Appendix: Summary of Best Papers Selected for the IMIA Yearbook 2018, Section Clinical Research Informatics

Caron A, Chazard E, Muller J, Perichon R, Ferret L, Koutkias V, Beuscart R, Beuscart JB, Ficheur G

IT-CARES: an interactive tool for case-crossover analyses of electronic medical records for patient safety

J Am Med Inform Assoc 2017;24(2):323-30

The increasing adoption of Electronic Healthcare Records (EHRs) is an opportunity for developing clinical epidemiological approaches based on the analysis of EHR data to evaluate the risk of adverse events following medical procedures. This paper describes the development and evaluation of an interactive tool to be used by clinical epidemiologists to systematically design case-crossover analyses of large electronic medical records databases for monitoring patient safety. Contrasting with the case-control design, in the case-crossover design the case and the control are one and the same person (albeit at different times). The advantage of the case-crossover design is to allow the investigator to control for time-constant confounding factors such as gender, age, weight, and lifestyle patterns.

The analytical tool IT-CARES implements a simple data model consistent with the case-crossover design in order to explore the association between exposures and outcomes. The exposures and outcomes are defined by clinical epidemiologists as lists of codes entered via a user interface screen.

IT-CARES is an interactive, freely-available, open source tool providing a user interface with three columns: (i) the outcome criteria in the left-hand column, (ii) the exposure criteria in the right-hand column, and (iii) the estimated risk (odds ratios, presented in both graphical and tabular formats) in the middle column. IT-CARES has been tested on data from the French national inpatient stay database which documents diagnoses and medical procedures for 170 million inpatient stays. Data collected between 2007 and 2013 were used to estimate the population-based risk of an acute thromboembolic or bleeding event (the primary outcome) following exposure to a medical procedure. The authors compared the results of their analysis with reference data from the literature and demonstrated that the estimated odds ratios were consistent with the literature data in terms of both the effect size and the persistence of risk over time. They also performed a negative control (carpal tunnel surgery) and as expected did not observe a significant elevation of the thromboembolic risk after this day-case surgery.

Although the risks of adverse events following medical procedures can be assessed (at least in part) in randomized controlled trials (RCTs), in some areas such as venous thromboembolism or bleeding, the RCTs conducted to date have failed to determine the long-term risk of adverse events following medical procedures. In addition, the external validity of RCTs is limited by the strict eligibility criteria and the short follow-up period, whereas a robust assessment of patient safety requires large population-based studies. In this context, the added value of IT-CARES is to allow clinical epidemiologists to design and rapidly execute in very large databases a complex case-crossover analysis which is the most suitable design for pharmaco-epidemiological population-based studies. Although IT-CARES provided reliable results in a test case, the authors will carry out additional research in order to evaluate the tool in additional patient safety studies and elaborate on its usability for advancing the end-user experience.

Girardeau Y, Doods J, Zapletal E, Chatellier G, Daniel C, Burgun A, Dugas M, Rance B

Leveraging the EHR4CR platform to support patient inclusion in academic studies: challenges and lessons learned

BMC Med Res Methodol 28 2017;17(1):36

The ability to perform protocol feasibility assessments on EHR data, initially to validate the likelihood of a protocol being able to recruit enough patients and subsequently to help a healthcare provider to target suitable candidate patients to screen, critically depends upon two success factors: the availability of EHR data of sufficient quality, and the computability of the eligibility criteria as EHR queries. The latter was the focus of the best paper by Girardeau et al., who examined the extent to which the criteria within three clinical research protocols in use at the Georges Pompidou European Hospital in Paris, France (HEGP) and at the Münster University Hospital, Germany (UKM) could be expressed as EHR queries. This study was undertaken as part of the EHR4CR (Electronic Health Records for Clinical Research) project, one of the largest Europe-funded public-private partnerships that has developed a computer platform to enable the reuse of data collected from EHRs.

