Key Contributions in Clinical Research Informatics

Christel Daniel[1, 2], Ali Bellamine[1], Dipak Kalra[3], Section Editors of the IMIA Yearbook Section on Clinical Research Informatics

[1] Information Technology Department, AP-HP, F-75012 Paris, France
[2] Sorbonne University, University Paris 13, Sorbonne Paris Cité, INSERM UMR_S 1142, LIMICS, F-75006 Paris, France
[3] The University of Gent, 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 2020.

Method: A bibliographic search using a combination of Medical Subject Headings (MeSH) descriptors and free-text 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. After peer-review ranking, a consensus meeting between two section editors and the editorial team was organized to finally conclude on the selected four best papers.

Results: Among the 877 papers published in 2020 and returned by the search, there were four best papers selected. The first best paper describes a method for mining temporal sequences from clinical documents to infer disease trajectories and enhancing high-throughput phenotyping. The authors of the second best paper demonstrate that the generation of synthetic Electronic Health Record (EHR) data through Generative Adversarial Networks (GANs) could be substantially improved by more appropriate training and evaluation criteria. The third best paper offers an efficient advance on methods to detect adverse drug events by computer-assisting expert reviewers with annotated candidate mentions in clinical documents. The large-scale data quality assessment study reported by the fourth best paper has clinical research informatics implications, in terms of the trustworthiness of inferences made from analysing electronic health records.

Conclusions: The most significant research efforts in the CRI field are currently focusing on data science with active research in the development and evaluation of Artificial Intelligence/Machine Learning (AI/ML) algorithms based on ever more intensive use of real-world data and especially EHR real or synthetic data. A major lesson that the coronavirus disease 2019 (COVID-19) pandemic has already taught the scientific CRI community is that timely international high-quality data-sharing and collaborative data analysis is absolutely vital to inform policy decisions.

Keywords

International Medical Informatics Association Yearbook, clinical research informatics, biomedical research, clinical trials as topic, real-world evidence generation, phenotyping

Yearb Med Inform 2021:233-8
http://dx.doi.org/10.1055/s-0041-1726514

1 Introduction

Within the 2020 International Medical Informatics Association (IMIA) Yearbook, the goal of the Clinical Research Informatics (CRI) section is to provide an overview of research trends from 2020 publications that demonstrate the progress in multifaceted aspects of medical informatics supporting research and innovation in the healthcare domain. New methods, tools, and CRI systems have been developed in order to enable real-world evidence generation and optimize the life-cycle of clinical trials. The CRI community has also addressed the important challenges of enabling the huge clinical research efforts to be deployed to inform doctors, epidemiologists and the public about coronavirus disease 2019 (COVID-19) patients and contributed to "Informatics in public health and pandemics" - this year's special theme for the IMIA Yearbook.

2 Paper Selection Method

A comprehensive review of articles published in 2020 and addressing a wide range of issues for CRI was conducted. The selection was performed by querying MEDLINE via PubMed (from National Center for Biotechnology Information (NCBI)) with a set of predefined Medical Subject Heading (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 Translational 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 9th, 2021 for papers published in 2020, considering the electronic publication date. Among an original set of 877 references, 401 papers were selected as being in the scope of CRI and their scientific quality was blindly rated as low, medium, or high by the section editors based on papers' title and abstract. Sixty-one references classified as high-quality contributions to the field by at least two of the three section editors were considered. These 61 papers were classified into the following eleven areas of the CRI domain (multiple classification choices were permitted):

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2021 in the section 'Clinical Research Informatics'. The articles are listed in alphabetical order of the first author's surname.

Section
Clinical Research Informatics
▪ Bell SK, Delbanco T, Elmore JG, Fitzgerald PS, Fossa A, Harcourt K, Leveille SG, Payne TH, Stametz RA, Walker J, DesRoches CM. Frequency and types of patient-reported errors in electronic health record ambulatory care notes. JAMA Netw Open 2020 Jun 1;3(6):e205867.
▪ Estiri H, Strasser ZH, Murphy SN. High-throughput phenotyping with temporal sequences. J Am Med Inform Assoc 2021 Mar 18;28(4):772–81.
▪ Geva A, Stedman JP, Manzi SF, Lin C, Savova GK, Avillach P, Mandl KD. Adverse drug event presentation and tracking (ADEPT): semiautomated, high throughput pharmacovigilance using real-world data. JAMIA Open 2020 Oct;3(3):413–21.
▪ Zhang Z, Yan C, Mesa DA, Sun J, Malin BA. Ensuring electronic medical record simulation through better training, modeling, and evaluation. J Am Med Inform Assoc 2020 Jan 1;27(1):99–108.

