

# Clinical Research Informatics: Contributions from 2017

Christel Daniel<sup>1,2</sup>, Dipak Kalra<sup>3</sup>, Section Editors for the IMIA Yearbook Section on Clinical Research Informatics

<sup>1</sup> AP-HP Direction of Information Systems, Paris, France

<sup>2</sup> Sorbonne University, University Paris 13, Sorbonne Paris Cité, INSERM UMR\_S 1142, LIMICS, Paris, France

<sup>3</sup> University of Gent, Belgium

## Summary

**Objectives:** To summarize key contributions to current research in the field of Clinical Research Informatics (CRI) and to select best papers published in 2017.

**Method:** A bibliographic search using a combination of MeSH descriptors and free terms on CRI was performed using PubMed, followed by a double-blind review in order to select a list of candidate best papers to be then peer-reviewed by external reviewers. A consensus meeting between the two section editors and the editorial team was organized to finally conclude on the selection of best papers.

**Results:** Among the 741 returned papers published in 2017 in the various areas of CRI, the full review process selected five best papers. The first best paper reports on the implementation of consent management considering patient preferences for the use of de-identified data of electronic health records for research. The second best paper describes an approach using natural language processing to extract symptoms of severe mental illness from clinical text. The authors of the third best paper describe the challenges and lessons learned when leveraging the EHR4CR platform to support patient inclusion in academic studies in the

context of an important collaboration between private industry and public health institutions. The fourth best paper describes a method and an interactive tool for case-crossover analyses of electronic medical records for patient safety. The last best paper proposes a new method for bias reduction in association studies using electronic health records data.

**Conclusions:** Research in the CRI field continues to accelerate and to mature, leading to tools and platforms deployed at national or international scales with encouraging results. Beyond securing these new platforms for exploiting large-scale health data, another major challenge is the limitation of biases related to the use of “real-world” data. Controlling these biases is a prerequisite for the development of learning health systems.

## Keywords

International Medical Informatics Association Yearbook; Clinical Research Informatics; Biomedical Research, Clinical Trials as Topic; Observational studies as Topic; Real-world data; Phenotyping

Yearb Med Inform 2018;177-83

<http://dx.doi.org/10.1055/s-0038-1641220>

via PubMed (from NCBI, National Center for Biotechnology Information) with a set of predefined MeSH descriptors and free terms: Clinical research informatics, Biomedical research, Nursing research, Clinical research, Medical research, Pharmacovigilance, Patient selection, Phenotyping, Genotype-phenotype associations, Feasibility studies, Eligibility criteria, Feasibility criteria, Cohort selection, Patient recruitment, Clinical trial eligibility screening, Eligibility determination, Patient-trial matching, Protocol feasibility, Real world evidence, Data Collection, Epidemiologic research design, Clinical studies as Topic, Multicenter studies as Topic, and Evaluation studies as Topic. Papers addressing topics of other sections of the Yearbook, such as Bioinformatics, were excluded based on the predefined exclusion of MeSH descriptors such as Genetic research, Gene ontology, Human genome project, Stem cell research, or Molecular epidemiology.

Bibliographic databases were searched on January 30, 2018 for papers published in 2017, considering the electronic publication date. Among an original set of 741 references, 160 papers were selected as being in the scope of CRI and their scientific quality was blindly rated as low, medium, or high by the two section editors based on papers' title and abstract. Seventy-two references classified as medium or high quality contributions to the field by at least one of the section editors were classified into the following eleven dimensions/sub areas of the CRI domain: *observational studies, reuse of electronic health record (EHR) data, feasibility studies, patient recruitment, data*

## Introduction

Within the 2018 International Medical Informatics Association (IMIA) Yearbook, the goal of the Clinical Research Informatics section is to provide an overview of research trends from 2017 publications that demonstrate excellent research about multifaceted aspects of medical informatics supporting clinical trials and observational studies. Clinical Research Informatics (CRI) continues to be developed and the CRI community has especially to address the important challenges of sharing health data with the best balance

“Between Access and Privacy” - this year's special theme for the IMIA Yearbook. New methods, tools, and CRI systems have been developed in order to collect, integrate and mine healthcare data for better care.

