Secondary Use of Patient Data: Review of the Literature Published in 2016

D. R. Schlegel[1], G. Ficheur[2], Section Editors for the IMIA Yearbook Special Section "Secondary Use of Patient Data"

[1] Department of Computer Science, SUNY Oswego, Oswego NY, USA

[2] Department of Medical Informatics, EA 2694, Lille University Hospital, France

Summary

Objectives: To summarize recent research and emerging trends in the area of secondary use of healthcare data, and to present the best papers published in this field, selected to appear in the 2017 edition of the IMIA Yearbook.

Methods: A literature review of articles published in 2016 and related to secondary use of healthcare data was performed using two bibliographic databases. From this search, 941 papers were identified. The section editors independently reviewed the papers for relevancy and impact, resulting in a consensus list of 14 candidate best papers. External reviewers examined each of the candidate best papers and the final selection was made by the editorial board of the Yearbook.

Results: From the 941 retrieved papers, the selection process resulted in four best papers. These papers discuss data quality concerns, issues in preserving privacy of patients in shared datasets, and methods of decision support when consuming large amounts of raw electronic health record (EHR) data.

Conclusion: In 2016, a significant effort was put into the development of new systems which aim to avoid significant human understanding and pre-processing of healthcare data, though this is still only an emerging area of research. The value of temporal relationships between data received significant study, as did effective information sharing while preserving patient privacy.

Keywords

Data reuse; secondary use; big data in medicine

Yearb Med Inform 2017:
http://dx.doi.org/10.15265/IY-2017-032
Published online August 18, 2017

Introduction

Reuse, or secondary use, of data concerns the use of clinical data for a different purpose than the one for which it was originally collected. The data being reused are usually those owned by hospitals and health systems - large databases containing administrative, claims, and patient health data. Oftentimes this data is reused for research and applications in quality of care and patient safety. Techniques relying on the reuse of data are in opposition to conventional clinical research using data collected prospectively using pre-defined cohorts.

The literature surrounding the reuse of data is large and continues to grow - the problem is difficult and remains interesting even after some success has been obtained. The difficulties faced include the need to refactor and manage the data (sometimes across sites using different data formats which must interoperate), large numbers of variables and categories to be aggregated, issues with data quality (e.g. missing data), and maintaining security and privacy. These difficulties continue to be the subject of research, and developing solutions is becoming increasingly interdisciplinary - recently we have seen the use of deep neural networks [1] from the field of artificial intelligence employed in solving data reuse classification problems in medicine.

The reuse of medico-administrative data has also, for several years, been of interest in epidemiology and in particular in pharmacoepidemiology. Projects such as the Observational Medical Outcomes Partnership (OMOP) have empirically demonstrated the value of this data compared to the more traditional pharmacovigilance databases [2]. These same projects also confirmed, for the control of confounding factors in this observational context, the quality of cohort designs with the use of high-dimensional propensity scores and the particular interest of cross-over designs. Moreover, randomized control trials (RCTs) and routinely collected data have recently been considered together through "registry-based RCTs" [3], which is an active area of research.

Achieving reliable data reuse is a challenge worthy of our time and research, allowing for the identification of patients of interest for retrospective research (electronic phenotyping [4]). Electronic Health Record (EHR) data is huge (and therefore has a high statistical power), can be used without interfering with patient care, and is real data that can play a central role within a learning healthcare system.

The papers selected as best papers involve the development of systems which aim to predict patient outcomes, explore the value of temporal relationships in data, and discuss effective information sharing while preserving patient privacy. The following sections discuss the best papers selection method and emphasize notable characteristics of the best papers in the context of the wider literature.

Paper Selection Method

The databases PubMed/Medline and Web of Science® were searched for peer-reviewed papers published in the English language and which pertain to data reuse. The following Boolean expression was used: "secondary use" OR "data reuse" OR ("big data" AND ("health" OR "medicine" OR "medical")).