The clinical studies selected for this research focused on (i) whether the pharmacokinetics of low molecular weight heparin is predictive of recurrent thromboembolism in cancer subjects, (ii) the effectiveness assessment of renal denervation in addition to standardised medical treatment in diabetic subjects with severe diabetic nephropathy, and (iii) a phase 3 study on Ewing Sarcoma. The three protocols had between seven and 10 inclusion criteria, and between five and 11 exclusion criteria, yielding a total of 67 distinct criteria. In the protocols, these criteria are expressed in free text, and the key step was to normalise the concepts within each criterion statement, in order to express these as distinct data items and set operators to connect them. The authors developed a six-step normalisation process to arrive at 114 distinct medical concepts, and as many computable expressions that could be executed as queries on the clinical data warehouses (research repositories derived from the hospital EHR systems) via the EHR4CR platform.

The authors found that 51 of the 67 criteria could be expressed computably. Of these, around 75% corresponded to data items mapped to locally used terminologies and were present in the structured data at the clinical data warehouses at HEGP and UKM. The authors discussed the challenge of nominating a suitable terminology to be the common mapping target for EHR queries, and why a multi-terminology approach may be most appropriate. Many of the criteria were complex, including concepts such as “at least”, “more than”, “if then”, which

require a suitable syntax for computable expression and query execution. The authors found that almost half of the criteria needed domain expert input to remove semantic ambiguity before a normalised expression could be derived.

The authors also discussed the necessity of having high quality EHR data in order to obtain accurate patient counts. Local knowledge may be required to interpret unexpected results in terms of missing data and other data quality impacts on the query results. They conclude that the exact reproducibility of the inclusion/exclusion criteria execution and a fair comparison of the query execution results are necessary when assessing the effectiveness of Clinical Trial Recruitment Support Systems.

Huang J, Duan R, Hubbard RA, Wu Y, Moore JH, Xu H, Chen Y

PIE: A prior knowledge guided integrated likelihood estimation method for bias reduction in association studies using electronic health records data

J Am Med Inform Assoc 2017 Dec 1

One of the greatest concerns about the reuse of EHR data in research is its fitness for purpose, since the data quality imperatives to support continuity of care and to support reuse for research are quite different. The assessment of data quality is an important topic in the CRI literature, in terms of assessment methodologies and specific forms of bias that may need correction to ensure valid scientific results. The paper by Huang et al., falls into the latter category, specifically considering the likelihood of misclassification of observational clinical data through algorithmic phenotype identification from coded data or natural language processing. They propose and provide evaluation evidence in favor of a statistical method termed PIE (prior knowledge-guided integrated likelihood estimation method). A prior distribution for the observation of interest is constructed using a realistic population of EHR data, and then algorithmic proxy measures are used to create the derived inclusion population. These populations (real and computed) are compared to define a distribution function for both sensitivity and specificity, and from this to derive an integrated likelihood. How this PIE function is calculated, and how it differs from conventional methods, is explained in detail in the paper.

The PIE method was evaluated using a Kaiser Permanente EHR data set of 2,022 patients with treated diabetes, the gold standard being two or more filled prescriptions for a diabetes medication. The surrogate measure to define the phenotype was calculated from several other diabetes diagnostic markers, and used to derive the sensitivity and specificity. Comparing the use of true sensitivity and specificity with the use of the PIE method showed that the latter reduced the over or under estimation bias by between 60% and 100% across a range of diabetes characteristics. In this evaluation, the authors were able to use a validation data set from Kaiser Permanente, but they proposed that in the absence of such a resource the literature and established data on prevalence rates may be used instead.

Jackson RG, Patel R, Jayatilleke N, Kolliakou A, Ball M, Gorrell G, Roberts A, Dobson RJ, Stewart R

Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project

BMJ Open 2017;7(1):e012012

The growth of literature and proliferation of clinical research platforms highlight the opportunity for reusing EHR data to identify suitable patients to be recruited into clinical trials. Severe Mental Illness (SMI) presents particular challenges for determining patient eligibility from EHRs because each mental health diagnosis spans considerable population heterogeneity, and the severity of an illness is better characterised by symptoms than by the diagnostic label. A further challenge is that mental health symptoms are usually documented in free text rather than encoded. The paper by Jackson et al., reports on research as part of the CRIS-CODE (Clinical Record Interactive Search Comprehensive Data Extraction) project, which has the long-term objective of offering comprehensive Natural Language Processing (NLP) models for mental health constructs. This study presents the capability of using NLP to extract mental health symptoms from clinical narratives at the South London and Maudsley NHS Foundation Trust, one of the largest mental healthcare organisations in Europe. The Clinical Record Interactive Search (CRIS) tool is in use at the Trust to maintain a pseudonymised shadow EHR, enriched with extracted concepts on 250,000 patients, which is being used for research.

