A comparative study of different methods for automatic identification of clopidogrel-induced bleedings in electronic health records

Hee-Jin Lee, PhD1, Min Jiang, MS1, Yonghui Wu, PhD1, Christian M. Shaffer, MS2, John H. Cleator, MD2,3, Eitan A. Friedman, MD4, Joshua P. Lewis, MD5, Dan M. Roden, MD2,3, Josh Denny, MD, MS3,6, Hua Xu, PhD1

1University of Texas Health Science Center at Houston, Houston, TX; 2Department of Pharmacology, Vanderbilt University School of Medicine, Nashville, TN; 3Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN; 4Division of Cardiovascular Medicine, Vanderbilt University, Nashville, TN; 5Division of Endocrinology, Diabetes and Nutrition, University of Maryland School of Medicine, Baltimore, MD; 6Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN

Abstract

Electronic health records (EHRs) linked with biobanks have been recognized as valuable data sources for pharmacogenomic studies, which require identification of patients with certain adverse drug reactions (ADRs) from a large population. Since manual chart review is costly and time-consuming, automatic methods to accurately identify patients with ADRs have been called for. In this study, we developed and compared different informatics approaches to identify ADRs from EHRs, using clopidogrel-induced bleeding as our case study. Three different types of methods were investigated: 1) rule-based methods; 2) machine learning-based methods; and 3) scoring function-based methods. Our results show that both machine learning and scoring methods are effective and the scoring method can achieve a high precision with a reasonable recall. We also analyzed the contributions of different types of features and found that the temporality information between clopidogrel and bleeding events, as well as textual evidence from physicians’ assertion of the adverse events are helpful. We believe that our findings are valuable in advancing EHR-based pharmacogenomic studies.

INTRODUCTION

Large electronic health records (EHRs) databases have emerged as valuable data sources for clinical and translational research12. Moreover, EHR databases have also been linked to archived biological materials to accelerate research in personalized medicine3, e.g., identifying common and rare genetic variants that contribute to risk of diseases or to drug efficacy and toxicity (pharmacogenomics, or PGx). For example, the eMERGE network (electronic MEdical Records and GEnomics)5 is a consortium of 10 institutions, each of which has DNA biobanks coupled to large comprehensive EHRs. The longitudinal nature of the data contained within the EHRs make them ideal for quantifying drug outcome events (both efficacy and toxicity). Biobanks that are linked to large EHRs become ideal resources for PGx studies5,6, as they could overcome key challenges to advancing PGx discovery, such as the infrequency of adverse drug events, need for study of diverse populations, and rare variants. Successful stories about new PGx findings based on EHR-linked biobanks have been reported for many drugs, including warfarin5,6 and clopidogrel1, and promising results for new PGx discoveries with heparin-induced thrombocytopenia5, vancomycin drug levels9, and ACE-inhibitor induced cough10. These studies typically used rule-based algorithms combined with focused manual review. Another recent study explored warfarin related bleeds, but was particularly challenging requiring significant manual review11.

One of the challenges for EHR-based PGx studies is to identify patients with specific adverse drug reactions (ADR) after a drug exposure. It is costly and time-consuming to recruit domain experts to review medical charts to identify such cases. Therefore, informatics methods that can automatically identify drug-induced adverse events have been investigated. Sai and colleagues52 developed an algorithm to detect statin-induced myopathy from EHRs. They mined structured data to extract lab test results, ICD codes and the dates of statin prescription. By incorporating the time course information of creatinine kinase values, the lab test indicator of myopathy, their rule-based method achieved 100% precision in identifying patients with statin-induced myopathy. Haerian and colleagues13 proposed identification of cases where the adverse event is due to a patient’s pre-condition(s) and medication(s). They
focused on two ADRs, rhabdomyolysis and agranulocytosis, and developed a rule-based system which employs natural language processing (NLP) and expert-generated list of known risk factors. Their system showed a sensitivity of 93.8% and a specificity of 91.8%. Lin and colleagues proposed an automatic identification method for methotrexate-induced liver toxicity among patients with rheumatoid arthritis. They provided their system with several pieces of temporal information; their SVM classifier was trained on 3-month-long episode level, using features that are designed to reflect temporality such as verb tense. They achieved precision of 80.0% and recall of 89.9%.