In addition, a complementary condition was used on Pubmed/Medline regarding the date of publication in 2016 (“2016/01/01”[Date - Publication]: “2016/12/31”[Date - Publication]). Three filters were added in the case of Web of Science® about the date of publication: “2016”, the field of interest: “Medical Informatics”, and the type of paper: “Article OR Proceedings”. Data reuse is an extremely large topic area and these query terms have the potential of excluding some papers, for example in epidemiology, in which data reuse is performed, but not discussed explicitly. Still, a total of 941 papers were retrieved. Papers were independently analyzed by the section editors on the basis of titles and abstracts. The documents were classified into two categories: “accept” or “reject” based on relevance and perceived impact. Each article labeled “accept” was examined in detail to finally reach a consensus list of 14 candidate best papers. In accordance with the IMIA Yearbook selection process [5], the candidate best papers were assessed by the two section editors and by two additional reviewers. Four papers were selected as the best papers (Table 1). A summary for each of them is given as an appendix.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2017 in the section ‘Learning from Experience: Secondary Use of Patient Data’. The articles are listed in alphabetical order of the first author’s surname.

Section Learning from Experience: Secondary Use of Patient Data
▪ Chen J, Podchiyska T, Altman R. OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records. J Am Med Inform Assoc 2016;23:339–48.
▪ Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep 2016;6:26094.
▪ Prasser F, Kohlmayer F, Kuhn KA. The Importance of Context: Risk-based De-identification of Biomedical Data. Methods Inf Med 2016;55:347-55.
▪ Saez C, Zurriaga O, Perez-Panades J, Melchor I, Robles M, Garcia-Gomez JM. Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: a systematic approach to quality control of repositories. J Am Med Inform Assoc 2016;23:1085-95.

Conclusions and Outlook

In 2016, noteworthy papers discussing secondary use of patient data focused on studying and improving the quality of clinical data, issues in sharing data, and predicting health outcomes using clinical data.

Secondary use of clinical data relies on the data being consistent across time and across departments and/or sites under study. A frequently used approach which continues to be applied to new domains is the use of standard ontologies and terminologies to apply a common semantic model to healthcare data. Sahoo *et al.*, implemented an informatics platform for epilepsy by surveying existing outcome data, identifying common data elements, and developing an epilepsy domain ontology to resolve issues of data heterogeneity [6]. Quality of clinical data goes hand-in-hand with the quality of the semantic resources the data is mapped with, which also has seen continued work this year (e.g., [7]). Sauer *et al.* [8], found that the United States Veterans Administration (VA) stores the results of pulmonary function tests (PFT) in structured, semi-structured, and unstructured forms. They used a natural language processing system to extract PFT results with a high degree of accuracy (F > .98). In this case, despite the different representations, they were able to avoid being formal about their semantic representations because of the narrow focus.

While the VA uses common data formats across sites, many who wish to share data between sites do not. Saez *et al.* [9], applied novel methods based on information theory and geometry to assess variability among multiple data sources and changes over time. In particular, they empirically studied data quality issues stemming from variation in probability distributions (due to population differences, biased practices, etc.) and time, concluding that “even if semantic and integration aspects are addressed in data sharing infrastructures, probabilistic variability may still be present.”

Work on identifying issues in sharing data between sites was another trend this year. Saez *et al.* [9], discussed the issue of data quality, but other issues addressed include the difficulty in understanding what’s in a large data collection, and issues in maintaining patient privacy when datasets are shared publically. Demner-Fushman’s paper on preparing radiology examination documents for distribution, including de-identification and indexing, noted that “an important step in facilitating secondary use of clinical document collections is easy access to descriptions and samples that represent the content of the collections” [10]. This is a part of the larger academic discussion of how researchers should handle scientific data in general [11], as is being studied by the CEDAR group at Stanford.

De-identification of data for sharing has become a significant concern as more clinical datasets are provided on the web for research. Prasser *et al.*, empirically examined the issue from the point of view of minimizing risk of re-identification, balancing increases in privacy with data quality, all while considering the data sharing context and the aims of the potential attackers [12]. Their use of risk models decreased information lost in de-identification by 10-24% depending on the strength of adversary they were protecting against.