## About the Paper Selection

A comprehensive review of articles published in 2017 and addressing a wide range of issues for CRI was conducted. The selection was performed by querying MEDLINE

*management and CRI systems, data/text mining and algorithms, data quality assessment or validation, security and confidentiality, ethical, legal, social, and policy issues and solutions, stakeholder participation, data integration and semantic interoperability, communicating study results.* The 72 references were reviewed jointly by the two section editors to select a consensual list of 13 candidate best papers representative of all CRI categories. Following the IMIA Yearbook process, these 13 papers were peer-reviewed by the IMIA Yearbook editors and external reviewers (at least four reviewers per paper). Five papers were finally selected as best papers (Table 1). A content summary of these best papers can be found in the appendix of this synopsis.

## Outlook and Conclusion

### Ethical, Legal, and Social Issues of Observational Studies and Reuse of EHR Data

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. The paper from Kim et al., selected as a best paper, reports the implementation of patient e-consent for the use of de-identified data of EHRs for research and demonstrates that considering patient preferences increases satisfaction and does not significantly affect participation in research [1].

### Data/Text Mining and New Technologies from Artificial Intelligence

Except for demographics, diagnoses, acts, and laboratory results, the major part of EHR data is unstructured, especially symptoms. There is a lack of instruments allowing to efficiently reuse symptomatology for research purposes. The paper from Jackson et al., selected as a best paper, describes an approach for using natural language processing to extract symptoms of severe mental illness from clinical text [2]. The authors demonstrate the performance of an information extraction approach from discharge summaries of routine mental health records from a database of 1.2 million residents in south London (UK) reporting an average F1-score of 0.88 of automatic extraction of

46 symptoms. Lucini et al., also used a text mining approach for optimizing the use of hospital resources in the context of emergency department overcrowding [3]. The authors report an average F1-score of 77.70% in predicting future hospitalizations from early information on short-term inward bed demand for patients receiving care at the emergency department. Text mining is also used in the field of adverse events identification from safety reports. For example, Marella et al., used a machine learning approach to retrieve technology-related and EHR-related adverse events from databases of legacy free-text safety reports and to distinguish them from reports in which the EHR was mentioned only in passing [4]. Of the four tested algorithms, a naive Bayes kernel performed best (AUC of  $0.927 \pm 0.023$  and F-score of  $0.877 \pm 0.027$ ) and was used to develop a semi-automated approach to screening cases from a mandatory, population-based, patient safety reporting system.

### Data Integration and Semantic Interoperability

The paper from Girardeau et al., selected as a best paper, describes the challenges and lessons learnt in formalising eligibility criteria to enable EHR querying using the EHR4CR (Electronic Health Records for Clinical Research)<sup>1</sup> platform in the context of an important collaboration between the private industry and public health institutions [5]. The authors defined metrics to assess the different steps of the formalisation of eligibility criteria and reported that the formal representation of 64.2% of a total of 67 computable inclusion and exclusion criteria was considered by experts to be satisfactory or higher. Alonso-Calvo et al., present a standards-based approach to ensure semantic integration of data to be used in clinico-genomic clinical trials [6]. This approach has been adopted in national and international research initiatives, such as the EURECA-EU research project<sup>2</sup>. The Knowledge Base workgroup of the Observational Health Data Sciences and Infor-

**Table 1** Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2018 in the section 'Clinical Research Informatics'. The articles are listed in alphabetical order of the first author's surname.