Our long-term goal is to develop automatic methods to accurately identify cases and controls for EHR-based PGx studies. In this study, we developed and compared different informatics approaches to identify clopidogrel-induced bleeds in Vanderbilt University Medical Center (VUMC). We selected clopidogrel-induced bleedings as a use case since an ongoing PGx study about clopidogrel already exists wherein the research team has reviewed patients’ medical records to determine cases of clopidogrel-induced bleedings. Clopidogrel is often used in association with aspirin (Dual anti-platelet therapy, DAPT) to inhibit platelet function thereby to prevent recurrent adverse cardiovascular events in patients with coronary artery disease (CAD) undergoing percutaneous coronary intervention. While DAPT is widely used as recommended by the current guidelines, the therapy is also associated with an increased risk of potentially life-threatening adverse bleeding events. It is known that there is a great inter-individual variation in clopidogrel response; approximately 70% of which is heritable. Although the common loss of function CYP2C19*2 variant is known to be a major determinant that explains about 12% of the variation in clopidogrel responses, the majority of the heritable variation remains to be discovered. Identification of such gene variants that determine clopidogrel responses, especially those that will lead to adverse bleeding events, will provide the opportunity for more effective anti-platelet therapy through genotype-informed prescribing. Since relevant gene variants can often be relatively rare, larger sample sizes are required to determine the roles of gene variants in clopidogrel responses.

Both structured (ICD-9 codes) and unstructured (clinical notes) data were collected from the EHR and NLP technologies were applied to extract bleeding and clopidogrel information from the clinical documents. The NLP methods that we developed and evaluated were of three different types: 1) rule-based methods; 2) machine learning-based methods; and 3) scoring function-based methods. Both machine learning and scoring methods were shown to be effective, while the scoring method showed a high precision with a reasonable recall. In addition, our analysis on the contributions of different types of features showed that the temporality information between clopidogrel and bleeding events, as well as textual evidence from physicians’ assertion of the adverse events are helpful. We believe such valuable findings will advance EHR-based pharmacological and PGx studies.

METHODS

Datasets

At VUMC, there is an ongoing PGx study that aims to determine associations with increased adverse bleeding events after clopidogrel exposure, using the DNA bank bioVU and the Synthetic Derivative (SD) database, a de-identified copy of VUMC’s EHR database. For this project, the research team identified a cohort of 2,268 patients who potentially had clopidogrel-induced bleedings. The cohort selection criterion was based on structured EHR data - it required a bleeding ICD-9 code (decided by domain experts) within the first and last mention of clopidogrel in a patient’s medical record. Then, nurses manually reviewed the medical charts of these 2,268 potential cases and identified 1,921 true cases that had bleedings caused by clopidogrel and 347 “non-cases” that do not have bleeding events caused by clopidogrel. In this study, our task is to use various types of EHR data to automatically identify cases among the 2,268 patients that had bleedings caused by clopidogrel. We divided the 2,268 candidate patients into a training set of 1,512 patients (two thirds), on which we developed and trained our methods, and a test set of 756 patients (one third), on which we evaluated our methods.

Feature extraction

For each patient in the cohort, we collected both structured (e.g., ICD-9 codes from all steps of clinical pipelines) and unstructured (e.g., clinical notes) data from SD. For example, the training set contained a total of 709,076 clinical notes, and the test set contained 343,674 notes. Every note was then processed using two NLP tools: MedEx and KnowledgeMap Concept Identifier (KMCI). MedEx extracts drug names along with its signature information such as ‘strength’, ‘route’, ‘frequency’, ‘form’, ‘dose’, and ‘duration’. KMCI is a general clinical
NLP system developed at VUMC that extracts clinical entities and maps them to Concept Unique Identifiers (CUIs) in the Unified Medical Language System (UMLS). For our purposes, negated concepts were removed from the outputs.

