Preliminary Exploratory Data Analysis of Simulated National Clinical Data Research Network for Future Use in Annotation of a Rare Tumor Biobanking Initiative

Alex S. Felmeister
College of Computing and Informatics
Drexel University
Philadelphia, USA
felmeistera@email.chop.edu*

Angela J. Waanders, MD, MPH
Division of Oncology
The Children's Hospital of Philadelphia
Philadelphia, USA

Sarah E.S. Leary, MD
Hematology-Oncology
Seattle Children's Hospital
Seattle, USA

Jeff Stevens
Cancer and Blood Disorders Center
Seattle Childlren's Hospital
Seattle, USA

Jennifer L. Mason
Division of Neurosurgery
The Children's Hospital of Phildelphia
Philadelphia, USA

Rachel Teneralli
Applied Clinical Research Center
The Children's Hospital of Phildelphia
Philadelphia, USA

Xiaohua Hu, PhD
Collegeo of Computing and Informatics
Drexel University
Philadelphia, USA

L. Charles Bailey, MD, PhD
Divisions of Oncology and Hematology
The Children's Hospital of Philadelphia
Philadelphia, USA

***Abstract*—Observational data resources based on the capture of clinical data in the electronic health record (EHR) have produced significant learning opportunities in many areas of medicine. These large data resources can span multiple hospital systems and employ common semantics, ontologies, and data models. They have uncovered critical safety issues for patients, and spurred observational research and clinical decision support. In the age of precision medicine there is also an increased need to obtain genomic and clinical data to discover novel treatments for the deadliest of diseases. With this, there are efforts to create deep-dive disease specific repositories that include tissue in biobanks. The latter require significant human annotation of biospecimens. Securing the data is especially critical in rare pediatric brain tumors. In the specific case of The Children's Brain Tumor Tissue Consortium (CBTTC) an international rare pediatric brain tumor repository, the number of patients that need to be followed prospectively is outpacing the ability of human annotation. In this preliminary study, we perform a prescribed data exploration analysis on simulation data in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) employed by the pediatric data network PEDSNet with the intention to ascertain feasibility in automatic annotation of patient records in the CBTTC.**

Keywords—Data Mining, Health Informatics, Translational Research, Health Data Acquisition

I. INTRODUCTION

This preliminary research is intended to evaluate analysis methods on observational clinical data to automatically obtain necessary data for tumor-tissue-based research related to diagnostics, tumor location, survival, treatment, and prognosis. The approach here will focus on predicting pertinent information from the large observational dataset and discovering data points that contribute to the data-driven phenotype of a diagnosed subject [1]. The larger research project we are pursuing involves two national organizations, PEDSnet[2] and The Children's Brain Tumor Tissue Consortium (CBTTC)[3]. PEDSnet, a PCORnet Clinical Data Research Network (CDRN)[4], is an information resource containing over 5.3 million open-ended pediatric observational data sets aggregated from clinical data collected from Electronic Health Records (EHRs).

The Children's Brain Tumor Tissue Consortium (CBTTC) is a growing, open-ended biologically focused pediatric brain tumor resource focused on creating easy-to-obtain cohorts around pediatric rare brain tumor specimens with rich longitudinal clinical annotation. Both projects look to bring data *to* the researchers and increase the [5] at which there can be high-impact clinical research created for a population in need. The patient sets overlap in both projects because two institutions participate in both projects; The Children's Hospital of Philadelphia and Seattle Children's Hospital. The two resources have different aims, but together they combine biologically focused research with clinical outcomes research. This manuscript is a report on the preliminary work to obtain information from the standard model adopted by PEDSnet and use simulation data to gauge the potential predictive capabilities of PEDSnet on longitudinal, event based data.

There are some differences between biologically based tissue research and strictly observational research: 1)

This continued research is supported by the iFellowship Program administered by the University of Pittsburgh as part of the iSchool Consortium. Funding is provided by The Andrew W. Mellon Foundation.

level of automation of data collection, 2) human burden of data abstraction, and 3) specifics of information necessary for research and operational logistics. Similarities include: 1) longitudinal, temporally and sequentially based data sets, 2) cross-institutional effort, 3) high national visibility, 4) overlapping specific patient and institutional participation and 5) common goals to make progress in disease-related research.