This year reuse of clinical data trended on predicting outcomes. The contrast of two opposing approaches becomes clear when examining the literature as a whole: (i) identification by experts of the required data elements to build predictors (electronic phenotyping) for outcome prediction, and (ii) the “firehose” approach without pre-processing in which many more variables than may be needed are used. As with Saez *et al.,* above, Goldstein, *et al.* [13], continued a theme of focusing on temporal issues, identifying different data elements most predictive of mortality in patients receiving hemodialysis over different time horizons: vital signs in the near term, demographics and comorbidities in the long-term. In contrast, the OrderRex system [14] takes the “firehose” approach - automatically ingesting around 1,500 of the most common data elements from inpatient notes and performing association statistics in order to predict next order recommendations and outcomes. Importantly, they found that using temporal relationships between orders in their database improves results, from a precision at 10 recommendations of 33% to 38%.

A team at Mount Sinai has also developed an unsupervised method for learning directly from EHR data, this time using state-of-the-art artificial intelligence (AI) techniques such as feature learning and deep neural networks, called Deep Patient [15]. This system was used to predict whether patients would develop various diseases using random forest classifiers after using a deep neural network for feature extraction. The system was found to outperform other unsupervised learning mechanisms such as Principal Component Analysis (PCA) and Gaussian mixture models. Accuracy of the system was found to be quite high (.929) but the F-Score was still rather low (.181), even though it was better than all comparison systems. These “firehose”-based approaches are sure to continue gaining popularity as more structured and free text EHR data is annotated with standardized semantic resources for input into such systems. This strategy is related to those used in many papers reviewed by the section editors based on the extraction of large number of quantitative features in medical images (i.e. radiomics), and on the use of raw EHR data to build predictors, as in the work of Singh, *et al.* [16], identifying novel predictors of kidney failure from concepts extracted directly from clinical notes.

Acknowledgements

We would like to acknowledge the support of Brigitte Séroussi and John H. Holmes, along with the reviewers who assisted with the selection process.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
2. Ryan PB, Stang PE, Overhage JM, Suchard MA, Hartzema AG, DuMouchel W, et al. A comparison of the empirical performance of methods for a risk identification system. Drug Saf 2013;36 Suppl 1:S143–58.
3. Li G, Sajobi TT, Menon BK, Korngut L, Lowerison M, James M, et al. Registry-based randomized controlled trials- what are the advantages, challenges, and areas for future research? J Clin Epidemiol 2016;80:16–24.
4. Shivade C, Raghavan P, Fosler-Lussier E, , Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014;21:221–30.
5. Lamy J-B, Séroussi B, Griffon N, Kerdelhué G, Jaulent MC, Bouaud J. Toward a formalization of the process to select IMIA Yearbook best papers. Methods Inf Med 2015;54:135–44.
6. Sahoo SS, Zhang G-Q, Bamps Y, Fraser R, Stoll S, Lhatoo SD, et al. Managing information well: Toward an ontology-driven informatics platform for data sharing and secondary use in epilepsy self-management research centers. Health Informatics J 2016;22:548–61.
7. Kamdar MR, Tudorache T, Musen MA. A systematic analysis of term reuse and term overlap across biomedical ontologies. Semantic Web 2016;:1–19.
8. Sauer BC, Jones BE, Globe G, Leng J, Lu CC, He T, et al. Performance of a Natural Language Processing (NLP) Tool to Extract Pulmonary Function Test (PFT) Reports from Structured and Semistructured Veteran Affairs (VA) Data. EGEMS (Wash DC) 2016;4:1217.
9. Sáez C, Zurriaga O, Pérez-Panadés J, Melchor I, Robles M, García-Gómez JM, et al. Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: a systematic approach to quality control of repositories. J Am Med Inform Assoc 2016;23:1085–95.
10. Demner-Fushman D, Kohli MD, Rosenman MB, Melchor I, Robles M, García-Gómez JM. Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inform Assoc 2016;23:304–10.
11. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018.
12. Prasser F, Kohlmayer F, Kuhn KA. The Importance of Context: Risk-based De-identification of Biomedical Data. Methods Inf Med 2016;55:347–55.
13. Goldstein BA, Pencina MJ, Montez-Rath ME, Winkelmayer WC. Predicting mortality over different time horizons: which data elements are needed? J Am Med Inform Assoc 2017;24:176–81.
14. Chen JH, Podchiyska T, Altman RB. OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records. J Am Med Inform Assoc 2016;23:339–48.
15. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep 2016;6:26094.
16. Singh K, Betensky RA, Wright A, Curhan GC, Bates DW, Waikar SS. A Concept-Wide Association Study of Clinical Notes to Discover New Predictors of Kidney Failure. Clin J Am Soc Nephrol 2016;11:2150–8.

