 2014. The extraction resulted in an XML file comprehendng different tags, including the “Citation/Article/Abstract” tag (article summary) and the “Citation/Article/PublicationType” tag (article type).

2.2. Methods

The abstracts were then classified into three categories according to the “PublicationType” tag: “intervention trial”, “observational study” and “other type of study”. Unfortunately, the “PublicationType” tag is an optional tag, so it was only filled for some of the abstracts: In order to overcome this difficulty we proposed a method to automatically identify the category of an abstract.

We defined a two-stage strategy. First, we constructed a predictive model to categorize each abstract into “observational”, “interventional” or “other” studies. In a second step, we built an algorithm based on regular expressions to automatically extract the number of patients enrolled.

Finally, we merged the results of the two previous steps to perform the analysis over time of the number of included patients, stratified by type of study.
2.2.1. Automatically classifying abstracts as “Interventional”, “Observational” or “Other”—The set of abstracts comprehended all the abstracts where the “PublicationType” tag was defined (n=9284). The existing values of this tag (n=72 different values) were classified by experts into three categories:

- “interventional” trial
- “observational” study
- “other” category of study: neither interventional nor observational

The abstracts were collected into a corpus, and the data were cleaned using several functions of the R’s “tm” package [5]. A documents-words matrix was constructed using the term frequency–inverse document frequency (tf-idf) weighting method to assign a weight to each of the terms and exclude the less frequent terms (frequency lower than 5%). As a result, 347 terms were selected to train the model.

The multiclass problem was managed following the One Vs One method as described by Pal et al. [6]. For each pair of classes, a learning sample (80%) and a test sample (20%) were randomly selected. A Support Vector Machine (SVM) classifier was used for the three pairs of classes using R’s “e1071” package [7] with a radial-based kernel (Table 1).

Finally, the class label that occurred the most was assigned to the corresponding abstract: the accuracy rate on the complete set for this algorithm was 95.5%.

2.2.2. Extracting the number of included patients—To extract the number of patients included in each study, we used ICU’s regexes, quite similar in form and behaviour to Perl’s regexes [8]. 133 different patterns were defined:

- [a number] then [a synonym for “patient”] (124 patterns)
- [a synonym for “including”] then [a number] (9 patterns)

In addition, 16 patterns were used to exclude numbers related to time, weight and distance.

Then we assessed the algorithm’s quality by comparing the numbers of patients extracted from a sample of 100 randomly selected abstracts (among those classified as interventional or observational studies according to MEDLINE) with the numbers found by an expert by reading those abstracts. An abstract was classified as true positive when both conditions were met: 1) the classifier could detect whether the number of included patients was reported or not, and 2) in case that number was reported, the classifier could automatically identify this number.

The recall on this sample was 68.5% [95% Confidence Interval (CI)] [58.0%; 77.8%] and the precision was 94.0% [95% CI] [85.4%; 98.3%].

2.2.3. Running the model over the entire dataset—The previously developed model was re-run for the entire dataset. The period was divided into four 5-year intervals.
A descriptive analysis was performed to compute the median (with first and third quartiles) of included patients for each of the four periods of interest. In addition, these results were plotted with a boxplot using a logarithmic scale for the number of patients.

Finally, the difference over time in the number of included patients between observational and interventional studies was analyzed: a rank-based linear model (which is efficient and robust to outliers) was performed to assess the association between, on the one hand, the “number of included patients” and, on the other hand, the interaction “five-year interval * type of study”.

3. Results

A total of 18,107 abstracts were analyzed: 11,456 were classified as interventional or observational. Among them, a number of included patients was found for 9,483 abstracts. Figure 1 and Table 2 show the comparative evolution of the number of patients by the type of study, over the period 1995–2014.

The median number of included patients remains similar until the period 2010–2014. During this last interval, the median number of included patients increased for observational studies, compared with interventional studies. This difference of evolution was confirmed by the rank-based estimation for linear model studying the interaction “five-year time interval * type of study” (p<0.001).