Section
Clinical Research Informatics
<ul style="list-style-type: none"> <li>Caron A, Chazard E, Muller J, Perichon R, Ferret L, Koutkias V, Beuscart R, Beuscart JB, Fichet G. IT-CARES: an interactive tool for case-crossover analyses of electronic medical records for patient safety. <i>J Am Med Inform Assoc</i> 2017;24(2):323-30.</li> <li>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. <i>BMC Med Res Methodol</i> 2017;17(1):36.</li> <li>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. <i>J Am Med Inform Assoc</i> 2017 Dec 1.</li> <li>Jackson RG, Patel R, Jayatilake N, Koliakou 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. <i>BMJ Open</i> 2017;7(1):e012012.</li> <li>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. <i>J Am Med Inform Assoc</i> 2017;24(2):380-7.</li> </ul>

<sup>1</sup> <http://www.ehr4cr.eu/>

<sup>2</sup> <http://www.eurorec.org/RD/eureca.cfm>

matics (OHDSI) collaboration describes the use of the Linked Data paradigm to facilitate the systematic and scalable integration of relevant evidence sources within LAERTES, an open scalable system operating within the larger software stack provided by the OHDSI clinical research framework<sup>3</sup> [7].

## Data Management and CRI Systems

Although research with structured EHRs is expanding, data science methodology enabling the rapid search/extraction, cleaning, and analysis of these large, often complex, datasets is less well developed. rEHR is an R package developed for manipulating and analyzing EHR data that has been tested with one of the largest primary care EHRs databases, the Clinical Practice Research Datalink (CPRD)<sup>4</sup> [8]. The paper from Caron et al., selected as a best paper, describes IT-CARES, a method and interactive tool for case-crossover analyses of EHRs for patient safety. More and more hospitals report their efforts to unlock clinical data for research and to demonstrate the validity of EHR-based research [9]. As an example, the Georges Pompidou University Hospital (Paris, France) reports its 8-year follow-up experience of development and use of clinical data warehouses (CDWs) enabling 74 research projects [10]. At a larger scale, the national CALIBER research platform<sup>5</sup> links major sources of EHR data across primary and secondary care in UK, and 33 studies were conducted in order to drive innovation and improve efficiency and quality of care in the cardiovascular domain [11].

## Security

There is research activity to develop reliable automated systems to de-identify patient notes in EHRs. Dernoncourt et al., used artificial neural networks to remove the 18 types of protected health information defined by the Health Insurance Portability and Ac-

countability Act (HIPAA) from notes and they report a better performance than previously published systems while requiring no manual feature engineering [12].

## Data Quality and Reproducibility in Biomedical Research

Ongoing works across several domains of science and policy currently aim at clarifying the reproducibility in biomedical research and to enhance the transparency and accessibility of research results. RepeAT is an assessment tool operationalizing key concepts of research transparency in the biomedical domain, specifically for secondary biomedical data research using EHR data [13]. RepeAT includes 119 unique variables grouped into five categories (research design and aim, database and data collection methods, data mining and data cleaning, data analysis, data sharing and documentation). The aim is to facilitate comparisons of research transparency and accessibility across domains and institutions. In the era of Big Data, the interest of large linked datasets is increasing. Linkage of large data sources often relies on probabilistic methods using a set of common identifiers (e.g. sex, date of birth, postcode). Variation in data quality at an individual or organisational level (e.g. hospitals) can result in the clustering of identifier errors, and potentially introduces a selection bias to the resulting linked dataset. Harron et al., measured variation in identifier error rates in a large English administrative data source, incorporated this information into match weight calculation, and demonstrated that attribute- and organisational-specific weights reduced the selection bias as compared with weights estimated using traditional probabilistic matching algorithms [14].

The paper from Huang et al., selected as a best paper, describes a method for reducing biases in the estimation of associations caused by imperfect phenotyping in EHR-derived data by incorporating prior information through integrated likelihood estimation [15]. The authors evaluated the proposed method on real EHR-derived data on diabetes from Kaiser Permanente, Washington, and they showed that the estimated associations using their method were

very close to the estimates from the gold standard method and they reduced biases by 60%-100% as compared to the two methods commonly used in current practice.

In conclusion, research in the CRI field continues to accelerate and to mature leading to tools and platforms deployed at national or international scales with encouraging promising results. Beyond securing these new platforms for exploiting large-scale health data, another major challenge is to control the biases related to the use of “real-world” data - a prerequisite for the development of learning health systems.

## Acknowledgement

We would like to acknowledge the support of Martina Hutter and the reviewers in the selection process of the IMIA Yearbook.