Though this manuscript includes many accessible methods, the outcome we are looking for is an ability to create set of data that adds up to a data-driven phenotype that has the power to predict the condition of a patient but conversely can be used to annotate biological tissue specimens. We take a data exploratory analysis approach to discovering features based on clinical expert advice and perform direct structured database querying to prepare the simulation data for multiple iterations of predictive analytics based on frequencies and sequences of events in observational health data. The data used for prediction is part of a patient data-driven phenotype and could be integrated with the CBTTC.

II. Motivation

With the dawn of genomics and precision medicine, there is an increased need to create accessible repositories of data and specimens for research[6]. This information is gathered without hypothesis to empower researchers to dive in and make advances in areas in which there is little advance[7]. A drive to create open data and collaborative initiatives in rare and deadly diseases is evidenced in the national spotlight from government to private industry[8]. A reaction to this need, specifically in cancer, is seen in programs like the CBTTC.

III. Related Works

A. Use of Observational Data

Observational data can be used in many different circumstances. In the literature, observational data sets are used to accomplish:

- Adverse event prediction for drug safety[9];
- Finding negative controls for studies when comparing patient clinical behavior[10][11];
- Epidemiological studies of co-varying events and environments[12];
- Quality issues around pharmacovigilance over time[13];
- The evaluation of definition of standards for consistency in evaluating visit types in health systems[14];
- Leveraging patient similarity versus genomics to create methods in personalized medicine[15]; and
- Discovering computable and data-driven phenotypes[16][17].

We are interested in discovering the computable phenotype or data-driven phenotype through predictive analytics on time-series and frequency in observational clinical records. The computable phenotype is a representation of all of the data and a combination of knowledge of a patient within the electronic health record (EHR) inclusive of all digital elements beyond tagged observations and conditions[16]. We are looking to find data points, that include a temporal component, to aid in the biologically based research in rare brain tumors.

IV. Methods

The first component of this research was to analyze and become familiar with the capabilities of the Observational Medical Outcomes Partnership (OMOP) common data model data model (CDM)[5]. This is a relational model representing observational data usually derived from the EHR. We are potentially gauging the utility of this data set and data model based on the predictive power of the data. Therefore, the data must be transformed and prepared for running multiple algorithms to predict a specific condition in a patient based on frequencies and sequences of events. For the purpose of this manuscript methods laid out here are for the sole purpose of evaluating the data in the OMOP model and setting up plug-and-play pipelines for transformations and analyses upon obtaining data sets. This initial research is to start to understand, transform, and run some simulations pertinent to data points required by brain tumor researchers of the CBTTC.

A. Data Source

The observational data is built on the OMOP data model adopted by Observational Health Data Sciences and Informatics (OHDSI). OHDSI serves many large-scale networked and harmonized observational data sets based on electronic medical record data[18]. For this preliminary work, we obtained a 1,000 person sample of simulated CMS SynPUF patient data[19]. This data is modeled on real observations from real data sources, but the patients have fictional medical histories and therefore the data cannot be used to make any research conclusions from analysis[20]. The data set is specifically intended to create and test applications for OHSDI. Therefore, we used this data to build and test methods for this research.

Vocabulary in the OMOP CDM specifically for diagnoses/conditions, observations and drugs are based on SNOMED-CT[21], ICD-9-CM[22] and RxNorm[23] respectively. Some of these concepts can map to each other and this is noted in the concept data file retrieved from the ATHENA standardized vocabulary tool from OHDSI[24]. We requested an extensive concept table with about 2.8 million concept descriptions. Table 1 is a list of some of the value counts of a selection of 41 specific sets of vocabularies used to describe objects in the OMOP data model.