Reuse of Electronic Healthcare Records (EHRs), Learning Healthcare System (LHS) data (n=22); Big data management, data integration, semantic interoperability and data quality assessment (n=14); Data science (data/text mining, Artificial Intelligence (IA), Machine Learning (ML)) (n=32); Security and data privacy (n=3); Feasibility studies, patient recruitment, improved user experiences of CRI systems (n=14) and Governance (ethical, regulatory, societal, policy issues, stakeholder participation, research networks, team science) (n=18).

The 61 references were reviewed jointly by the section editors to select a consensual list of 15 candidate best papers representative of all CRI categories. In conformance with the IMIA Yearbook process, these 15 papers were peer-reviewed by the IMIA Yearbook editors (the two section editors, and two editors in chief), and external reviewers. Four 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.

Conclusions and Outlook

The 15 candidate best papers for 2021 illustrate recent efforts and trends in different CRI areas such as real-world evidence generation; Data integration, semantic interoperability and data quality assessment; security, confidentiality and data privacy; data/text mining, Artificial Intelligence (AI) and Machine Learning (ML); feasibility studies, patient recruitment, data management and CRI systems; ethical, legal, social, policy issues and solutions, stakeholder participation.

3.1 Real-world Evidence Generation, Electronic Phenotyping

Although EHRs, often linked to biorepositories, are seen as important data sources for translational research, the lack of precise phenotype labels still limits their use. The traditional approach for electronic phenotyping comprises manually defining inclusion and exclusion criteria based on structured data, such as diagnosis codes, laboratory results, and medications to build cohorts of patients with (or without) a certain phenotype or clinical condition of interest. Although several collaborative networks exist for sharing phenotype definitions, creating rule-based phenotypes in each site is typically labor-intensive, requiring multiple rounds of review by domain experts. Interestingly, Kashyap *et al.* demonstrated the good portability of phenotype algorithms between sites within the USA, using APHRODITE (Automated PHenotype Routine for Observational Definition, Identification, Training and Evaluation), an open-source framework for sharing phenotype classifiers across sites within the OHDSI network [1]. Unfortunately, the portability decreased beyond the USA indicating geographic limitations to generalizability and the authors consequently consider that sharing the classifier-building recipe is still nowadays more useful for facilitating collaborative observational research than sharing the pretrained classifiers. Ahuja *et al.* developed a fully automated multi disease phenotyping algorithm suited toward large-scale highly multiphenotype applications such as phenome-wide association studies [2]. The first best paper of the CRI section, from Estiri *et al.* demonstrates how the use of a vector sequential representation of the EHR data combined with a novel sequential pattern mining (SPM) algorithm to characterize temporal relationships between EHR data instances - in particular diagnoses, complications and medications - significantly improves the performance of a high throughput algorithm to predict 10 phenotypes [3].

3.2 Data Integration, Semantic Interoperability, and Data Quality Assessment

The quality of real-world data has been featured in previous Yearbook editions and we continue to see new studies, methodologies and tools to assess data quality. In the second best paper of the section, Bell *et al.* invited patients registered at three US healthcare providers to access their EHRs online, from home, and to report errors [4]. The important message is that errors frequently do exist in patient records and sometimes can be serious enough to mislead clinical decision-making and the accuracy of real-world evidence. Building on the literature on data quality dimensions, Bian *et al.* examined over 200 data quality assessment studies within the PCORnet network, to study what dimensions were used and how these were each defined [5]. They observed inconsistency in their definitions, which would make comparisons between assessments unreliable. Utilizing standard tools to assess data quality can help ensure consistency in the way each dimension is applied to the data. Dixon *et al.* have demonstrated, using 100 million clinical messages, how the Observational Health

Data Sciences and Informatics (OHDSI) open source software can be used to assess completeness, timeliness and entropy, as data quality dimensions [6].

3.3 Security, Confidentiality, and Data Privacy

One of the main bottlenecks of data-driven research is privacy protection preventing and delaying broad and sustainable medical data sharing. In addition, to address privacy protection challenges, researchers often make trade-offs on data utility. Raisaro *et al.* describe the efforts of the international consortium for Secure Collective Research (SCOR) to reconciliate privacy/utility conflicts for collaborative data sharing and analysis [7]. Leveraging the latest progress in modern security and federated machine learning algorithms the consortium proposes a ready to-deploy secure infrastructure addressing the urgency of data sharing by reducing administrative and regulatory barriers driven by privacy and security concerns while respecting patient privacy and maximizing data utility. In the third best paper, Zhang *et al.* leveraged a well-established method for generating synthetic EHR data by learning and reproducing patterns from real EHR data using generative adversarial networks (GANs) [8]. They found this can be substantially improved through more appropriate training, modelling, and evaluation criteria. The new GAN generator that is able to learn from smaller training data sets, shows greater capability to incorporate low-prevalence concepts and outperforms the state-of-the-art approaches offering a novel and evaluated advanced method to generate accurate simulated EHR data.

3.4 Data/Text Mining, Artificial Intelligence, and Machine Learning

CRI researchers are progressively and intensively applying modern machine learning (ML) algorithms in the real-world and especially with EHR data. The fourth best paper from Geva *et al*. describes the development of a natural language processing (NLP) pipeline for detecting potential adverse drug events (ADEs) in EHRs combined with a tool for human review and adjudication of true ADEs [9]. The ADE presentation and tracking (ADEPT) solution appears as an efficient advance on the use of real-world data for pharmacovigilance by computer-assisting expert reviewers with annotated candidate mentions in clinical documents.

Efforts to discover clinical knowledge from EHR are hampered by the fact that EHR observations often reflect a complex set of processes (e.g., workflows of the provider and payer organizations) rather than patients' true health state. Machine learning must account for potential biases introduced through the recording process especially when mining the temporal dimension of the disease progression and treatment outcomes. The aim of the paper from Pokharel *et al.* was to address the representation of the temporal aspect of the patient record while computing similarities between patient records to support cohort selection, patients stratification or medical prognosis [10]. Brajer *et al.* not only developed a machine learning model that predicts in-hospital mortality for all adult patients at the time of hospital admission but they also evaluated the algorithm prospectively and carried out an external validation [11]. They demonstrated that machine learning models predicting in-hospital mortality can be implemented on live EHR data with prospective performance if highly curated research data sets are used in the development and external validation of the algorithm. The next challenge is now to understand how to best integrate such models into the clinical workflow, identify opportunities to scale, and quantify the impact on clinical and operational outcomes.

3.5 Feasibility Studies, Patient Recruitment, Data Management, and CRI Systems

There are now many tools and products that can query real-world data to optimize clinical trial protocol designs and recruitment. However, important clinical information is often captured in free text. Liu *et al.* have designed a free text information retrieval method that extracts information from a hospital clinical data warehouse and transfers it into the Observational Medical Outcomes Partnership Common Data Model [12]. This improves the accuracy and completeness of patient matching. A complementary challenge is how researchers can query this information. Dobbins *et al.* have developed a user-friendly graphical data exploration and query tool for researchers to navigate semantic hierarchies and assemble query constructs using a drag and drop interface [13]. This tool generates Structured Query Language (SQL) queries that can then be mapped to any clinical database schema. Once patients have given their consent and are being recruited into a trial a time-consuming and error prone re-capture of patient histories usually occurs. Zong *et al*. have developed a pipeline that utilizes the Health Level 7 - Fast Healthcare Interoperability Resources (HL7 FHIR) Application Programming Interfaces (APIs), which are increasingly being adopted in EHR systems, and shown that a colorectal cancer research dataset can be accurately mapped to the case report forms used at the Mayo Clinic [14].