## References

1. Kim H, Bell E, Kim J, Sitapati A, Ramsdell J, Farcas C, et al. iCONCUR: informed consent for clinical data and bio-sample use for research. *J Am Med Inform Assoc* 2017;24(2):380-7.
2. Jackson RG, Patel R, Jayatilake N, Kolliakou A, Ball M, Gorrell G, et al. 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.
3. Lucini FR, S Fogliatto F, C da Silveira GJ, L Neyeloff J, Anzanello MJ, de S Kuchenbecker R, et al. Text mining approach to predict hospital admissions using early medical records from the emergency department. *Int J Med Inf* 2017;100:1-8.
4. Marella WM, Sparnon E, Finley E. Screening Electronic Health Record-Related Patient Safety Reports Using Machine Learning. *J Patient Saf* 2017;13(1):31-6.
5. Girardeau Y, Doods J, Zapletal E, Chatellier G, Daniel C, Burgun A, et al. Leveraging the EHR4CR platform to support patient inclusion in academic studies: challenges and lessons learned. *BMC Med Res Methodol* 2017;17(1):36.
6. Alonso-Calvo R, Paraiso-Medina S, Perez-Rey D, Alonso-Oset E, van Stiphout R, Yu S, et al. A semantic interoperability approach to support integration of gene expression and clinical data in breast cancer. *Comput Biol Med* 2017;87:179-86.
7. Knowledge Base workgroup of the Observational Health Data Sciences and Informatics (OHDSI) collaborative. Large-scale adverse effects related to treatment evidence standardization (LAERTES): an open scalable system for linking pharmacovigilance evidence sources with clinical data. *J Biomed Semant* 2017;8(1):11.

<sup>3</sup> <https://github.com/OHDSI/KnowledgeBase/tree/master/LAERTES>

<sup>4</sup> <https://www.cprd.com/home/>

<sup>5</sup> <https://www.ucl.ac.uk/health-informatics/caliber>

8. Springate DA, Parisi R, Olier I, Reeves D, Kontopantelis E. rEHR: An R package for manipulating and analysing Electronic Health Record data. *PloS One* 2017;12(2):e0171784.
9. Caron A, Chazard E, Muller J, Perichon R, Ferret L, Koutkias V, et al. 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.
10. Jannot A-S, Zapletal E, Avillach P, Mamzer M-F, Burgun A, Degoulet P. The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience. *Int J Med Inform* 2017;102:21-8.
11. Hemingway H, Feder GS, Fitzpatrick NK, Denaxas S, Shah AD, Timmis AD. Using nationwide 'big data' from linked electronic health records to help improve outcomes in cardiovascular diseases: 33 studies using methods from epidemiology, informatics, economics and social science in the ClinicAl disease research using Linked Bespoke studies and Electronic health Records (CALIBER) programme [Internet]. Southampton (UK): NIHR Journals Library; 2017 [cité 10 juin 2018]. (Programme Grants for Applied Research). Accessible at: <http://www.ncbi.nlm.nih.gov/books/NBK414778/>
12. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. *J Am Med Inform Assoc* 2017;24(3):596-606.
13. McIntosh LD, Juehne A, Vitale CRH, Liu X, Alcoser R, Lukas JC, et al. Repeat: a framework to assess empirical reproducibility in biomedical research. *BMC Med Res Methodol* 2017;17(1):143.
14. Harron K, Hagger-Johnson G, Gilbert R, Goldstein H. Utilising identifier error variation in linkage of large administrative data sources. *BMC Med Res Methodol* 2017;17(1):23.
15. Huang J, Duan R, Hubbard RA, Wu Y, Moore JH, Xu H, et al. 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.

**Correspondence to:**

Christel Daniel, MD, PhD

WIND DSI - Assistance Publique - Hôpitaux de Paris

5 rue Santerre - 75 012 PARIS

France

Tel: +33 1 48 04 20 29

E-mail: [christel.daniel@aphp.fr](mailto:christel.daniel@aphp.fr)



## 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-cross-over 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* 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 avail-

ability 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, Jayatilake 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.