In this paper the authors describe applying the TextHunter (machine learning) NLP tool suite on top of a ConText (context extraction) algorithm, and they document how these tools were configured to optimise the extraction of SMI symptoms, which symptom lexical strings were targeted, and how gold standard training annotation sets were developed. A total of 37,211 instances of the chosen 50 SMI symptoms were annotated from 32,767 documents to create gold standards and training data for each symptom, in order to develop and refine the extraction models.

The study itself comprised the analysis of 23,128 discharge summaries on patients with SMI and 13,496 discharge summaries on patients known to have no SMI. Due to the poor performance of four of the extraction models, these four symptoms were excluded from the final study. The authors demonstrated the ability to extract data for one or more of the 46 symptoms in 87% of patients with SMI and 60% of patients with non-SMI diagnosis, with a median F1 score of 0.88 (the harmonic average of the precision and recall, range 0 = worst to 1 = best).

The challenges and limitations of the work are clearly discussed. Perhaps the greatest one that still needs to be addressed is establishing the temporal progression of symptoms, in order to properly determine the most recent mental health state of patients and therefore more precisely predict their eligibility for a clinical trial.

By tackling arguably one of the more difficult semantic aspects of clinical documentation, the authors have not only advanced the opportunity to reuse EHRs for recruitment to mental health trials, but they demonstrated the potential of NLP to augment coded EHR data in general, across therapeutic areas.

Kim H, Bell E, Kim J, Sitapati A, Ramsdell J, Farcas C, Friedman D, Feupe SF, Ohno-Machado L

iCONCUR: informed consent for clinical data and bio-sample use for research

J Am Med Inform Assoc 2017;24(2):380-7

The deployment of EHRs associated with the emergence of Big Data technologies and new machine learning methods is an opportunity to exploit data at scale for generating new knowledge. In current practices, an opt-out approach is used for reusing de-identified data for research, explicit consent being considered as unnecessary or impractical for implementation in clinical settings. In the context of the recent General Data Protection Regulation (GDPR) promulgated by the European Union, approaches ensuring citizen-controlled dynamic, traceable and transparent consent management for processing, and exchanging EHR data are gaining interest.

This paper describes an implementation of a web-based tiered informed consent tool (iCONCUR) collecting patient preferences regarding the use of de-identified EHR data and bio-samples for research. The consent tool was installed in four outpatient clinics of an academic medical center and patients' preferences about the use of their data have been evaluated (394 participating patients, along which, 126 patients specified their preferences). The analysis is stratified by the demographic characteristics of the participants, data type sharing, and intent of use of the shared data. The majority consented to share most of their data and specimens with researchers. Their willingness to share varied according to the type of pathology (greater among participants from a Human Immunodeficiency Virus (HIV) clinic than for those from internal medicine clinics), the recipient of the data (higher number of items declined for for-profit institution recipients), and the type of the data (patients are most willing to share demographics and body measurements and least willing to share family history and financial data). Participants indicated that having granular choices for data sharing was appropriate, and that they liked being informed about who was using their data for what purposes, as well as about outcomes of the research.

The paper illustrates the implementation of an electronic informed consent system and reports the results of an excellent study on patient preference on data sharing. The study demonstrates that taking into account patient preferences increased satisfaction, and did not significantly affect participation in research. Dynamic consent was also proposed as a new approach that better serves both patients (i.e., data donors) and researchers (i.e., data receivers) in terms of promoting trust around data use and facilitating the recruitment and continuous management of study participants.