After discussion with the domain experts, we identified the following features that may help identify clopidogrel-induced bleedings, including:

**Bleeding information:** Three types of bleeding information were extracted: 1) A set of 109 ICD9 codes about bleedings, as identified by domain experts; 2) A set of 465 CUIs related to bleeding, by expanding the ICD9 bleeding codes following the UMLS hierarchy. We checked for the presence of any of these 465 CUIs in a patient’s clinical notes, based on KMCI’s outputs; 3) A set of 51 keywords that are related to bleedings, as defined by domain experts. When we reviewed bleeding CUIs extracted by KMCI, we noticed that some frequently mentioned bleeding terms such as ‘epistaxis’ and ‘melaena’ were missed by KMCI. Therefore, we collected 51 keywords and developed a regular expression program to extract these keywords (non-negated) from clinical notes.

**Clopidogrel information:** Based on MedEx’s outputs, we extracted several features about clopidogrel exposure, e.g., how many times clopidogrel was mentioned, how many times it was mentioned with a dose, or frequency.

**Temporality information between bleeding and clopidogrel events:** It is critical to know how closely a bleeding event occurred to a clopidogrel mention. Therefore, we collected timestamps of all the bleeding and clopidogrel mentions extracted earlier. The timestamp information was obtained either from the date of ICD-codes or the date of clinical notes that mentioned bleeding (recognized through CUIs or keywords). Once we had the timeline of all bleeding and clopidogrel mentions of a patient, we calculated the shortest duration between a clopidogrel mention and a bleeding event as a feature. Domain experts also pointed out that other clinical events such as ‘chest tube’, ‘stab wound’, and ‘menses’ could also cause bleedings. Therefore, we compiled a list of these events (13 keywords) and identified their timestamps on the timeline as well. We then calculated durations between these events and bleedings which were then supplied as additional temporality features.

**Explicit evidence from physicians’ assertion:** We also noticed that clinical notes sometimes contain physicians’ assertions about bleedings caused by clopidogrel. For instance, the sentence “Bleeding was substantial from all raw surfaces owing to the fact that the patient was on double antiplatelet therapy with aspirin and clopidogrel” clearly states that the bleeding was caused by clopidogrel. Therefore, we deemed that it would be helpful to identify clopidogrel-induced bleedings by recognizing such explicit statements by the physicians. We employed a simple heuristic rule to detect such explicit evidence: if a sentence mentioned both clopidogrel and bleeding and had no negation phrase, we considered it to be explicit evidence and used it as a feature for classification.

**Classification methods**

After we extracted the above mentioned features, we developed three different types of methods to classify if a patient was a case (having clopidogrel-induced bleedings) or not by utilizing these features:

**Rule-based methods:** The rule-based methods employed two rules. The first rule was based on the temporality information. Given a patient’s timeline with dates of clopidogrel, bleeding, and other bleeding causes, we identified whether there was bleeding after a mention of clopidogrel without any other event(s) that may cause bleeding in-between. The second rule was based on explicit evidence. If a patient’s clinical notes contained a sentence that mentioned both clopidogrel and bleeding without any negation, the patient was classified as a ‘case’. We evaluated the performance of each rule individually as well as both rules together.

**Machine learning-based methods:** In this approach, we trained support vector machine (SVM) classifiers using different sets of features. We used LibSVM package as the implementation of SVM and optimized the parameters using the training set. We evaluated the contribution of different types of features as described below: 1) features containing basic bleeding and clopidogrel information only, e.g., the numbers of bleeding CUIs, ICD codes, and keywords, and the number of clopidogrel mentions; 2) temporality information between clopidogrel and bleeding, as described above; and 3) explicit evidence feature extracted from clinical notes; and 4) combined features of both temporality information and explicit evidence.

**Scoring-based method:** The scoring-based system calculated a score for each patient to indicate the likelihood of the patient’s having a clopidogrel-induced bleeding. If the score was equal to or greater than a pre-set cut-off value, the patient was classified as a ‘case’, otherwise as ‘non-case’. The system considered the temporality information
and the explicit sentence evidence only. The assumption was that shorter the time duration between a clopidogrel and a bleeding event, more likely was the patient a ‘case’. The scoring function is shown below:

\[ s = \sum_{f_{time} \in F_{time}} w_f (1 - f_{time} / 10,000) + w_{sent} f_{sent} \]

\( F_{time} \) denotes the set of all temporality features, whose member \( f_{time} \) has the length of minimum time duration from clopidogrel exposure to bleeding as its value. \( f_{time} \) differs from one to another on the method through which we identify the minimum time duration (e.g., the bleeding dates can be identified either by ICD codes, by CUI mentions, or by keywords, and the time duration may or may not contain a date with other events that may cause bleeding). \( f_{sent} \) refers to the explicit evidence feature, had value 1 when there is explicit sentence evidence and 0 otherwise. \( w_f \) and \( w_{sent} \) represent weights assigned to each temporality feature and the explicit evidence feature, respectively. The weight values were chosen empirically based on the training set.