B. Transformation

We took a multi-modal approach to transforming the data through direct structured database querying and in-memory

TABLE I. SELECTION OF CONCEPTS USED IN OMOP

Concept	Number of Unique Concepts
SNOMED-CT	785,712
NDC	679,579
RxNORM Extension	617,949
RxNorm	257,977
LOINC	123,717
ICD10CM	106,666
NDFRT	37,344
ICD9CM	18,672
CPT4	14,579
…	

module-based functions available in common statistics packages. This multi-modal approach as discussed in Peissing et al. gave us the ability to continue to evaluate the data during the transformation process until the data was consistent across concepts for analyses[25].

The concepts were matched to each data table based on the specific instruction set forth in the OHDSI documentation. In the first data transformation, we created verbose data tables for the patient, condition, observation and drug exposure objects in the OMOP data model and joined this with patient demographics (gender, race, date of birth etc.). The data model contains other information pertaining to billing and hospital administration in the electronic health record, but these were not included in our evaluation. We focused only on the exploration of the objects associated with the progression of the patient and their interaction with hospital services and clinical observations. The *patient*, *conditions*, *observations* and *drug exposures* were put through a simple exploratory analysis pipeline consisting of transforms, labeling concepts, and joining with patient demographics to create verbose versions of objects for further computational and human interrogation of the data. All transformation and provenance of the data are maintained in code repositories utilizing Python notebooks to allow us to transparently navigate backwards and forward through the data transformation process.

Temporal and frequency data was then derived using the date information and counts across multiple tables. For observations, the verbose data tables created were further processed to join an age of the patient at every patient-observation event. Finally, a frequency matrix was created for every patient and the number of times an observation took place for that patient. We performed a similar process for drug exposures. The "condition occurrence" object are the times when a patient's diagnosis or current condition is entered into the medical record at a visit or encounter. This table was joined on to frequency matrices where applicable to categorize patient records with a diagnosis or condition.

C. Preparation and Prediction

The next step was to identify patient sets based on observations or conditions recorded. We took a supervised approach to this process utilizing subject matter experts in Oncology, Epidemiology and the PEDSnet network to get a better understanding as to what features are important to the domain. There was no need, at this point, to use unsupervised approaches to find factors for this research. Further, those would be quite impossible with the current simulation data[1]. We identified some general diagnoses that could categorize the patient population in the simulation data that will be similar to how we categorize patients in the pediatric component of this project. The following categorizations were established in the simulation data to run the analysis algorithms in Table 2. This table also describes the data source of the derived categorization.

We then established feature combinations of observations and drug exposures and time point driven data sets of observations and drug exposures. These sets of categorized and feature frequency or time-point data were then iteratively put through a variety of predictive methods and visualizations. Though prediction performance is not an indicator of any real-world situation in this case, it is important that the data worked with the pipelines being developed, so we are ready for the real data when it is available to us. Data was transformed into the following tables/matrices:

- Matrix of observation frequencies.
- Matrix of drug exposure frequencies.
- Matrix of drug exposure frequencies and observation frequencies.
- Attribute list of Conditions and age at first diagnosis.
- Attribute list of observations and age at observation.
- Attribute list of drug exposures and age at drug exposure.
- Sequence list of observations with first observation as 0.
- Sequence list of drug exposure with first observation as 0.
- Sequence list of drug exposure and observations with first of either at 0.

TABLE II. FOUR CATEGORIZATIONS DERIVED FROM THE DATA

Categorization Description	Possible Values	Source
Person History of *Specific* Cancer Reported (not specifically in the health system)	Thyroid, Breast, Prostate, Lung, Brain, Cervix, Ovary, Larynx, GI tract, stomach …	Occurrence Object: ICD-9-CM/ICD-10-CM
Personal History of Cancer Reported (not specifically in the health system)	Cancer, Non-Cancer	Occurrence object: occurrence description based on ICD-9-CM/ICD-10-CM
Specific primary malignancy frequently observed across the population	Lung, Breast, Prostate, Colon	Condition-occurrence object: Condition Occurrence Description based on SNOMED-CT Condition occurrence description based on SNOMED-CT
Primary Malignancy diagnosis in the health system	Cancer, Non-cancer	Condition-occurrence object: Condition occurrence description based on SNOMED-CT

Each analysis-ready table can be toggled based on population based frequencies. For example, the matrix of observations can show only those found frequently in the database population to avoid rare conditions or vice-versa[26]. This toggling method is shown in the literature when exploring event based data looking for patterns or predictive capability[27].

