Lister Hill National Center for
Biomedical Communications
An Intramural Research Division of the U.S. National Library of Medicine

Clinical Data Science

A Report to the Board of Scientific Counselors April 2017

Vojtech Huser, MD, PhD

Cognitive Science Branch

U.S. National Library of Medicine, LHNCBC
8600 Rockville Pike, Building 38A
Bethesda, MD 20894

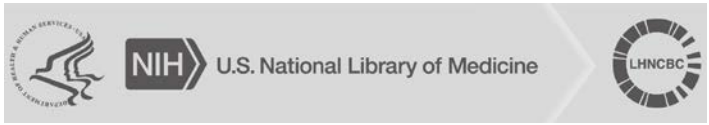

Table of Contents

Abstract

The focus of Dr. Vojtech Huser's research is on integrated health data repositories of routine healthcare data or clinical trial data. While state-of-the-art healthcare data warehouses offer the ability to generate some valid insights into health care processes, they still pose significant informatics challenges related to repository design and analytical interoperability of repositories. We present a research project organized along four dimensions: (1) generating insights from data repositories, (2) expertise with currently available repositories, (3) characterizing data repositories, and (4) integrating data repositories. We highlight three of our publications to illustrate various aspects of our research portfolio. The first study evaluated a data quality tool and qualitatively compared data quality assessment practices across seven organizations. The second study analyzed the suitability of a clinical research informatics standard for capturing research protocol and case report forms data. And finally, the third study examined drug treatment pathways in three chronic diseases (diabetes, hypertension, and depression) conducted within the Observational Health Data Sciences and Informatics consortium.

1 Introduction

In the past decade, large integrated data repositories with healthcare data have become crucial to studying healthcare services utilization and answering observational research questions.[1-3] Many isolated claims dataset were merged into much larger databases, such as Truven MarketScan Commercial Claims and Encounters database of more than 150 million covered patients. Similarly, distributed networks or repositories of administrative claims and electronic health records (EHRs) are also growing larger. Examples include the FDA Sentinel Network, the National Patient-Centered Clinical Research Network (PCORNet), Observational Health Data Sciences and Informatics (OHDSI) and Vizient University Health System Consortium. Informatics efforts to standardize disparate local storage models (including semantic integration) have culminated with the emergence of common data models (CDMs) that do not try to capture every possible detail, but target limited and pragmatic data standardization optimized for a research analytical scenario.[2]

Before health data repositories can be exploited for clinical insights, robust methods for their creation and maintenance need to exist,[4] standardization of their content in reference to standard terminologies and data elements needs to be performed,[5] their quality needs to be assessed,[6] and the appropriate statistical and data mining methods need to be researched.[7]

Data repositories of interest to our project include the NIH Biomedical Translational Research Information System (BTRIS); large claims repositories, such as the Medicare database; license-based resources, such as the Truven Commercial Claims and Encounters database; and datasets following a Common Data Model (CDM) within a research network, such as the Observational Health Data Sciences and Informatics (OHDSI) consortium. As they become available, patient-level data from clinical trials would also be of interest.[8]

2 Project objectives

The main objective of our project is to generate insights from integrated data repositories (IDRs). In order to support this objective, we need to have expertise with available repositories and address the informatics challenges of these repositories, namely characterizing their content and integrating these repositories.[4]

Generating insights from data repositories*: Using the repositories available to us, such as the CMS Virtual Research Data Center (VRDC), our objective is to answer concrete clinical questions, taking into account not only the features of the repository (e.g., size and data elements available), but also its limiting*

characteristics (i.e., data granularity and dataset population). We intend to explore both hypothesis- and data-driven approaches to investigating clinical questions. Additionally, by collaborating with external institutions with access to rich EHR data (e.g., Observational Health Data Science and Informatics collaborative; OHDSI), as we already have, we will be able to access a larger set of repositories and investigate a broader set of clinical questions. Finally, by co-investigating data repositories with domain experts from NIH, we will make it possible to test hypotheses arising from pre-clinical or basic biological research using the appropriate data repositories.

***Expertise with available repositories**:* Central to this project are integrated data repositories (IDRs), including datasets collected for healthcare and for research. Data collected for healthcare include data from Electronic Health Record (EHR) and administrative claims data. Data collected for research include patient-level data from human clinical trials. Such repositories are generally separate, but can be integrated. In terms of clinical datasets and increased availability of licensed data repositories and growth of research networks, researchers often need to select from several big data resources to license or to participate in a consortium for the collaborative analysis of clinical data across repositories. We have already acquired expertise with several available healthcare databases, including the GE Centricity outpatient dataset, the Truven Commercial Claims dataset, the Truven Medicare and Medicaid datasets, and the NIH intramural research data warehouse, called Biomedical Translational Research Information System (BTRIS). *We plan to acquire practical expertise with both clinical and research datasets whenever possible, and expertise through published reports otherwise.*

***Characterizing data repositories**:* Not all repositories can be used for testing a given clinical hypothesis.[9] For example, because it mostly records the prescriptions made to elderly patients, a drug dataset from Medicare Part D would not be appropriate for analyzing the frequency of teratogenic drugs prescribed to pregnant women. Issues including size, population characteristics, data quality, and, more generally, suitability of a given repository for a specific research question need to be considered carefully. Therefore, being able to characterize existing data repositories is an important aspect of this project. *In addition to methods, we intend to develop tooling (e.g., code libraries and packages) to support dataset characterization. We also expect to contribute to the development of best practices for repository creation and maintenance through dataset characterization.*

***Integrating data repositories**:* While it is valuable to analyze individual repositories, more benefits may come from integrating individual repositories into larger repositories, for example to support large-scale analyses, meta-analyses, and comparisons across repositories (e.g., for reproducibility testing).[2] Integrating repositories rests, in a large part, on the transformation of local repositories using a homegrown data model into repositories based on a common analytical model, supporting federated queries across repositories. The emergence of common data models (CDMs) for an analytic purpose reflects a vision for analytical interoperability.[2] Integrated data repositories not only share a harmonized information model, but also commit to target terminologies for coding biomedical entities (e.g., RxNorm for drugs, SNOMED CT for diagnoses, and LOINC for clinical observations). *We intend to keep contributing to the development of common data model, such as the Observational Medical Outcomes Partnership (OMOP) model. Moreover, we want to support the integration of several routine healthcare clinical repositories, research repositories, and repositories across healthcare and research in support of translational research. Finally, we want to investigate the role of emerging standards, such as Fast Healthcare Interoperability Resources (FHIR), for integrating repositories.*

3 Project Significance

In the past few years, NIH has increased its investment in data science, in particular through the Big Data to Knowledge (BD2K) program. NLM is also increasingly involved in data science activities, following the report from the NIH Advisory Committee to the Director in June 2015.[10] This report articulated the role of

NLM in the support and dissemination of EHR systems, healthcare research data repositories and the standards that underpin them. This project's research focus on data repositories (and their role in data sharing and data analysis) is therefore directly aligned with both this NIH-wide emphasis on data science and NLM's contribution.

***Generating insights from data repositories**:* Insights gained from analyzing large clinical repositories are expected to improve care. Broadly speaking, such insights embody the learning health system paradigm[11], i.e., a virtuous cycle in which the analysis of observational data collected for healthcare can provide evidence to improve care. Examples include pragmatic clinical trials,[12] supported through routine healthcare data, rather than dedicated clinical research data. More generally, comparative effectiveness studies provide patients with comparison of all current treatment choices as opposed to clinical trials focused on market approval for new drugs. Another example is computational drug repositioning (i.e., finding new uses for existing drugs, such as use of anticonvulsion drug *topiramate* for inflammatory bowel disease[13]), which can speed up drug development and reduce costs by taking advantage of already well-established knowledge of pharmacokinetics, pharmacodynamics and side-effects.

***Expertise with available repositories**:* Knowledge about existing data repositories – if possible hands-on, practical knowledge – is a prerequisite to being able to advise other researchers (e.g., from other NIH institutes and centers) interested in analyzing observational datasets to validate hypotheses arising from pre-clinical or basic biological research. More specifically, as a co-investigator in such studies, broad expertise with datasets, including knowledge of their features and limitations, will help us guide domain experts in selecting appropriate datasets for specific investigations.

***Characterizing data repositories**:* "Big data" is insufficient by itself if it is not also "good data". Emphasis on mere size of the dataset is downplayed, and progressively replaced by focus on quality. Improved methods for data characterization and automated tools for data quality assessments can lower the barrier for researchers to report on data quality in their publications[6] and help reduce the chances of reporting biased or false observational evidence.[7] Characterizing data repositories also underpins the ability to pick the most appropriate dataset for a given investigation and leads to an improved research output.

***Integrating data repositories**:* Researchers have already started to report clinical insights derived from large integrated repositories of clinical data. For example, the association of lower short- and long-term mortality with overweight and obesity in adult intensive care unit patients[14] or repository support for running a pragmatic randomized clinical trial determining the optimal dose of aspirin for atherosclerosis prevention.[12] The benefit of improved integration of human clinical trials data lies in improved ability to conduct meta-analyses (e.g., analysis of collection of Alzheimer disease trials)[15] and possibly improved design of patient-level trial results repositories that undergo rapid evolution.[16]
Key to these research activities is a common data model supporting analytics across repositories, including both a harmonized information model and standard terminologies. As common data models become more sophisticated, they will support a larger set of analyses across a wider variety of datasets. And as the adoption of common data models increases, integration of new datasets will become easier. For example, the adoption of the OMOP CDM by the *All Of Us* one-million patient cohort will make this dataset directly compatible with the many repositories of observational data already conformant with the OMOP model.

Contribution
Over the past few years, we have made a number of contributions to various aspects of this project, published in the scientific literature. Our publications are listed below, organized by the underlying clinical dataset investigated. Three of these publications[17-19] are highlighted in the next section.

- Datasets in the Research Lab of the Innovation in Medical Evidence Development and Surveillance (IMEDS) program of the Reagan-Udall Foundation for the Food and Drug Administration (FDA): predominantly claims datasets created by (and licensed from) IBM Truven Healthcare[20-23]
- Biomedical Translational Research Information System (BTRIS): NIH Clinical Center integrated data repository for intramural clinical trials data[24-28]
- Collaboration with sites within the Observational Health Data Sciences and Informatics (OHDSI) Consortium that adopted the OMOP Common Data Model[17,19,28]
- Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II and III: collection of de-identified data of from intensive care unit patients[29,30]
- Center for Medicare and Medicaid (CMS) Virtual Research Data Center: Medicare and Medicaid claims data[31] (in collaboration with other LHNCBC investigators)

4 Highlighted publications

In this section, we briefly present three recent investigations illustrating various aspects of our research portfolio. The first two reflect informatics dimensions of our research project, namely data quality (under ***Characterizing data repositories***) and research data integration (under ***Integrating data repositories***), while the last one illustrates insights gained from clinical repositories (under ***Generating insights from data repositories***).

4.1 Research focus #1: Data Quality

4.1.1 Research portfolio

The validity of observational research based on large healthcare repositories, such as FDA's Sentinel initiative or CMS's virtual research data center, depends on the quality of the data in those repositories. While data quality can be defined in many different ways, our broad definition includes aspects, such as completeness and conformance with the requirements of the underlying data model. Data quality assessment (DQA) tools have started to emerge for several common data models.[32] In 2015, an informatics Data Quality Collaborative was formed and has produced a number of data quality related initiatives.[33]

Our work on data quality includes the following studies: (1) In 2014 and 2015, we conducted a pilot study that assessed the size and completeness of 17 large datasets (Iris OHDSI tool)[34] and several studies related to the size and representativeness of deceased patient subsets within a data repository.[27,35,36] (2) In 2016, we conducted an evaluation study of the Achilles Heel tool, which is presented below. (3) Later in 2016 we initiated a follow-up data quality study that extends Achilles Heel's functionality using data from several OHDSI data partners (currently ongoing).[37]

4.1.2 Multi-site Evaluation of a Data Quality Tool for Patient-Level Clinical Datasets

Huser V, DeFalco FJ, Schuemie M, Ryan PB, Shang N, Velez M, Park RW, Boyce RD, Duke J, Khare R, Utidjian L, Bailey L. Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Data Sets. EGEMS (Wash DC). 2016 Nov 30;4(1):1239. doi: 10.13063/2327-9214.1239. PMID: 28154833

This study is an evaluation conducted in 2016 of a software tool developed by the OHDSI Consortium, called Achilles Heel. In this study, we compared the output from a DQA for 24 datasets (originating from seven sites). We fully designed the study and conducted the necessary data collection and analysis. The impact of this work was the subsequent addition of several new features, measures and rules to this data quality tool (used by many institutions that maintain a healthcare data repository) and the first published comparison of data quality outputs for a large set of healthcare datasets.

Methodology
The Achilles tool employs a two-step approach. Step 1 performs over 170 different pre-computations that characterize the data. Step 1 results in thousands of aggregated counts stratified by different parameters. To preserve privacy, aggregations that produce small counts (typically under 10 patients per cell count) are not exported. To facilitate cross-dataset and even cross-model interoperability, each pre-computed analysis is assigned an identifier (analysis_id) and a short description of the pre-computed analysis. For example, "715: Distribution of days_supply by drug_concept_id" or "506: Distribution of age at death by gender". The Achilles data model allows storage of the results of all pre-computations in a single table (achilles_results) organized by up to five analysis dimensions (called strata within Achilles). Step 1 pre-computed analyses are driven not only by data quality questions, but also by data visualization needs of a data exploration application (either AchillesWeb or Data Sources tab within the Atlas[38] application). The step-1 pre-computations allow fast data density visualizations and tabular views of data availability by data domain in general and by individual event concept (such as individual diagnosis, procedure, medication, laboratory result, or observation; sometimes further stratified by age decile or gender).

Once step-1 pre-computations are completed, step 2 consists of the execution of data quality rules and some optional data transformation procedures. Step 2 only uses aggregated counts created in step 1 and does not require any patient-level data. Step-2 data quality rules can be classified into two categories: (1) *CDM conformance rules* that check whether the OMOP CDM specifications are being followed (for example, whether valid SNOMED CT codes are used to capture diagnoses); and (2) *data quality rules* that investigate data completeness (are data values present?), data plausibility (are data values believable?) and general data conformance (adherence to an expected data format). The data quality part of the Achilles application is formally referred to as Achilles Heel. At the time of our evaluation, the rules originated from the OHDSI community (including from an earlier OMOP tool called Oscar). An important part of our Achilles Heel evaluation study was, in fact, an assessment of how good and how useful this rule set is.

Results
Our study demonstrated that 24 datasets from 7 sites (converted to the OMOP data model) can be analyzed by a common software tool that examines data quality. Compared with previous approaches that typically employ site-specific data models and site-specific DQA scripts, demonstrating execution of a common data quality assessment framework across multiple datasets represents a significant advance in data quality evaluation.

The data partner sites in our study included single academic medical centers, a pharmaceutical industry research department, a clinical data research network and a research program of a medical research foundation. Most sites provided data quality assessments for a single dataset, while three sites provided data for multiple datasets.

The median number of errors identified by the Achilles Heel tool was 17 (in each dataset), while the total number of distinct errors analyzed was 982 (pooled error data across all datasets). Our study provided an overview of the most common data quality errors identified in at least 10 datasets (see Table 1). We also analyzed the overlap among data quality rules. For example, an implausible entry in birth year will trigger several individual DQA rules that target a given domain, such as medication, condition or procedure, resulting in the same underlying error being reflected across several domains (e.g., medication data prior birth; procedure data prior birth, etc.).

Error ID	Count of datasets with error	Count of all error instances	Error description
103	15	n/a	Distribution of age at first observation period; age should not be negative
206	13	18	Distribution of age by visit_concept_id; age should not be negative
406	13	31	Distribution of age by condition_concept_id; min(age) should not be negative
600	13	14	Number of persons with at least one procedure occurrence, by procedure_concept_id; concepts in data are not in correct vocabulary (CPT4; HCPCS,ICD9P)
717	12	3173	Distribution of quantity by drug_concept_id; max(quantity) should not be > 600
114	11	n/a	Number of persons with observation period before year-of-birth; should not be > 0
410	11	n/a	Number of condition occurrence records outside valid observation period; should not be > 0

Table 1: Subset of the most common errors found

In addition to the quantitative and descriptive comparison of Achilles Heel outputs across many datasets, the study included a qualitative survey of how data quality assessments are executed at each site. Most sites executed the Achilles Heel tool after their first data conversion to the common data model. With regards to the impact of the tool, most sites found Achilles Heel output helpful in discovering extract-transfer-load (ETL) errors. Many sites used this information to improve their ETL code. The intent was to eliminate all or some of the Achilles Heel errors and warnings by revising the data transformation code. When asked how frequently CDM datasets are refreshed, the answers ranged from never (static CDM data; 1 site) to biweekly, with most sites refreshing it once a year. The amount of resources dedicated to initial data quality evaluation and ongoing data quality monitoring also varied widely. At one site, where a CDM dataset is tied to a health information exchange, data quality is monitored by a committee of five people that meets monthly. All sites used additional data quality tools besides Achilles Heel. Two sites routinely compare the overall data volume in source and converted data and investigate significant variation in volume trends over time.

Conclusion

Computational tools for automated data quality assessment represent an important emerging focus for healthcare research institutions and distributed research networks. Despite the fact that data quality assessment is task-dependent ("fitness of data for what?"), many existing tools and efforts indicate that some general DQA rules and methodologies indeed exist. The current use of the tool by the community and feature requests (submitted to the Achilles Heel GitHub issue tracker) indicate that automated computational approaches to DQA are highly requested by researchers. Our existing and future efforts focus predominantly on this computerized and general DQA scenario.

4.2 Research focus #2: Integration of research data with routine healthcare data

4.2.1 Research portfolio

With wider availability of larger repositories of routine healthcare data, many scientists have turned their attention to the aggregate of all data collected by human interventional and observational trials as the next source of rich clinical data.[39] Recent growth of patient-level research data repositories (such as Database of Genotypes and Phenotypes – dbGaP, Project DataSphere, TrialShare, DataShare platform of the National Institute of Drug Abuse, and the DASH database of the National Institute of Child Health and Human Development) has highlighted the need to standardize how researchers access de-identified clinical trial data.[40,41] In addition to data themselves, computable formats for capturing the study context, such as the study design, study protocol, informed consent and data collection instruments (Case Report Forms; CRFs) are also of interest to many clinical research informaticians.

Considering this challenge and the fact that the NIH Clinical Center is a major research site for hundreds of intramural clinical trials, we conducted in 2015 an evaluation of the Operational Data Model (ODM) standard created by the Clinical Data Interchange Standards Consortium (CDISC).[18] ODM is a standard for capturing study context and data in an XML-based computable format. For example, since 2016, the REDCap data capture system allows export of complete study metadata and data in the ODM format.[42] Our team's role was to fully design the evaluation approach and execute all the study steps.

In addition to the study presented below, the following activities also reflect our involvement with clinical research issues: (1) In 2016, we co-authored a comprehensive literature review of the past and current use of the ODM standard.[43] (2) We organized a panel at the 2016 AMIA Annual Symposium in the clinical research informatics track. Our panel discussed the state of the art of common data elements and their use in patient-level trial data repositories.[41] (3) In 2014, we organized a tutorial on CDISC standards at the AMIA Joint Summits.[44]

4.2.2 Representation of clinical research study protocol and case report forms

Huser V, Sastry C, Breymaier M, Idriss A, Cimino JJ. Standardizing data exchange for clinical research protocols and case report forms: An assessment of the suitability of the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM). J Biomed Inform. 2015 Oct;57:88-99. doi: 10.1016/j.jbi.2015.06.023. Epub 2015 Jul 15. PMID: 26188274

Methodology

The ODM standard evaluation study[18] used an intramural clinical trial as a case study to analyze the standard's strengths and weaknesses. The institutional background was important to consider and analyze. The NIH intramural research program typically has over 2300 active studies. On average, 231 new protocols are initiated every year. Since 1953, the NIH intramural research program has registered a total of 8017 completed studies and maintains a repository[45] of data collected in those studies. We investigated the study metadata needs of all relevant research systems, including a web-based protocol authoring system, electronic Institutional Review Board (eIRB) systems, a protocol management system, research electronic data capture systems, an electronic health record (EHR) system, and the research data repository/warehouse. We evaluated the best standard and mechanism that would support moving protocol data and metadata across these systems (see Figure 1)

In addition to analyzing the needs of the various research IT systems, we have divided the study life-cycle into several stages (considering primarily a data capture system perspective).

1. *Study drafting:* The goal of the study drafting stage is to generate study registration information and the full protocol that describes the steps and procedures of the protocol. These data are needed to either support a study funding decision process or to communicate the study to the larger research team. Internal systems and internal representation formats are involved during this stage.
2. *Study registration:* Federal and internal NIH policies require registration of the trial in the ClinicalTrials.gov registry, which uses an XML schema based standard.
3. *Study initiation and execution:* The study initiation stage starts upon IRB approval and requires the existence of one or more CRFs. In a multi-site trial, each site may be using a different electronic data capture system and the ability to import and export CRFs is an important function of a protocol representation format and can save time spent on duplicate entries of the same CRFs into multiple systems.
4. *Study termination:* The study termination phase begins when the last patient's data are collected. During the study termination phase, a data format that can capture study results is needed, in addition to mere capture of the study protocol. Export of data collected during the study for

statistical or other analysis is the most important step in the termination phase. A standardized export of study data is helpful to statisticians, who deal with multiple studies, or to data repositories responsible for long-term storage of clinical study data.

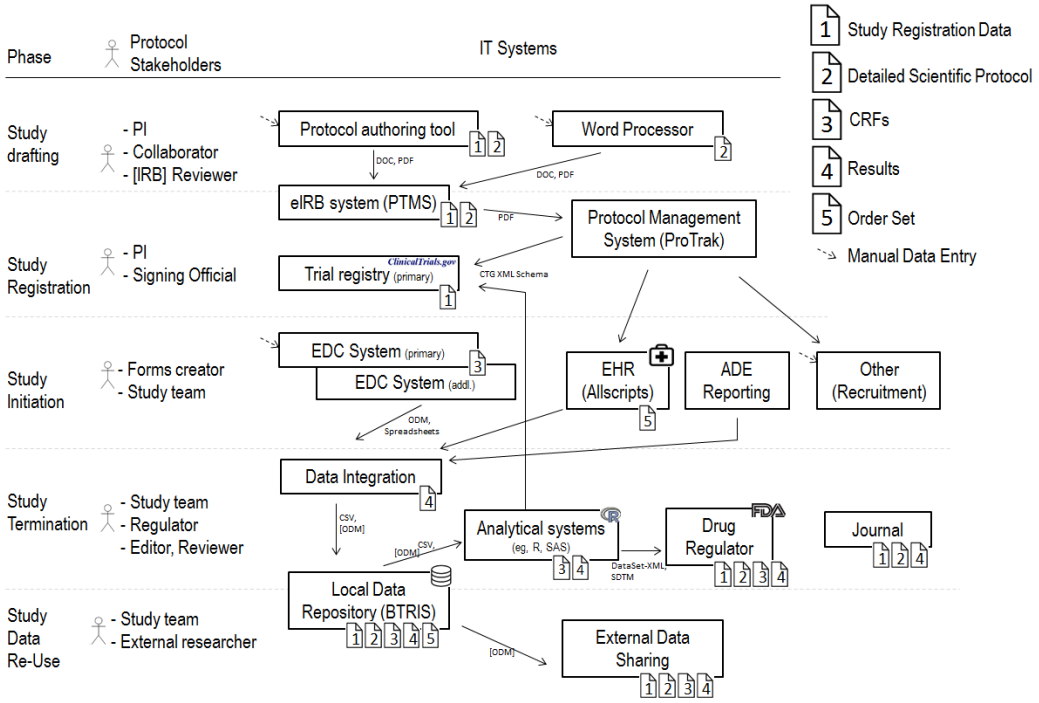

Figure 1: Overview of research systems at the NIH intramural research program

Results

The project followed a case study approach where a single trial was followed though all the stages defined above. The selected trial was NCT00001848: "The Safety and Effectiveness of Surgery With or Without Raloxifene for the Treatment of Pelvic Pain Caused by Endometriosis". The trial examined whether 6 months of raloxifene was effective in the treatment of chronic pelvic pain in women with endometriosis. Women with chronic pelvic pain underwent laparoscopy and were randomly allocated to raloxifene or placebo. A second laparoscopy was performed at 2 years, or earlier, if pain returned.[46]

The raloxifene trial had 25 forms defined with a total number of 686 questions. The trial had 10 study milestones defined (e.g., Screening, Baseline, 3 Month, 6 Month, 9 Month, 24 Month) with several case report forms assigned for repeated collection at multiple milestones.

Our published results included all possible computable representation formats relevant to various stages. (1) *RaloxifeneStudy-Draft.ODM.XML* file capturing study drafting stage data in ODM XML format; (2) *RaloxifeneStudy-Protocol-outline.docx* files that follow the CDISC Protocol Representation Model outline template and *RaloxifeneStudy-SDTM.csv* file with the corresponding Study Data Tabulation Format (SDTM) data elements; (3) *RaloxifeneStudy-registration.ODM.XML* file aiming at the requirements for registration in the ClinicalTrials.gov registry; and (4) two *RaloxifeneStudy-SurgicalFindings-Form.ODM.xml* files aimed at capturing CRF data.

Using the context of the raloxifene trial, we discussed in detail the strengths and weaknesses of existing standards (focusing in greater detail on the ODM standard) and informatics challenges with clinical trials data integration and management. The impact of the study was the creation of draft specifications for a new

version of ODM (v2.0) by CDISC (with our input) and increased harmonization of the ODM standard with trial registries via an ODM extension.[47,48]

Conclusion
The *All Of Us* one-million patient cohort program of the Precision Medicine Initiative is an exemplary demonstration of the recent trend of observational studies that combine long-term routine healthcare data (originating from EHR or claims data) with research-specific data (originating from research case report forms). The informatics vision is to seamlessly integrate routine healthcare data with research data. This requires good understanding of both of those parallel data domains. Our efforts to understand and properly store case report form data from clinical trials data (with the background knowledge of routine healthcare data repositories) aim at bridging this existing gap.

4.3 Research focus #3: Insights from data repositories

4.3.1 Research portfolio

The emergence of common data models together with agreement on target terminologies (e.g., RxNorm for drug exposure data) have created an opportunity to perform analyses across multiple healthcare datasets using a single analysis script. In 2016 we participated in one such study that analyzed treatment pathways in diabetes mellitus, hypertension and depression within the OHDSI consortium.[19] This study is presented below. Our team's role in the study was to execute the analytical code created by two of the study co-authors on a predominantly outpatient dataset (within the IMEDS research program) and contribute to the presentation of the multi-site aggregated data.

Along the same lines, the following analyses were also carried out: (1) In 2016, we analyzed drugs used during pregnancy[21] with respect to multiple risk classification schemes (Briggs classification and legacy FDA classification) in the Truven Commercial Claims and Truven Medicaid datasets. These datasets were available to us via the IMEDS research program of the Raegan-Udall Foundation for the FDA; (2) In 2015, we analyzed genomic testing events[20] that utilized recently introduced molecular pathology procedural codes for genomic testing, such as MLH1 gene sequencing for Lynch syndrome, and pharmacogenomics testing, such as CYP2C19 genotyping analysis for common variants.

4.3.2 Characterizing treatment pathways at scale using the OHDSI network

Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, Suchard MA, Schuemie MJ, DeFalco FJ, Perotte A, Banda JM, Reich CG, Schilling LM, Matheny ME, Meeker D, Pratt N, Madigan D. Characterizing treatment pathways at scale using the OHDSI network. Proc Natl Acad Sci U S A. 2016 Jul 5;113(27):7329-36. doi: 10.1073/pnas.1510502113. Epub 2016 Jun 6. PMID: 27274072

Methodology
The treatment pathway study[19] analyzed the variability of pharmacological treatment interventions over three years across three diseases. The patient inclusion criteria were exposure to an antidiabetic, antihypertensive, or antidepressant medication, as well as presence of at least one diagnostic code for the corresponding disease (type 2 diabetes mellitus, hypertension, or depression). There were additional temporal and data requirements for one year prior and three years after the index date (date of the first exposure to the qualifying drug class). The cohort definition also included exclusion criteria based on diagnostic data, such as exclusion of schizophrenia patients within the depression cohort.

The study analyzed the sequence of medications with some limiting assumptions. Patients who switched off from a medication and back to it were only recorded for the first exposure. The sequence approach did not distinguish switching medications from adding medications. Sequences were limited to 20 medications.

A total of eleven datasets representing a total of 255 million patients were analyzed, with our team contributing aggregated study data for the de-identified GE Centricity outpatient EHR dataset (consisting of 33 million patients) included in the IMEDS program of the Reagan-Udall Foundation for the FDA.

Results

The treatment pathways for the three diseases demonstrated great heterogeneity in terms of preferred first-line therapy (see figure 2). Whereas in diabetes, 75% of the patients used the most common first-line therapy (*metformin*), in depression, only 17% of the patients used the most common first-line therapy (*citalopram*). In addition to aggregate ingredient-based analysis, differences among datasets (that originated from different geographic regions) were apparent from the results, such as *gliclazide* (diabetes drug) used only in the United Kingdom. Abstracted sequences also allowed assessment of monotherapy (defined within this study as use of a single medication in the entire three-year window) versus using multiple drugs or drug-classes. Besides clinical results, an interesting informatics result was the fact that that 10% of diabetic patients, 24% of hypertension patients and 11% of depression patients followed a treatment pathway that was unique in the entire collection of 255 million patients. That means that if a patient would be asking what patients are similar to them, the answer would be no one (if entire drug sequence match would be required).

The impact of the study was in demonstrating feasibility of studies across multiple continents and sampling from a very large patient population. Another important study impact was revealing a variety of drug ingredients used worldwide and the need to use a drug terminology that has international reach (which eventually led to creation of OHDSI's extension to RxNorm).

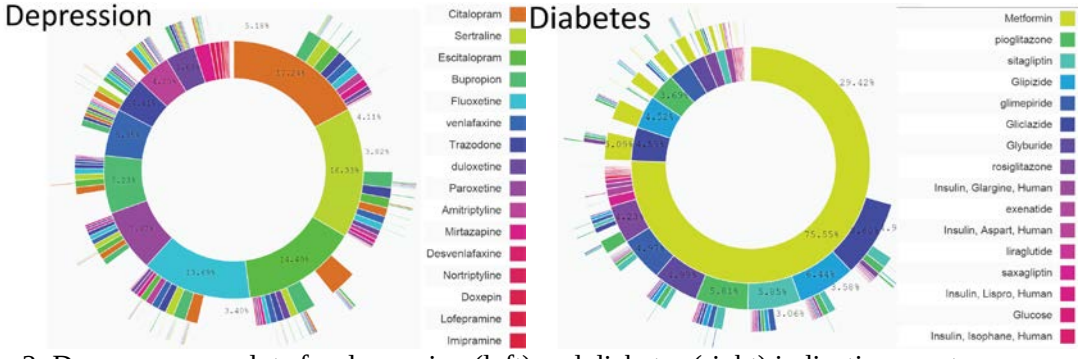

Figure 2: Drug sequence plots for depression (left) and diabetes (right) indicating greater consensus on first line treatment in diabetes

Conclusion

This study is an example of analysis for which existing large healthcare datasets offer sufficient data and support meaningful comparisons between different sites or geographic regions. Whereas advancing data quality often leads to demonstration of a "data absence problem", it is important to pursue example studies that clearly demonstrate the current value and possibilities of healthcare big data with a "data presence problem" mindset. This study also exemplifies a data science trend of big questions that focus on analytical methods and data models that operate across a range of medical domains. The clinical discovery dimension of our project intends to follow these two high-level principles.

5 Summary and Future Plans

We have presented a vision for clinical data science research at the LHNCBC, with two major objectives: to generate new clinical insights from repositories and to address informatics challenges with repositories.

We have shown our past contributions to various aspects of our project, especially data quality and clinical insights gained from large clinical repositories.

Our plans for the near future include additional data quality studies and developing new data quality measures. For example, we are working on the applicability of the Achilles data quality rules to data in other non-OMOP formats. We are also considering new measures that would aggregate data across a distributed research network and provide assessment of data size and quality at the network level. In terms of research data integration, we continue to evaluate emerging clinical research informatics standards (e.g., FHIR). In collaboration with other divisions of NLM, we have also started working on automated and manual annotation of research data repositories with Common Data Elements (CDEs). We hope to develop guidelines for CDE annotation that would prevent content duplication and aid in data discovery to help researchers reuse datasets. We want to keep exploring new datasets and repositories (e.g., use the Symphony Health dataset previously acquired by LHNCBC as a pilot environment for drafting analyses). Additionally, the Optum dataset may become available to us via a pre-established contractual framework set up by the IDEA lab at the Department of Health and Human Services. In terms of use of repositories to generate discoveries, we have started identifying published observational clinical research findings that can be replicated for the purpose of validating such findings on different datasets. We plan to model and execute clinical analyses against datasets directly available to us, especially the Medicare claims data licensed by LHNCBC via the CMS VRDC platform. We will also work on hypothesis generation using data-driven approaches.

This project would also directly contribute to a possible NIH Center for Observational Investigations based at NLM in partnership with intramural researchers across NIH institutes and centers. Such a center would greatly facilitate the investigation of hypotheses stemming from basic, pre-clinical research and provide a framework for co-investigation between NLM data scientists and NIH biologists and medical experts.

6 Acknowledgements

We would like to thank NLM colleagues that work with the CMS Virtual Research Data Center database for their advice about this resource (Clem McDonald, Seo Hyon Baik, Fabricio Kury). We would like to thank NIH Clinical Center (CC) BTRIS team for their help with understanding data integration issues at the NIH CC (EHR system data combined with electronic data capture systems for clinical trials), including former BTRIS chief James Cimino. We would like to thank NIH CC Department of Clinical Research Informatics team for help with interpreting data structures used by Allscripts Sunrise Clinical Manager (EHR system used at NIH CC). We would like to thank Chandan Sastry, Matt Breymaier from the Clinical Trials Database team for their help in understanding and extending a home-grown electronic data capture system developed by the National Institute of Child Health and Human Development. We would like to thank student Alok Sagar and special volunteer Yohan Sumithipala.

7 Glossary

BD2K	Big Data to Knowledge (NIH research initiative)
BTRIS	Biomedical Translational Research Information System
CDISC	Clinical Data Interchange Standards Consortium
CDM	Common Data Model
CMS	Center of Medicare and Medicaid Services
DQA	Data Quality Assurance
EHR	Electronic Health Record
ETL	Extract, Transform, Load (set of processing steps for data manipulations)
FDA	Food and Drug Administration
IDR	Integrated Data Repository
IMEDS	Innovation in Medical Evidence Development and Surveillance (research program)
IRB	Institutional Review Board
LOINC	Logical Observation Identifiers Names and Codes
MIMIC	Multiparameter Intelligent Monitoring in Intensive Care (clinical dataset name)
ODM	Operational Data Model (a standard defined by CDISC in existence since 1999)
OHDSI	Observational Health Data Sciences and Informatics
PCORNet	Patient Centered Outcomes Research Network

8 References

1. Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-World Evidence - What Is It and What Can It Tell Us? *N Engl J Med* 2016; **375**(23): 2293-7.
2. Weng C, Kahn MG. Clinical Research Informatics for Big Data and Precision Medicine. *Yearb Med Inform* 2016; (1): 211-8.
3. Embi PJ, Payne PR. Evidence generating medicine: redefining the research-practice relationship to complete the evidence cycle. *Med Care* 2013; **51**(8 Suppl 3): S87-91.
4. Mackenzie SL, Wyatt MC, Schuff R, Tenenbaum JD, Anderson N. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. *J Am Med Inform Assoc* 2012; **19**(e1): e119-e24.
5. Richesson RL, Chute CG. Health information technology data standards get down to business: maturation within domains and the emergence of interoperability. *J Am Med Inform Assoc* 2015; **22**(3): 492-4.
6. Kahn MG, Brown JS, Chun AT, et al. Transparent reporting of data quality in distributed data networks. *EGEMS (Wash DC)* 2015; **3**(1): 1052.
7. Madigan D, Stang PE, Berlin JA, et al. A systematic statistical approach to evaluating evidence from observational studies. *Annual Review of Statistics and Its Application* 2014; **1**: 11-39.
8. Hudson KL, Collins FS. