Published in final edited form as:

J Pediatr Gastroenterol Nutr. 2018 October ; 67(4): 434–435. doi:10.1097/MPG.0000000000002021.

Electronic Medical Records Enable Precision Medicine Approaches for Celiac Disease

Idit Kosti^{1,2} and Marina Sirota^{1,2}

¹Institute for Computational Health Sciences, University of California, San Francisco, CA 94158, USA

²Department of Pediatrics, University of California, San Francisco, CA 94158, USA

Precision medicine is an emerging integrative approach for disease prevention, early detection and treatment that takes into account individual variability in genetics, medical history, environmental exposures, and lifestyle. The electronic health record (EHR) captures extensive information on patients over time from all aspects of medical care. The captured structured data can be clinical diagnosis in the form of International Classification of Disease-9/10 (ICD-9/10), measurements collected from the patients in the form of vital signs, laboratory tests (urine, blood, and swabs), imaging sessions, encounters with the patients by medical staff, medication orders and prescriptions. Some of the data captured, such as doctors' notes, is either semi or un-structured. EHR provides a new source of data for biomedical research, allowing the researchers to ask new questions about care standards, disease diagnostics and improve disease therapeutics. An example for a suggested change in care standards is a study by Rappoport et al. [1]. In this paper, the authors show how a data-driven approach on healthy individuals finds that intervals used in to determine abnormalities of laboratory tests results are different between male and females for some laboratory tests. An example of leveraging EMR to improve predictive disease diagnostics is shown in a recent study by Miotto et al. [2]. Here, the authors apply a deep learning framework to pre-process raw EHR data to improve performance when predicting future disease. Finally, an example showing improved therapeutics was recently published by Boland et al. [3] where the authors categorize and identify harmful drugs from medication group C (classified as "risk not ruled out" for pregnant patients) using a machine learning approach, and by that shed light on which drugs from this group shouldn't be used during pregnancy. Access to EHR also allows researchers to identify and define precise subsets of patients to study further and apply precision medicine strategies for disease diagnostics and therapeutics.

Celiac disease (CD) is an immune-mediated condition triggered in genetically susceptible individuals by the ingestion of gluten containing grains and associated with human leukocyte antigen (HLA) DQ2 and DQ8 haplotypes [4]. In the pediatric population, CD ranges from classic CD, diagnosed in at the early age of 6–24 months and non-classic CD that is usually diagnosed at a later age [5]. Better understanding of the disease and co-occurring conditions is crucial for better diagnosis and treatment. In the past, there were

several attempts to study CD using EHR. CD is especially challenging in the EHR context since the value of ICD-9 code related to the disease has been shown in the past to poorly identify patients with CD [6]. A recent paper described the co-occurrence of CD and autoimmune diseases using EHR [7]. Prinzbach et al. are leveraging EHR to analyze associations of phenotypes and to test known comorbidities with celiac disease in children. They furthermore use this approach to discover new associations of ICD-10 codes with CD. They identify 45 ICD-10 codes that are significantly associated with CD, among them 13 are known comorbidities including autoimmune conditions such as Crohn's Disease and Type 1 Diabetes and nine are expected symptoms of CD such as abdominal and pelvic pain and nausea and vomiting. Novel ICD-10 include personal traits, other conditions such as eosinophilic esophagitis, abnormal immunological and serum enzyme levels. Although deeper investigation is required to verify and validate these findings, this paper presents the first step towards leveraging electronic health records to study CD.

With the complexity and sparsity of EHR data, there is a huge need for new computational methods and approaches. Taking longitudinal aspect of the data into account and developing patient trajectories is one obvious next step. Applying innovative machine learning and clustering approaches to classify and stratify patients will allow us to study disease heterogeneity in the context of treatment response and outcomes. Furthermore, leveraging EHR data repositories and integrating clinical phenotyping data together with molecular measurements provides an unprecedented opportunity to advance the precision medicine in the context of disease diagnostics and therapeutics and will be key in informing a range of clinical and translational precision medicine research. Phenome Wide Association Study (PheWAS) [8] is a method of combining of genomic and EHR data. In the past, PheWAS led to validation of known genotype-phenotype associations and better definitions of disease subtypes. PheWAS method on based on genomic and EHR data has shown an association between HLA haplotype DQB1*02:01-DQA1*05:01 and CD [9]. The integration of genomic data with EHR data has been applied extensively in projects such as eMERGE [10] and Geisinger MyCode project [11], which was able to leverage whole exome sequencing with EHR data to confirm existing and identify new associations with metabolic traits. [9] There is no doubt that in the near future, efforts to gather EHR data, combine it with other molecular measurements and advanced computational approaches will improve the quality and accuracy of patient care, and lead to new discoveries in precision medicine.

Acknowledgments

Funding Sources: MS is supported by K01LM012381 and 1P30AR070155-01.

References

1. Rappoport N, Paik H, Oskotsky B, Tor R, Ziv E, Zaitlen N, et al. Creating ethnicity-specific reference intervals for lab tests from EHR data. *bioRxiv*. 2017; :213892.doi: 10.1101/213892
2. 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:srep26094.doi: 10.1038/srep26094
3. Boland MR, Polubriaginof F, Tatonetti NP. Development of A Machine Learning Algorithm to Classify Drugs Of Unknown Fetal Effect. *Sci Rep*. 2017; 7:12839.doi: 10.1038/s41598-017-12943-x [PubMed: 28993650]

4. Fasano A, Berti I, Gerarduzzi T, Not T, Colletti RB, Drago S, et al. Prevalence of Celiac Disease in At-Risk and Not-At-Risk Groups in the United States: A Large Multicenter Study. *Arch Intern Med*. 2003; 163:286–292. DOI: 10.1001/archinte.163.3.286 [PubMed: 12578508]
5. Fasano A. Clinical presentation of celiac disease in the pediatric population. *Gastroenterology*. 2005; 128:S68–73. [PubMed: 15825129]
6. Tanpowpong P, Broder-Fingert S, Obuch JC, Rahni DO, Katz AJ, Leffler DA, et al. Multicenter study on the value of ICD-9-CM codes for case identification of celiac disease. *Ann Epidemiol*. 2013; 23:136–142. DOI: 10.1016/j.annepidem.2012.12.009 [PubMed: 23313264]
7. Escudié J-B, Rance B, Malamut G, Khater S, Burgun A, Cellier C, et al. A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease. *BMC Med Inform Decis Mak*. 2017; 17:140.doi: 10.1186/s12911-017-0537-y [PubMed: 28962565]
8. Denny JC, Bastarache L, Roden DM. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. *Annu Rev Genomics Hum Genet*. 2016; 17:353–373. DOI: 10.1146/annurev-genom-090314-024956 [PubMed: 27147087]
9. Karnes JH, Bastarache L, Shaffer CM, Gaudieri S, Xu Y, Glazer AM, et al. Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. *Sci Transl Med*. 2017; 9:eaai8708.doi: 10.1126/scitranslmed.aai8708 [PubMed: 28490672]
10. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet Med*. 2013; 15:761–771. DOI: 10.1038/gim.2013.72 [PubMed: 23743551]
11. Carey DJ, Fetterolf SN, Davis FD, Faucett WA, Kirchner HL, Mirshahi U, et al. The Geisinger MyCode community health initiative: an electronic health record–linked biobank for precision medicine research. *Genet Med*. 2016; 18:906–913. DOI: 10.1038/gim.2015.187 [PubMed: 26866580]