D. Survival Data

The final method we wanted to apply to the simulation data were survival metrics. As discussed in the introduction of this paper, the domain of interest for this research will include longitudinal components of a child's treatment, adverse events and survival when diagnosed with rare brain tumors which all contribute to overall and event-free survival metrics[28], [29]. A Kaplan-Meier curve is a method to deal with incomplete observations with differing survival times and used to compare survival between groups. To create this curve, we needed to obtain the patient's 1) serial time, 2) their status at the end of the serial time (with potential censoring) and 3) a categorical item or study group[30].

TABLE III. FOUR CATEGORIZATIONS DERIVED FROM THE DATA

Person	T	E	Group	Disease
389	27.0	1	Sim.	Cancer
153	19.0	1	Sim.	Non-cancer
176	21.0	1	Sim.	Cancer
261	22.0	1	Sim.	Non-cancer
9	10.0	1	Sim.	Non-cancer
…				

In the simulation data, the first time a patient has an occurrence of a primary malignancy, we mark that as month 0. Then the latest event is calculated from the amount of time from event 0. If that event was death, we mark the event column. We also needed to be able to add any patient category for survival comparisons. Below in Table 3 is an example of the data prepared for Kaplan-Meier analysis. We used our most general categorization from the frequency analysis to build out survival analyses. We applied the LifeLines python module which has the ability to give average views of the population[31]. The module also has tools to evaluate survival regression when working with individual patient records. For this example we used the module's fitter specifically stating

patient death if known and right censoring patients without a known date of death.

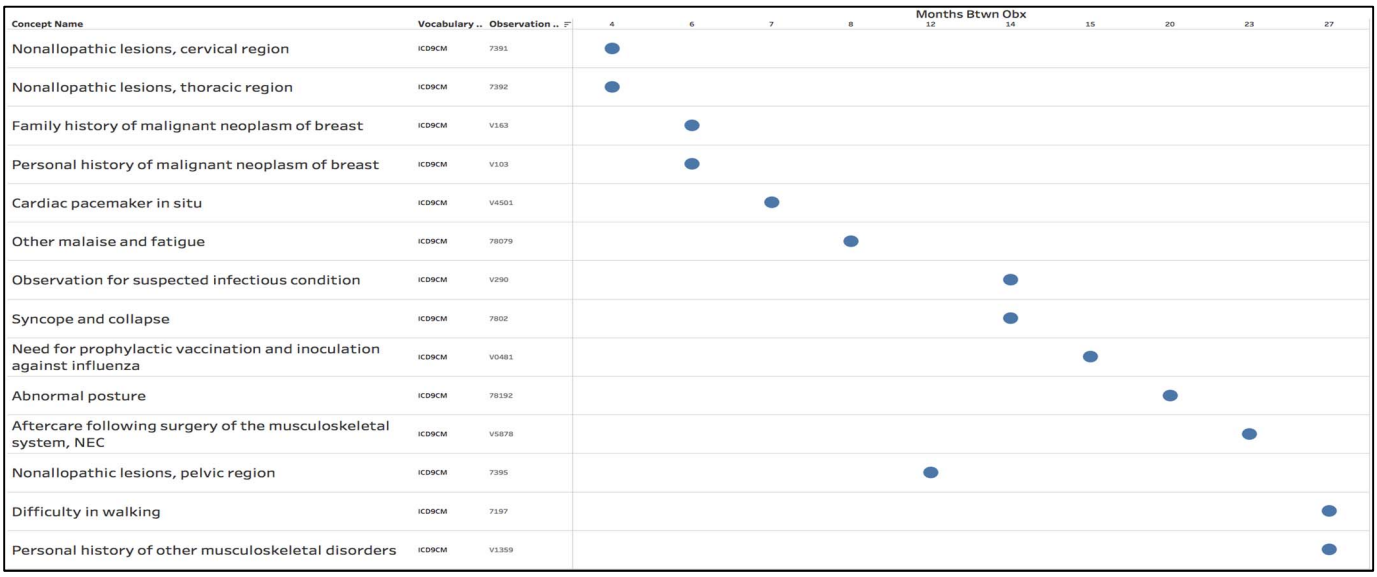

Fig 1. Single patient view of longitudinal drug exposures and observations.