Correspondence to:
Daniel R. Schlegel
Department of Computer Science
396 Shineman Center
SUNY Oswego
Oswego NY, 13126, USA
E-mail: daniel.schlegel@oswego.edu

Summary of the Best Papers Selected for the 2017 Edition of the IMIA Yearbook, Special Section “Learning from Experience: Secondary Use of Patient Data”

Chen J, Podchiyska T, Altman R

OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records

J Am Med Inform Assoc 2016;23:339-48

Compliance with evidence-based guidelines is low and a majority of clinical decisions are not supported by randomized control trials. Thus, a large part of medical practice is thus driven by individual expert opinion. The authors present a clinical order recommender system which operates on a database which has been mined from existing patient data. The input to the data mining system is around 1,500 common electronic medical record (EMR) data elements (out of 5.4 million structured data elements) from labs results, orders, and diagnosis codes, including temporal separation in the form of patient timelines. This data was extracted for 18 thousand patients and stored in an association matrix. Queries to the database come in the form of clinical terms for the captured data elements for a patient. A ranking of suggested orders based on the input data and the association matrix is output to the user. By mixing outcomes such as death and hospital readmission in with the order results, the system also acts as a predictor of outcomes. The authors observe that including the temporal data increased precision from 33 to 38%, but also note that continued work is required to differentiate simply common behaviors on certain data from the correct ones.

Miotto R, Li L, Kidd BA, Dudley JT

Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records

Sci Rep 2016;6:26094

Proposed in this paper is a novel unsupervised deep feature learning method to derive a patient representation from EHR data that facilitates the prediction of clinical outcomes. Deep learning techniques, using neural networks with more than one hidden layer, have not previously been broadly used with EHR data. The authors used aggregated medical records from the Mount Sinai data warehouse with a stack of denoising autoencoders to capture stable structures and regular patterns from pre-processed EHR data. Then, they implemented random forest classifiers (one-vs.-all learning) to predict the probability that patients might develop a certain disease. On 76,214 test patients comprising 78 diseases from diverse clinical domains and temporal windows, the results significantly outperformed those achieved using representations based on raw EHR data and alternative feature learning strategies such as principal component analysis and Gaussian mixture models.

Saez C, Zurriaga O, Perez-Panades J, Melchor I, Robles M, Garcia-Gomez JM

Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: a systematic approach to quality control of repositories

J Am Med Inform Assoc 2016;23:1085-95

The authors propose the evaluation of variability in data distributions as a criterion which could be used systematically in assessing data quality. This variability is assessed first on different sources of data (i.e., from different sites), and second, over time. The authors proposed a novel statistics-based assessment method providing data quality metrics and exploratory visualizations. The method is empirically driven on a public health mortality registry of the region of Valencia, Spain, with >500,000 entries from 2000 to 2012, separated into 24 health departments. The repository was partitioned into two temporal subgroups following a change in the Spanish National Date certificate in 2009. Several types of data quality issues were identified including punctual temporal anomalies, and outlying or clustered health departments. The authors note that these issues can occur because of biases in practice, different populations, and changes in protocols or guidelines over time - none of which are solved through usual techniques of mapping to standard semantics.

Prasser F, Kohlmayer F, Kuhn KA

The Importance of Context: Risk-based De-identification of Biomedical Data

Methods Inf Med 2016;55:347-55

As data sharing becomes more common, concerns about maintaining the privacy of patients in such data sets is growing as well. International laws, such as HIPAA, and European Directive on Data Protection emphasize the importance of context when implementing measures for data protection. With methods of de-identification such as k-anonymity (dataset is transformed in such a way that each record is not different from k-1 other records), the degree of protection is high, but it is associated with a loss of information content. Indeed, a major challenge of data sharing is the adequate balance between data quality and privacy. The authors propose a generic de-identification method based on risk models, which assesses the risk of re-identification. An experimental evaluation was performed to assess the impact of different risk models and assumptions about the background knowledge/context of an attacker. Compared with reference methods, the loss of information was between 10% and 24% less, depending on the strength of the adversary being protected against.