3.6 Ethical, Legal, Social, Policy Issues and Solutions, Stakeholder Participation, and Research Networks

The CRI community contributed to "Informatics in public health and pandemics" – this year's special theme for the IMIA Yearbook, and the focus of the CRI survey paper in this chapter, which is on COVID-19 and the ethical considerations of data protection. There is an urgent need of additional research in order to build robust and scalable infrastructures with state-of-the-art security, interoperability and data curation technology to enable large-scale automated and unbiased federated data analysis to accelerate scientific discoveries and combat not only existing but also emerging diseases such as the SARS-CoV-2 outbreak and future pandemics. High quality data sets are key assets for predicting prognosis and drug response from phenotypic, genotypic, and epigenetic data through innovative clinical trials and large-scale observational studies. However, in the rapidly growing field of real-world evidence generation, it is crucial to

more critically evaluate EHR-driven studies the veracity of the data used to support the conclusions in order to prevent harm from misleading studies [15, 16].

Acknowledgements
We would like to acknowledge the support of Adrien Ugon, Martina Hutter, Kate Fultz Hollis, Lina Soualmia, Brigitte Séroussi, and the whole Yearbook editorial team as well as the reviewers for their contribution to the selection process of the Clinical Research Informatics section for the IMIA Yearbook 2021.

References

1. Kashyap M, Seneviratne M, Banda JM, Falconer T, Ryu B, Yoo S, et al. Development and validation of phenotype classifiers across multiple sites in the observational health data sciences and informatics network. J Am Med Inform Assoc 2020 Jun 1;27(6):877–83.
2. Ahuja Y, Zhou D, He Z, Sun J, Castro VM, Gainer V, et al. sureLDA: A multidisease automated phenotyping method for the electronic health record. J Am Med Inform Assoc 2020 Aug 1;27(8):1235–43.
3. Estiri H, Strasser ZH, Murphy SN. High-throughput phenotyping with temporal sequences. J Am Med Inform Assoc 2021 Mar 18;28(4):772–81.
4. Bell SK, Delbanco T, Elmore JG, Fitzgerald PS, Fossa A, Harcourt K, et al. Frequency and Types of Patient-Reported Errors in Electronic Health Record Ambulatory Care Notes. JAMA Netw Open 2020 Jun 1;3(6):e205867.
5. Bian J, Lyu T, Loiacono A, Viramontes TM, Lipori G, Guo Y, et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. J Am Med Inform Assoc 2020 Dec 9;27(12):1999–2010.
6. Dixon BE, Wen C, French T, Williams JL, Duke JD, Grannis SJ. Extending an open-source tool to measure data quality: case report on Observational Health Data Science and Informatics (OHDSI). BMJ Health Care Inform 2020 Mar;27(1).
7. Raisaro JL, Marino F, Troncoso-Pastoriza J, Beau-Lejdstrom R, Bellazzi R, Murphy R, et al. SCOR: A secure international informatics infrastructure to investigate COVID-19. J Am Med Inform Assoc 2020 Nov 1;27(11):1721–6.
8. Zhang Z, Yan C, Mesa DA, Sun J, Malin BA. Ensuring electronic medical record simulation through better training, modeling, and evaluation. J Am Med Inform Assoc 2020 Jan 1;27(1):99–108.
9. Geva A, Stedman JP, Manzi SF, Lin C, Savova GK, Avillach P, et al. Adverse drug event presentation and tracking (ADEPT): semiautomated, high throughput pharmacovigilance using real-world data. JAMIA Open 2020 Oct;3(3):413–21.
10. Pokharel S, Zuccon G, Li X, Utomo CP, Li Y. Temporal tree representation for similarity computation between medical patients. Artif Intell Med 2020 Aug;108:101900.
11. Brajer N, Cozzi B, Gao M, Nichols M, Revoir M, Balu S, et al. Prospective and External Evaluation of a Machine Learning Model to Predict In-Hospital Mortality of Adults at Time of Admission. JAMA Netw Open 2020 Feb 5;3(2):e1920733.
12. Liu S, Wang Y, Wen A, Wang L, Hong N, Shen F, et al. Implementation of a Cohort Retrieval System for Clinical Data Repositories Using the Observational Medical Outcomes Partnership Common Data Model: Proof-of-Concept System Validation. JMIR Med Inform 2020 Oct 6;8(10):e17376.
13. Dobbins NJ, Spital CH, Black RA, Morrison JM, de Veer B, Zampino E, et al. Leaf: an open-source, model-agnostic, data-driven web application for cohort discovery and translational biomedical research. J Am Med Inform Assoc 2020 Jan 1;27(1):109–18.
14. Zong N, Wen A, Stone DJ, Sharma DK, Wang C, Yu Y, et al. Developing an FHIR-Based Computational Pipeline for Automatic Population of Case Report Forms for Colorectal Cancer Clinical Trials Using Electronic Health Records. JCO Clin Cancer Inform 2020 Mar;4:201–9.
15. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ 2018 Apr 30;361:k1479.
16. Kohane IS, Aronow BJ, Avillach P, Beaulieu-Jones BK, Bellazzi R, Bradford RL, et al. What Every Reader Should Know About Studies Using Electronic Health Record Data but May be Afraid to Ask. J Med Internet Res 2021 Mar 2;23(3):e22219.

Correspondence to:
Christel Daniel, MD, PhD
Data and Digital Innovation Department, Information Systems Direction – 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

Appendix: Summary of Best Papers Selected for the 2021 Edition of the IMIA Yearbook, CRI Section

Bell SK, Delbanco T, Elmore JG, Fitzgerald PS, Fossa A, Harcourt K, Leveille SG, Payne TH, Stametz RA, Walker J, DesRoches CM

Frequency and types of patient-reported errors in electronic health record ambulatory care notes

JAMA Netw Open 2020 Jun 1;3(6):e205867

The accuracy of electronic health record data matters more than ever, especially due to the proliferation of clinical decision support, workflow systems and learning health systems. Research, usually in general practice, has established that patients can discover errors and advise on the correction that is needed. This paper reports on a major US study involving 22,889 patients who have had access to their hospital or community practice EHRs and read them. They were invited to review their online health data and to respond by means of a questionnaire. (A larger pool of patients were invited, and 21.7% responded.) Of those, around 21% identified an error in their records and 40% of those regarded those errors as serious (one in ten indicated "very serious"). These errors were mostly related to errors in documented diagnoses (incorrect or missing), also in medication, allergies and procedures. The authors note that "Older and sicker patients were twice as likely to report a serious error compared with younger and healthier patients, indicating important safety and quality implications". These errors were rated by patients themselves, which may not be the same as clinician judgement of importance (which is not discussed). The respondents reported mixed reactions from healthcare professionals to these errors, sometimes they were corrected promptly and another times repeated requests still resulted in no rectification of the error. The authors recognise the low response rate and discuss this. There will inevitably be unknown selection biases in terms of health status and social diversity. The significance of this paper is to highlight the relatively high frequency with which health records contain errors, and also the value of patients in helping to correct them.

Estiri H, Strasser ZH, Murphy SN

High-throughput phenotyping with temporal sequences

J Am Med Inform Assoc 2021 Mar 18;28(4):772–81

Personalized medicine research and machine learning require the characterization of patients into stratified subpopulations, accurately enough to allow differentiation of disease course, treatment effectiveness and outcomes. However, clinical documentation is intended primarily to serve continuity of care decision-making and medico-legal record-keeping, and scientific research purposes are not usually the priority. It can therefore be difficult to infer facts like the date when a health condition first arose in a patient or to link a causal series of events such as treatment changes and complications. Most advanced phenotyping initiatives developed for cohort building do not utilize the temporal dimension of the disease progression or treatment outcomes. Estiri *et al.* aimed at utilizing a vector sequential representation of the EHR data combined with a novel sequential pattern mining (SPM) algorithm to characterize temporal relationships between EHR data instances – in particular diagnoses, complications and medications – and to significantly improve the performance of a high throughput feature selection algorithm to predict phenotypes. A representation mining algorithm was first developed to construct from EHR diagnosis, and medication records five classes of feature sets: a baseline representation for computational phenotyping (aggregated vector representation (AVR), temporal sequential representations using two different algorithms (SPM and tSPM) and combined classes (AVR+SPM, AVR+tSPM). A computational phenotyping algorithm was then trained on the five feature sets extracted from the data from the Mass General Brigham Biobank to predict 10 phenotypes and evaluated against the gold standard labels from validated disease cohorts. The results show improved performance across all 10 phenotypes compared to existing classifiers published in the literature. This paper demonstrates that sequencing the diagnoses and medications results in rich feature sets having the capability to enhance the performance of downstream phenotyping algorithms. This new method enables new insights in disease trajectories and help to improve the accuracy of future machine learning and the delivery of personalised medicine.

Geva A, Stedman JP, Manzi SF, Lin C, Savova GK, Avillach P, Mandl KD

Adverse drug event presentation and tracking (ADEPT): semiautomated, high throughput pharmacovigilance using real-world data

JAMIA Open 2020 Oct;3(3):413–21

Pharmacovigilance based on real world data is challenging because clinicians rarely document adverse drug events (ADEs) in a structured form, might not document a symptom as being due to a medication item in free text, and might not recognise a symptom as being attributable to a medicine. Extracting this information in a fully computable form by natural language processing (NLP) is not usually accurate enough. In this research, Geva *et al.* report on an optimised workflow for computer assisted annotation of the indicators of ADE in EHR data. The authors have developed a methodology and user-facing tool (ADEPT – source code available) to offer experts a visualisation of candidate occurrences of an adverse drug event within clinical narratives, with annotation tools to facilitate rapid decision making on each presented candidate match. Extracted concepts were mapped to the Unified Medical Language System Concept Unique Identifier (UMLS CUI), and colour coded by CUI class to assist with visual interpretation. The authors' method incorporated two independent reviewers and an adjudication user for cases of non-agreement. ADEPT was validated by searching for occurrences of seizure as an ADE (and not as a co-morbidity or unrelated event) while taking sildenafil in 416 patients. Using NLP, 72 candidate mentions were identified, and screened by the expert reviewers who were on average able to arrive at a decision in

less than four minutes per patient, although only nine seconds per document per reviewer before adjudication. This is substantially less time than existing methods and would appear to be flexibly extensible to other drugs and possible ADEs. This research has been selected as a best paper because it offers an efficient advance on methods to detect ADEs by computer-assisting expert reviewers with annotated candidate mentions in clinical documents.

Zhang Z, Yan C, Mesa DA, Sun J, Malin BA

Ensuring electronic medical record simulation through better training, modeling, and evaluation

J Am Med Inform Assoc 2020 Jan 1;27(1):99–108

There is rapidly expanding interest for re-using health data for the spectrum of learning health system purposes including quality improvement, public health screening and interventions and various kinds of research and innovation. Valid inferences from clinical data require relatively precise fine-grained information, which poses challenges for data protection, since most of these re-uses are not occurring on the basis of informed consent but on the basis that the data are considered de-identified. There have been decades of research into anonymization methods, and ways of establishing whether a dataset is sufficiently de-identified that individuals cannot be recognized within it. These methods require a difficult balance to be found between scientific utility and data protection. Various approaches have been developed to mitigate risk, including record simulation via a well-established method for generating synthetic EHR generative adversarial networks (GANs). These have the ability to generate realistic synthetic data from real records but with the loss of certain statistical properties of the real data. The objective of Zhang *et al.* was to enhance the learning model of GANs for generating diagnoses and procedure codes and evaluate the resulting pipeline on real EHRs, based on new evaluation measures. The new GAN generator developed in this work is able to learn from smaller training data sets and with greater capability to incorporate low-prevalence concepts, utilising Wasserstein divergence. The method includes cycles that compare the generated data with real EHR data to verify their similarity whilst verifying the preservation of privacy. Two evaluation measures were designed to compare the utility and privacy of the new and existing GANs for generating categorial data, using a large billing code data set of 1 million real EHRs at Vanderbilt University Medical Center. The proposed model outperformed the state-of-the-art approaches with significant improvement without sacrificing the privacy provided by such models. This best paper shows that EHR data simulation through GANs can be substantially improved. The limitation of the method is to generate only categorial data and in a static manner.