V. Results

A. Data Preparation

In this study, the simulation data, based on a real-world larger data set of adult Medicare patients was sufficient to perform complex transformations, data preparations and pipeline multivariate prediction simulations. The data was also labeled verbosely to explore the various diagnostic, observation and drug exposures during transformation and processing. The model lent itself to transformations that included temporal features, drug exposure based frequency matrices and combinations of the two. All of these could be crossed with patient demographics and diagnoses from the condition table. For example, we were able to create verbose frequency tables which included human readable along with machine computable components.

B. Longitudinal Represenation of Events

The data also lent itself to straightforward visualizations at the individual and population levels. This lent itself to deep patient-based analyses for rare disease and comparison with populations. The series of events for a specific patient set can be visualized in a similar way to other observational studies either looking at rare features or large epidemiological features[32]. Fig. 1 is an example of a patient event timeline of coded events and drugs represented from the simulation data. Coding is a mix of ICD-9-CM codes and RxNorm. The actual label is inconsequential because the data is simulated. The figure shows the ability to lay out an observation timeline for an individual patient.

C. Predictive Power

Though prediction methods will be used for validation with data from PEDSnet, it was important to set up scafolding for this specific data model and build pipelines ready for larger data sets. For each set of transformed and labeled data, the population was split into 10 parts and our model was trained on 9. Tests were done on 1 leftover set. This process was then repeated for mix of linear and non-linear algorithms. A percentage was created by comparing the correct predictions versus the full population in the data set[33]. This rudimentary approach yields very low performance scores. This is to be expected from the known limitations of the simulation data set used for this preliminary research. We know that the simulation data is based on actual real-world Medicare records, but they are collapsed into one patient for the purpose of learning the OMOP CDM and not for predictive power.

Surprisingly, there was a set of data in the simulation model that yielded some good predictive capability, but it was categorized at a very high level. The reason could be that the simulation data is actually based on real examples from other OMOP databases. In the literature describing predictive power in observational databases, there are examples of combining drug exposures and observations to get higher predictive capability[26]. Fig. 2 displays the percent of correctly identified cases from the different multi-variate machine learning algorithms in a box plot representation. Though inconsequential, this rating system will be used against the real PEDSnet data. Here we see K-Nearest Neighbor (KNN) has the highest success rates of prediction. This data with high predictive scores was labeled with a simple binary, “personal history of cancer” or “no personal history of cancer” versus patient-frequencies of highest population frequent events of all potential observations and drug exposures. When using the condition of a “primary malignancy,” the accuracy rates on the algorithms has very low performance even if factoring in all observations, drugs, their frequencies and sequence in the patients visit. Again, this was an expected result. The purpose of putting this data through these algorithms is the set-up the pipeline for real data to be evaluated in our future research.

D. Survival Analysis

The data in the CDM is highly temporal. Each event at any level of the model is tagged with time points or time ranges. The specific domain that will be addressed in future research is overall survival. The data lent itself to creating Kaplan-Meier curves because the data includes the date of initial primary diagnosis, the last date of patient contact and a date of death. These data points display Kaplan-Meier curves for any qualification category necessary to explore. The fact that this method also allows for right censoring is important because we will not always know if a subject is living or lost to follow up. Though Fig. 3 is an arbitrary representation; all data points are available to create these curves that are key to unlocking a larger picture for rare disease. Fig. 1 is a fitted Kaplan-Meier curve based on the entire population in the simulation data. It can be categorized with a general personal history of malignancy as previously discussed in the method section was also performed, but because of the nature of the simulation data it is not described or pictured in this paper because it does not fit a real-life circumstance and could be confusing. If this data were real, it would be describing an estimate of patient survival across the database population. In the end, we have data in the OMOP data model that will provide the information for survival analysis specifically the analysis done in cancer research. We will also have the opportunity to provide researchers with comparative survival