**Evaluation**

We trained and developed all the systems using the training set, and evaluated them using the test set. In the past studies of EHR phenotyping algorithms, precision (also known as positive predictive value, or PPV) is often used to evaluate systems’ performance. Therefore, we used PPV as the primary measure in this study. In addition, we also reported recall (also known as sensitivity), to ensure the system did not lose too many true cases. Please note that the recall measure here does not represent the true recall from the entire dataset, as we do not know how many true cases exist in the complete dataset (i.e., the Vanderbilt SD) and the current dataset is not a random sample of the entire population (i.e., the current dataset is selected from the entire population based on the ICD-9 codes; See Datasets for more detail). Finally, we calculated F-score, harmonic mean of PPV and sensitivity, to compare different classification methods.

**RESULTS**

In this section, we report the performance of the different classification methods evaluated on the 756 candidate patients in the test set (Table 1). Among the rule-based methods, it is not surprising that the rule based on the explicit evidence of ADR assertions by physicians achieved the highest precision (96.9%), but with a much lower recall than the performance of the rule using temporality information. Among ML-based systems, the combined feature sets of temporality information and explicit evidence achieved best performance (88.7% precision and 96.6% recall). For the scoring-based approach, increasing the cutoff value increased precision but lowered recall. Overall, ML-based and scoring-based methods outperformed the simple rule-based methods when both precision and recall are considered. If our primary goal is to achieve high precision (e.g., over 95%), the scoring-based method would be the best option, as it achieved a precision of 95% with a reasonable recall of 65.3%.

**Table 1.** The performance of different approaches on the test set.

| Category | System Description | Precision (PPV) | Recall (Sensitivity) | F-score |
|----------|-------------------|----------------|----------------------|--------|
| Rule     | Rule-based with temporality information | 92.1% | 80.2% | 85.74% |
|          | Rule-based with explicit evidence | 96.9% | 28.9% | 44.52% |
|          | Rule-based with both temporality and explicit evidence | 91.83% | 82.5% | 86.92% |
| ML       | ML with bleeding and clopidogrel information | 84.6% | 99.8% | 91.57% |
|          | ML with temporality information | 86.8% | 95.3% | 90.85% |
|          | ML with explicit evidence | 96.9% | 28.9% | 44.52% |
|          | ML with temporality and explicit | 88.7% | 96.6% | 92.48% |
| Scoring  | Scoring-based system (cutoff = 1) | 86.4% | 98.0% | 91.84% |
|          | Scoring-based system (cutoff = 3) | 90.4% | 90.2% | 90.30% |
|          | Scoring-based system (cutoff = 6) | 95.0% | 65.3% | 77.40% |
DISCUSSION

In this study, we developed and compared different NLP methods for automatic identification of patients with ADEs from EHRs, using clopidogrel-induced bleeding as our use case. We extracted various feature sets from both structured data and narrative clinical notes. The uses of features derived from information about bleeding, clopidogrel, their temporality relations, and explicit evidence from sentences were investigated, and we found that the temporality information and explicit evidence were helpful in identifying clopidogrel-induced bleedings. We also tested different classification strategies, i.e., rule-based, ML-based, and scoring-based methods. Although both ML-based and scoring-based methods showed better performance than the rule-based method, the scoring-based method seems to be more suitable if a high precision is desired. Our findings not only demonstrate the feasibility of automatic identification of clopidogrel-induced bleedings, but also provides valuable insights to handle the ADR case detection in general, thus to advance the EHR-based PGx studies.

Temporality between drug and ADR events is critical in this task. Despite the reasonable performance of our systems, our current strategies for temporal information extraction, representation, and analysis are very limited. For example, we simply assigned document time to events in a clinical note, which may generate false timestamps of clopidogrel and bleeding, as some events could have occurred in the past or could be indicated as future possibilities (though many of these should be filtered out using our implementation of NegEx). There are NLP tools that recognize a concept’s temporal relation to the document time. We plan to utilize such tools to produce more accurate temporality features, as investigated by Lin and colleagues for methotrexate–induced liver-toxicity. Another limitation is that we currently assume that a patient is on clopidogrel if clopidogrel is mentioned in a clinical note. However, medication mentions could refer to different status such as “start”, “stop”, “hold”, and “dosage change”, as described in Liu and colleagues. To further improve the performance, we plan to adapt and incorporate such deep drug exposure modeling methods into our systems. Furthermore, in our study, we simply used durations between events for building the ML model, which could be replaced by more sophisticated methods that consider entire timelines, e.g., time-series data mining techniques.

The following false positive example (misclassified as a ‘case’ while being a ‘non-case’) highlights the need for comprehensive analysis of temporality and drug status: a note on September 20th 2011 says “Recent episode of gross hematuria at OSH with resultant demand ischemia and NSTEMI.”, mentioning a bleeding event (hematuria) that occurred before the document creation time (Sep. 20th). The note also says “Per Urology, no obvious source of hematuria and has not reoccurred in house, will proceed with anticoagulation and LHC.”, from which one can identify that the bleeding is no longer on-going as of Sep. 20th. The note goes on with “Please take plavix for three more days total.”, where the use of the word ‘more’ indicates that the patient was taking clopidogrel (also known as plavix) for some time. Based on the note, our annotators classified the patient as a ‘non-case’, since they inferred that the patient had no bleeding episode whilst on clopidogrel (at least during hospitalization). However, our classification methods classified the patient as a ‘case’, since hematuria was mentioned in several notes from September 20th to 23rd together with clopidogrel, as it did not take into consideration the detailed analyses in terms of temporality and drug status. Note that the above example was misclassified as a ‘case’ by all three classification methods. Other false positives misclassified by all three methods showed similar characteristics (i.e., lack of proper analysis of temporality and/or drug status).

A novel and interesting finding of this study is the utilization of explicit evidence from sentences that contain physicians’ assertions about ADRs. This new type of information was shown to be highly effective in increasing the precision. The rule-based system using explicit evidence showed a precision of 97%. The system also showed a recall of 29%, indicating that more than one fourth of the patients with clopidogrel-induced bleedings have an explicit description about the causal relation between clopidogrel and bleeding in their clinical notes. Although the explicit evidence feature already produces high precision, we expect to have further improvement by employing NLP techniques that can precisely target the causal relation. Furthermore, explicit evidence can also be utilized to identify ‘non-case’ patient, e.g., we can identify sentences that explicitly describe that a bleeding is caused by other medical events rather than clopidogrel. For example, the sentence “He is having significant intermittent lower GI bleeding from radiation proctitis”, can be a strong signal to classify a patient as a ‘non-case’, thus improving the overall performance of the classification system.
This study has limitations. The gold standard was generated by an existing PGx study, which used only structured data to generate the candidate cases. If we use different criteria for candidate case selection, both the precision and recall reported here could change, although precision is probably more reliable measure of the two. Moreover, we only evaluated these methods on one use case (clopidogrel-induced bleeding). We believe that PGx studies on other similar drug adverse reactions would benefit from the findings of this study (i.e., the use of temporal information, filtering with other possible causes of adverse event, and the use of explicit mentions of adverse reactions). However, the generalizability of the findings from this study may need further validation. In addition, we want to mention that application of the proposed methods to other ADR types would require additional manual work, which includes collecting relevant ICD/CUI codes, identifying other possible causes, and selecting appropriate parameters for classification methods (e.g., time duration for rule-based system, weights for scoring-based system). For future work, besides method improvements on temporality and explicit evidence, we also plan to extend the work to additional PGx studies other than clopidogrel-induced bleeding.

CONCLUSION
Identifying patients with specific ADRs is a critical step in EHR-based PGx study, but the time and labor costs for manual chart review is huge. We investigated different informatics approaches for identification of patients with ADRs, using clopidogrel-induced bleeding as our case study. From our experiments with different computational methods and different types of features, we found that (1) while both ML and scoring methods are effective in identifying patients with bleedings caused by clopidogrel, the scoring method can achieve a high precision with a reasonable recall, and (2) temporality information between clopidogrel and bleeding events, as well as textual evidence from physicians’ assertion of the adverse events are valuable features. In our future work, we will further investigate the methods to improve the performance, and will apply our methods to ADRs other than clopidogrel-induced bleeding.

ACKNOWLEDGEMENT
This study is supported in part by grants from NIGMS 1R01GM103859 and 1R01GM102282, CPRIT R1307, and NHLBI U19HL065962.

References
1. Shea S, Hripcsak G. Accelerating the Use of Electronic Health Records in Physician Practices. N Engl J Med. 2010;362(3):192-195.
2. Jha AK, DesRoches CM, Campbell EG, et al. Use of Electronic Health Records in U.S. Hospitals. N Engl J Med. 2009;360(16):1628-1638. doi:10.1056/NEJMsa0900592.
3. McCarty CA, Wilke RA. Biobanking and pharmacogenomics. Pharmacogenomics. 2010;11(5):637-641. doi:10.2217/pgs.10.13.
4. Manolio TA. Collaborative genome-wide association studies of diverse diseases: programs of the NHGRI’s office of population genomics. Pharmacogenomics. 2009;10(2):235-241. doi:10.2217/14622416.10.2.235.
5. Roden DM, Xu H, Denny JC, Wilke RA. Electronic Medical Records as a Tool in Clinical Pharmacology: Opportunities and Challenges. Clin Pharmacol Ther. 2012;91(6):1083-1086. doi:10.1038/clpt.2012.42.
6. Wilke RA, Xu H, Denny JC, et al. The Emerging Role of Electronic Medical Records in Pharmacogenomics. Clin Pharmacol Ther. 2011;89(3):379-386. doi:10.1038/clpt.2010.260.
7. Delaney JT, Ramirez AH, Bowton E, et al. Predicting clopidogrel response using DNA samples linked to an electronic health record. Clin Pharmacol Ther. 2012;91(2):257-263. doi:10.1038/clpt.2011.221.
8. Karnes JH, Cronin RM, Rollin J, Teumer A. A genome-wide association study of heparin-induced thrombocytopenia using an electronic medical record. Thromb Haemost. 2015.
9. Van Driest SL, McGregor TL, Edwards DRV, et al. Genome-Wide Association Study of Serum Creatinine Levels during Vancomycin Therapy. Thameem F, ed. PloSone. 2015;10(6):e0127791. doi:10.1371/journal.pone.0127791.
10. Mosley JD, Shaffer CM, Van Driest SL, et al. A genome-wide association study identifies variants in KCNIP4 associated with ACE inhibitor-induced cough. Pharmacogenomics J. 2015;16(3):231-237. doi:10.1038/pj.2015.51.
11. Kawai VK, Cunningham A, Vear SI, et al. Genotype and risk of major bleeding during warfarin treatment.
12. Sai K, Hanatani T, Azuma Y, et al. Development of a detection algorithm for statin-induced myopathy using electronic medical records. *J Clin Pharm Ther*. 2013;38(3):230-235. doi:10.1111/jcpt.12063.

13. Haerian K, Varn D, Vaidya S, Ena L, Chase HS, Friedman C. Detection of Pharmacovigilance-Related Adverse Events Using Electronic Health Records and Automated Methods. *Clin Pharmacol Ther*. 2012;92(2):228-234. doi:10.1038/clpt.2012.54.

14. Lin C, Karlson EW, Dligach D, et al. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. *J Am Med Inform Assoc*. 2015;22(e1):e151-e161. doi:10.1136/amiajnl-2014-002642.

15. Steinhubl SR, Berger PB, Mann JT III, et al. Early and Sustained Dual Oral Antiplatelet Therapy Following Percutaneous Coronary Intervention: A Randomized Controlled Trial. *JAMA*. 2002;288(19):2411-2420. doi:10.1001/jama.288.19.2411.

