CellPress
REVIEWS

Forum

Electronic Health Records Are the Next Frontier for the Genetics of Substance Use Disorders

Sandra Sanchez-Roige[1,*] and Abraham A. Palmer[1,2]

Compared with other psychiatric disorders of similar heritabilities, the progress of substance use disorders (SUD) genetics has been slow. With the growing availability of large-scale biobanks with extensive phenotypes from electronic health records (EHR) and genotypes across millions of individuals, this platform is the next tool to accelerate SUD genetics research.

Although SUDs and depression have similar heritabilities, genome-wide association studies (GWAS) of depression have been more productive than GWAS of SUDs. For smoking, both nicotine dependence and the number of cigarettes per day have been extensively studied, yielding a number of findings, most notably a cluster of nicotinic receptor subunits [1]. For alcohol dependence, alcohol metabolizing enzymes (*ADH1B*, *ALDH2*) have been identified [1]. Less clinically focused phenotypes, such as alcohol consumption and tobacco and cannabis initiation, have also been productive but those phenotypes are so general that they provide only limited utility for developing better strategies for treatment and prevention of SUDs. The success of depression GWAS required very large sample sizes. Thus, it is critically important to identify the most efficient and economical approach to increasing sample size for SUDs GWAS. Whereas it would be possible to continue aggregating and meta-analyzing SUD cohorts, this approach is slow and expensive and will inevitably introduce heterogeneity because of differences in phenotype definition and ascertainment.

Genetic studies can ascertain subjects either by identifying individuals with SUD phenotypes and then genotyping them, or identifying genotyped cohorts in which phenotype information is available. The latter approach can utilize genotyped cohorts for whom EHR are also available; these include the Electronic Medical Records and Genomics (eMERGE) Network and the Million Veterans Program, and population-based cohorts such as UKBiobank and Generation Scotland. Even larger projects, particularly the All of Us Research Program, are underway. SUD information is available in all of these cohorts.

Regrettably, EHR are not designed for genetic research. Instead, they contain extensive longitudinal data that are the by-product of routine clinical care. EHR contain a variety of data types, including structured data from billing, laboratory test results, and unstructured data from physician notes. Thus, the analysis of EHR represents a 'Big Data' problem and has the potential to support more sophisticated phenotyping.

EHR-based phenotyping has already been used for genetic analyses of other medical phenotypes. Psychiatric phenotypes have proved to be more challenging because they rely on symptoms [2], not objective, quantifiable, laboratory results. EHR-phenotyping for psychiatric disorders will likely benefit from multiple sources of diagnostic data and the use of natural language processing (NLP) techniques to parse unstructured data contained in the physician's notes. Members of PsycheMERGE, a recent initiative from the eMERGE network to validate algorithms for psychiatric phenotypes, have recently developed EHR-phenotypes ('algorithms') for bipolar disorder [3], which used a combination of diagnostic codes and clinical notes to produce phenotypes that were equivalent to semi-structured interviews by qualified clinicians and showed similar heritability as traditionally ascertained cases [4].

The field of EHR offers tremendous opportunities for SUDs genetics; however, to realize this potential, EHR-based phenotypes for SUDs will need to undergo rigorous validation. There are existing algorithms that have already been validated for SUDs (alcohol use disorders (e.g., [5]), nicotine use disorders (e.g., [6]), cannabis use [7], and nonmedical opioid use (e.g., [8])). Although the common predictors included in each algorithm vary for each drug class, these generally include claim or billing codes. Billing data often contains information on indirect indicators of drug use, including: diagnosis [i.e., International Classification of Diseases (ICD) codes of abuse or dependence], record of counseling visits, and, especially for nonmedical opioid abuse, prescription medication, pharmacy shopping, multiple prescribers, total day's supply, and number of prescriptions dispensed, which altogether could be used as proxy measures for SUDs. Nonetheless, although using claim data can be enough to identify individuals with SUDs with high specificity (the proportion of true cases detected; i.e., ICD code yes/no) and sensitivity (the proportion of true controls detected), in the absence of other reliable means, claims-based algorithms should be used with caution (or followed-up in longitudinal studies). For example, in one study, claim-based algorithms identified smokers with high specificity but limited sensitivity [9], potentially misclassifying some individuals that have not yet been diagnosed. Similarly, nonmedical opioid use represents a unique challenge, since it is

seldom explicitly recorded as such. Thus, in addition to using diagnoses of abuse or dependence, definitions could benefit from inclusion of additional information, such as other mental health conditions, chronic pain, and hepatitis C. Augmenting structured data with unstructured text may also help to achieve the best possible performance [10]. NLP can capture more granular phenotypes (e.g., patterns of drug use or severity of use) and serve to 'confirm' a billing code via clinical documentation or detect exclusions (e.g., exclude control subjects with a family history of SUDs).