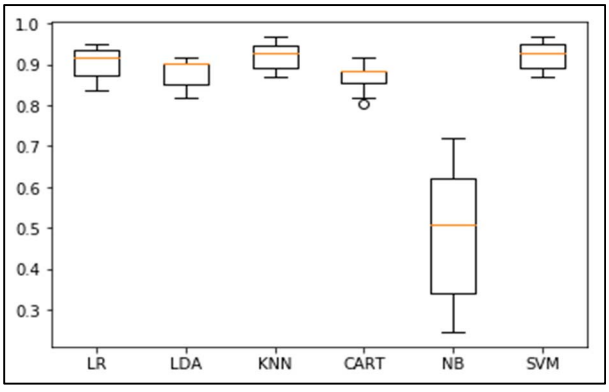

Fig 2. Algorithem performance comparision showing percent of correctly identified cases from different multi-variate machine learning algorithms; Logistic Regression Linear Discriminant Analysis, K Neighbors Classifier, Decision Tree Classifier, Gaussian and SVM.

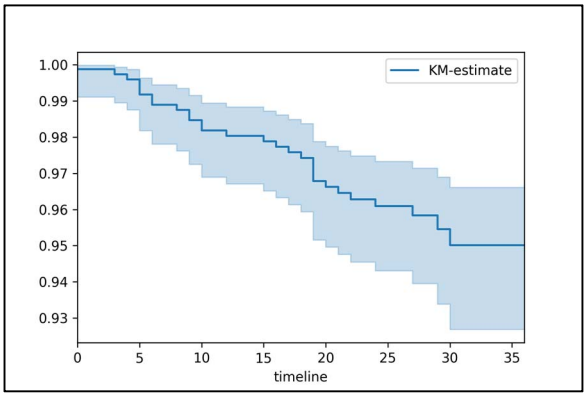

Fig 3. A Kaplan-Meier estimator representing a survival function from longitudinal data available in the simulation data.

curves from different resources along with the curves of populations discovered in either research repository which could become part of a data-driven phenotype that further classifies a disease.

VI. CONCLUSION AND DISCUSSION

One of the lessons learned in this exploratory analysis is that the PEDSnet data in the OMOP model could provide valuable temporal based information to the rare tissue repository. The ability to make population based survival estimates alone is a good result. At this point in this project pipelines are built in Python, and ready to accept any other set of data based on the OMOP common data model. As this research moves forward, the exploratory data analysis will continue to be narrowed in until the domain of interest is validated. Continued work on this project will include data that will be used to create consistent patient-to-patient data driven phenotypes in rare pediatric brain tumor cases. Of which, the components of the data-driven phenotypes will be used to supplement human curated national biobanking consortium that includes longitudinal clinical data. The initial exploratory data analysis performed here shows that the data in the OMOP common data model has great potential for use in rare cancer biomedical research because of its ability to quickly be processed into multiple common machine learning algorithms, visualizations and common survival and time-based analyses.

VII. FUTURE WORK

This is new research and expected to be completed in December 2018. The next steps are to receive pediatric data from a cohort of patients diagnoses with aggressive forms of cortical tumors, and the transformations and methods will be iteratively applied to data as it is received. The overall intention of this research is to discover if data in this particular clinical data research network could help automate the voluminous longitudinal data collection in the CBTTC with a more complete data-driven phenotype.

VIII. ACKNOWLEDGEMENTS

The authors would like to thank the generous opportunity from the Mellon Foundation though The Coherence at Scale Program; the talented staff and researchers of The Children's Brain Tumor Tissue Consortium (CBTTC); the Enterprise Informatics Group at the Department of Biomedical and Health Informatics (DBHi) at The Children's Hospital of Philadelphia and PEDSnet. A.S.F. would like to also acknowledge his doctoral advisor, Xiaohua "Tony" Hu, PhD and the guidance of K. David Harrison, PhD.

REFERENCES

[1] T. A. Lasko, J. C. Denny, and M. A. Levy, "Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data," *PLoS One*, vol. 8, no. 6, 2013.