4. Discussion

Through this study, we have shown that in recent years, the number of included patients has strongly increased for observational studies compared to interventional studies. This increase in the number of included patients is observed concomitantly with the emergence of health data warehouses and reuse of claims data in medical research.

This study has been performed by automatically analyzing abstracts from 3 leadings journals, by SVM and regular expressions. This algorithm was accurate and useful. Nevertheless, the accuracy of this algorithm could be improved in several ways, in particular in the extraction of the number of included patients. For example, one could propose to train a model to automatically identify the patterns allowing the extraction of the number of patients included in the studies. Alternatively, the method may be improved by using full natural language processing [9] rather than just regular expressions to more accurately identify these values.

This work could also be supplemented by the extension of our research to a larger number of journals, as well as to journals dealing with other types of data, in order to compare changes between these different research areas. Moreover, we analyzed only abstracts but such a work could be carried out on entire articles. We could also consider the retrieval of a greater number of metadata in order to characterize more precisely the research type.

Finally, the recent development of “randomized registry trials” combining, on the one hand, an intervention and, on the other hand, the identification of the outcome through data reuse,
could also have an impact, over the next decade, on the number of patients included in randomized clinical trials [10]. Thus, the routine storage of an increasing amount of data (from data warehouses or claims data) has had an impact in recent years on the number of patients included in observational studies, but may also soon have an impact on the number of patients included in interventional studies.

References

[1]. Safran C, Using routinely collected data for clinical research, Stat Med. 10 (1991) 559–564. [PubMed: 1905417]

[2]. O’Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S, Using text mining for study identification in systematic reviews: a systematic review of current approaches, Systematic Rev (2015).

[3]. Bekhuis T, Demner-Fushman D, Screening nonrandomized studies for medical systematic reviews: A comparative study of classifiers, Artificial Intelligence in Medicine 55 (2012), 197–207. [PubMed: 22677493]

[4]. Jonnalagadda SR, Goyal P, Huffman MD, Automating data extraction in systematic reviews: a systematic review, Systematic Rev (2015).

[5]. Feinerer I, Hornik K, Package ‘tm’, CRAN, 2017.

[6]. Pal M, Multiclass Approaches for Support Vector Machine Based Land Cover Classification, arXiv:0802.2411, 2008.

[7]. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chang C, Lin C, Package ‘e1071’, CRAN, 2017.

[8]. Wickham H, Package ‘stringr’, CRAN, 2018.

[9]. Schlegel DR, Crowner C, LeHoullier F, Elkin PL, HTP-NLP: A New NLP System for High Throughput Phenotyping, Informatics for Health 2017 (2017), 276–280.

[10]. Lauer MS, D’Agostino RB Sr, The randomized registry trial--the next disruptive technology in clinical research?., N Engl J Med 369 (17) (2013), 1579–1581. [PubMed: 23991657]
Figure 1.
Comparative evolution of the number of patients by type of study and by period
Table 1.

Parameters of the three algorithms on the test dataset

| Pairs                      | Accuracy (test sample) |
|----------------------------|------------------------|
| Interventional vs Observational | 95.9%                  |
| Interventional vs Other    | 96.5%                  |
| Observational vs Other     | 94.7%                  |
Table 2.
Number of patients included in each study, by type of study and by period

| Time interval | Observational studies | Interventional studies |
|---------------|-----------------------|------------------------|
|               | n  | Median | Q1 | Q3  | n  | Median | Q1 | Q3  |
| [1995,1999]   | 866 | 368    | 75 | 2611| 2885| 282    | 79 | 1150|
| [2000,2004]   | 739 | 494    | 82 | 2042| 2316| 422    | 116| 1664|
| [2005,2009]   | 611 | 856    | 131| 6825| 1953| 701    | 202| 2422|
| [2010,2014]   | 477 | 2078   | 160| 31022| 1609| 629    | 182| 2303|