16. Kushner FG, Hand M, Smith SC, et al. 2009 Focused Updates: ACC/AHA Guidelines for the Management of Patients With ST-Elevation Myocardial Infarction (Updating the 2004 Guideline and 2007 Focused Update) and ACC/AHA/SCAI Guidelines on Percutaneous Coronary Intervention (Updating the 2005 Guideline and 2007 Focused Update): A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol*. 2009;54(23):2205-2241. doi:10.1016/j.jacc.2009.10.015.

17. Aronow HD, Steinhubl SR, Brennan DM, Berger PB, Topol EJ, CREDO Investigators. Bleeding risk associated with 1 year of dual antiplatelet therapy after percutaneous coronary intervention: Insights from the Clopidogrel for the Reduction of Events During Observation (CREDO) trial. *Am Heart J*. 2009;157(2):369-374. doi:10.1016/j.ahj.2008.09.011.

18. Bhatt DL, Fox KAA, Hacke W, et al. Clopidogrel and Aspirin versus Aspirin Alone for the Prevention of Atherothrombotic Events. *N Engl J Med*. 2006;354(16):1706-1717. doi:10.1056/NEJMoa060989.

19. Shuldiner AR, O’Connell JR, Bliden KP, et al. Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy. *JAMA*. 2009;302(8):849-857. doi:10.1001/jama.2009.1232.

20. Holmes MV, Perel P, Shah T, Hingorani AD, Casas JP. CYP2C19 Genotype, Clopidogrel Metabolism, Platelet Function, and Cardiovascular Events: A Systematic Review and Meta-analysis. *JAMA*. 2011;306(24):2704-2714. doi:10.1001/jama.2011.1880.

21. Wallentin L, James S, Storey RF, et al. Effect of CYP2C19 and ABCB1 single nucleotide polymorphisms on outcomes of treatment with ticagrelor versus clopidogrel for acute coronary syndromes: a genetic substudy of the PLATO trial. *Lancet*. 2010;376(9749):1320-1328. doi:10.1016/S0140-6736(10)61274-3.

22. Roden DM, Pulley JM, Basford MA, et al. Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clin Pharmacol Ther*. 2008;84(3):362-369. doi:10.1038/clpt.2008.89.

23. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc*. 2010;17(1):19-24. doi:10.1197/jamia.M3378.

24. Denny JC, Smithers JD, Miller RA, Spickard A. “Understanding” medical school curriculum content using KnowledgeMap. *J Am Med Inform Assoc*. 2003;10(4):351-362. doi:10.1197/jamia.M1176.

25. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol*. 2011;2(3):27-27. doi:10.1145/1961189.1961199.

26. Ritchie MD, Denny JC, Crawford DC, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet*. 2010;86(4):560-572. doi:10.1016/j.ajhg.2010.03.003.

27. Tang B, Wu Y, Jiang M, Chen Y, Denny JC, Xu H. A hybrid system for temporal information extraction from clinical text. *J Am Med Inform Assoc*. 2013;20(5):828-835. doi:10.1136/amiajnl-2013-001635.

28. Lin C, Dligach D, Miller TA, Bethard S, Savova GK. Multilayered temporal modeling for the clinical domain. *J Am Med Inform Assoc*. 2015;23(2):ocv113-ocv119. doi:10.1093/jamia/ocv113.

29. Liu M, Jiang M, Kawai VK, et al. Modeling drug exposure data in electronic medical records: An application to warfarin. *AMIA Annu Symp Proc*. 2011,2011:815-823.

30. Chatzis SP, Kosmopoulos DI, Varvarigou TA. Robust Sequential Data Modeling Using an Outlier Tolerant Hidden Markov Model. *IEEE Trans Pattern Anal Mach Intell*. 2009;31(9):1657-1669. doi:10.1109/TPAMI.2008.215.

31. Warren Liao T. Clustering of time series data—a survey. *Pattern Recognition*. 2005;38(11):1857-1874. doi:10.1016/j.patcog.2005.01.025.
32. Rindflesch T, Libbus B, Hristovski D, Aronson AR, Kilicoglu H. Semantic relations asserting the etiology of genetic diseases. *In proceedings of AMIA annual symposium*. 2003:554-558. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1480275/.

33. Girju R, Moldovan DI. Text mining for causal relations. *In proceedings of FLAIRS conference*. 2002:360-364.