[2] C. B. Forrest, P. a Margolis, L. C. Bailey, K. Marsolo, M. a Del Beccaro, J. a Finkelstein, D. E. Milov, V. J. Vieland, B. a Wolf, F. B. Yu, and M. G. Kahn, "PEDSnet: a National Pediatric Learning Health System.," *J. Am. Med. Inform. Assoc.*, vol. 21, no. 4, pp. 602–6, 2014.

[3] A. R. Alex Felmeister, Rishi Lulla, Angela Waanders, Pichai Raman, Mariarita Santi, Jena Lilly, Jennifer Mason, Javad Nazarian, "GENE-12. THE CHILDREN'S BRAIN TUMOR TISSUE CONSORTIUM (CBTTC) INFRASTRUCTURE FACILITATES COLLABORATIVE RESEARCH IN PEDIATRIC CENTRAL NERVOUS SYSTEM TUMORS," *Neuro Oncol*, vol. 19, no. suppl_4, p. iv20-iv21, 2017.

[4] R. L. Fleurence, L. H. Curtis, R. M. Califf, R. Platt, J. V Selby, and J. S. Brown, "Launching PCORnet, a national patient-centered clinical research network.," *J. Am. Med. Inform. Assoc.*, vol. 21, no. 4, pp. 578–582, 2014.

[5] OHSDI, "OMOP common data model; 2015." .

[6] J. B. Vaught, M. K. Henderson, and C. C. Compton, "Biospecimens and biorepositories: From afterthought to science," *Cancer Epidemiol. Biomarkers Prev.*, vol. 21, no. 2, pp. 253–255, 2012.

[7] J. J. Boklan, "Little patients, losing patience: pediatric cancer drug development.," *Mol. Cancer Ther.*, vol. 5, no. 8, pp. 1905–1908, 2006.

[8] A. Shaikh and A. Butte, "Collaborative Biomedicine in the Age of Big Data The Case of Cancer," *J. Med. ...*, 2014.

[9] G. N. Norén, T. Bergvall, P. B. Ryan, K. Juhlin, M. J. Schuemie, and D. Madigan, "Empirical performance of the calibrated self-controlled cohort analysis within temporal pattern discovery: Lessons for developing a risk identification and analysis system," *Drug Saf.*, vol. 36, no. SUPPL.1, 2013.

[10] S. E. Simpson, D. Madigan, I. Zorych, M. J. Schuemie, P. B. Ryan, and M. A. Suchard, "Multiple self-controlled case series for large-scale longitudinal observational databases," *Biometrics*, vol. 69, no. 4, pp. 893–902, 2013.

[11] M. Hauben, J. K. Aronson, and R. E. Ferner, "Evidence of Misclassification of Drug–Event Associations Classified as Gold Standard 'Negative Controls' by the Observational Medical Outcomes Partnership (OMOP)," *Drug Saf.*, vol. 39, no. 5, pp. 421–432, 2016.

[12] S. Schneeweiss, W. Eddings, R. J. Glynn, E. Patorno, J. Rassen, and J. M. Franklin, "Variable Selection for Confounding Adjustment in High-dimensional Covariate Spaces When Analyzing Healthcare Databases," *Epidemiology*, vol. 28, no. 2, pp. 237–248, 2017.

[13] I. Zorych, D. Madigan, P. Ryan, and A. Bate, "Disproportionality methods for pharmacovigilance in longitudinal observational databases," 2011.

[14] E. A. Voss, Q. Ma, and P. B. Ryan, "The impact of standardizing the definition of visits on the consistency of multi-database observational health research," *BMC Med. Res. Methodol.*, vol.

15, no. 1, p. 13, 2015.

[15] P. Zhang, F. Wang, J. Hu, and R. Sorrentino, "Towards personalized medicine: leveraging patient similarity and drug similarity analytics.," *AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci.*, vol. 2014, pp. 132–6, 2014.

[16] G. Hripcsak and D. J. Albers, "Next-generation phenotyping of electronic health records.," *J. Am. Med. Inform. Assoc.*, vol. 20, no. 1, pp. 117–21, 2013.

[17] R. L. Richesson, J. Sun, J. Pathak, A. N. Kho, and J. C. Denny, "A survey of clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods," *Artif. Intell. Med.*, vol. 71, no. 2016, pp. 57–61, 2016.

[18] G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, P. R. Rijnbeek, J. Van Der Lei, N. Pratt, G. N. Norén, Y. C. Li, P. E. Stang, D. Madigan, and P. B. Ryan, "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers," *Stud. Health Technol. Inform.*, 2015.

[19] Centers for Medicare & Medicaid Services, "CMS 2008-2010 Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF)," 2014. [Online]. Available: http://www.ltscomputingllc.com/downloads/.

[20] R. E. Murray, P. B. Ryan, and S. J. Reisinger, "Design and validation of a data simulation model for longitudinal healthcare data.," *AMIA Annu. Symp. Proc.*, vol. 2011, pp. 1176–85, 2011.

[21] "SNOMED CT The Global Language of Healthcare," 2015. [Online]. Available: http://ihtsdo.org/snomed-ct/. [Accessed: 07-Feb-2015].

[22] N. C. for H. Statistics, "International Classification of Diseases,Ninth Revision, Clinical Modification," 2013. [Online]. Available: https://www.cdc.gov/nchs/icd/icd9cm.htm.

[23] R. Kleinsorge, C. Tilley, and J. Willis, "Unified Medical Language System (UMLS)," *Encyclopedia of Library and Information Science*, 2002. [Online]. Available: https://www.nlm.nih.gov/research/umls/rxnorm/.

[24] Observational Health Data Science and Informatics, "Athena Standardized Vocabularies." 2016.

[25] P. L. Peissig, L. V Rasmussen, R. L. Berg, J. G. Linneman, C. a McCarty, C. Waudby, L. Chen, J. C. Denny, R. a Wilke, J. Pathak, D. Carrell, a N. Kho, and J. B. Starren, "Importance of multi-modal approaches to effectively identify cataract cases from electronic health records," *J Am Med Inf. Assoc*, vol. 19, no. 2, pp. 225–234, 2012.

[26] Wael Farhan Zhimu Wang Yixiang Huang Shuang Wang Fei Wang Xiaoqian Jiang, "A predictive model for medical events based on contextual embedding of temporal sequences," *J. Med. Internet Res.*, vol. 4, 2016.

[27] G. N. Norén, J. Hopstadius, A. Bate, K. Star, and I. R. Edwards, "Temporal pattern discovery in longitudinal electronic patient records," *Data Min. Knowl. Discov.*, no. August 2016, pp. 1–27, 2009.

[28] P. Lambin, E. Roelofs, B. Reymen, E. R. Velazquez, J. Buijsen, C. M. L. Zegers, S. Carvalho, R. T. H. Leijenaar, G. Nalbantov, C. Oberije, M. Scott Marshall, F. Hoebers, E. G. C. Troost, R. G. P. M. Van Stiphout, W. Van Elmpt, T. Van Der Weijden, L. Boersma, V. Valentini, and A. Dekker, "'Rapid Learning health care in oncology' - An approach towards decision support systems enabling customised radiotherapy," *Radiother. Oncol.*, vol. 109, no. 1, pp. 159–164, 2013.

[29] B. Craig and G. Han, "Simulating the contribution of a biospecimen and clinical data repository in a phase II clinical trial: a value of information analysis," *Stat. methods ...*, no. 813, 2013.

[30] J. T. Rich, J. G. Neely, R. C. Paniello, C. C. J. Voelker, B. Nussenbaum, and E. W. Wang, "A practical guide to understanding Kaplan-Meier curves," *Otolaryngol. - Head Neck Surg.*, vol. 143, no. 3, pp. 331–336, 2010.

[31] "Lifelines." Cam Davidson-Pilon, 2014.

[32] C. Liu, F. Wang, J. Hu, and H. Xiong, "Risk Prediction with Electronic Health Records: A Deep Learning Approach," *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '15*, pp. 705–714, 2015.

[33] J. Brownlee, *Machine Learning Mastery with Python*, vol. 53, no. 9. 2013.