A previous study of ten medical diseases found that combining billing codes, clinical notes, and medication provided superior phenotypic performance [11]. The same approach can be used for SUDs. For example, alcohol use disorders were most commonly identified using progress reports or correspondence notes [12]. Similarly, for nonmedical opioid use, including information from the physician's notes improved performance [11]. Machine learning techniques, such as concept-extraction and deep learning, will be essential tools in this area. When available, inclusion of screening questionnaires, such as the Alcohol Use Disorder Identification Test, and structured physical health assessment (i.e., quantity and frequency of consumption) may also be helpful.

Finally, EHR have the potential to identify biologically meaningful subgroups, which may enhance the power of GWAS. There are multiple data-driven approaches for parsing phenotypic complexity, including clustering and principal component analyses, which can extract phenotypic signatures ('clusters' or subphenotypes) from a heterogeneous mixture of clinical syndromes. This offers an unprecedented opportunity to accelerate pharmacogenetic research by extracting treatment outcomes and drug response data, and test the association to general and specific patient characteristics. Furthermore, we may identify clusters or subphenotypes that may be associated with each drug of abuse (e.g., different comorbid conditions in each cluster). In addition, EHR has the potential to provide longitudinal insights into the onset and (in some cases) recovery from SUDs.

These new approaches are not without limitations. The stigma of SUD diagnosis may limit available documentation. Furthermore, EHRs are extremely heterogeneous and data about a single patient is often fragmented across different health systems. In addition, individuals may be erroneously classified as controls due to lack of available data.

Despite the challenges, we believe that incorporating EHR phenotyping is the next frontier for genetic studies of SUDs. This approach can vastly increase sample size by fundamentally changing the method of ascertainment. Analyses of EHR have utility beyond GWAS, they will also permit for creation of prediction models that take advantage of the longitudinal structure of the data to allow prediction of risk for SUDs and treatment response. These efforts will require the development of novel techniques by multidisciplinary teams with complementary expertise and access to suitable datasets.

[1]Department of Psychiatry, University of California San Diego, La Jolla, CA, 92093, USA
[2]Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, 92093, USA

*Correspondence:
sanchezroige@ucsd.edu (S. Sanchez-Roige).

https://doi.org/10.1016/j.tig.2019.01.007

References

1. Hancock, D.B. *et al.* (2018) Human genetics of addiction: new insights and future directions. *Curr. Psychiatry Rep.* 20, 8
2. Smoller, J.W. (2017) The use of electronic health records for psychiatric phenotyping and genomics. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 177, 601–612
3. Castro, V.M. *et al.* (2015) Validation of electronic health record phenotyping of bipolar disorder and controls. *Am. J. Psychiatry* 172, 363–372
4. Chen, C.-Y. *et al.* (2018) Genetic validation of bipolar disorder identified by automated phenotyping using electronic health records. *Transl. Psychiatry* 8, 86
5. Chen, E. and Garcia-Webb, M. (2014) An analysis of free-text alcohol use documentation in the electronic health record. *Appl. Clin. Inform.* 5, 402–415
6. Wiley, L.K. *et al.* (2013) ICD-9 tobacco use codes are effective identifiers of smoking status. *J. Am. Med. Inform. Assoc.* 20, 652–658
7. Jackson, R.G. *et al.* (2014) TextHunter – a user friendly tool for extracting generic concepts from free text in clinical research. *AMIA Annu. Symp. Proc.* 2014, 729–738
8. Canan, C. *et al.* (2017) Automatable algorithms to identify nonmedical opioid use using electronic data: a systematic review. *J. Am. Med. Inform. Assoc.* 24, 1204–1210
9. Huo, J. *et al.* (2018) Sensitivity of claims-based algorithms to ascertain smoking status more than doubled with meaningful use. *Value Health* 21, 334–340
10. Liao, K.P. *et al.* (2015) Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ* 350, h1885
11. Wei, W.-Q. *et al.* (2016) Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J. Am. Med. Inform. Assoc.* 23, e20–e27
12. Bell, J. *et al.* (2013) Use of electronic health records in identifying drug and alcohol misuse among psychiatric in-patients. *Psychiatrist* 37, 15–20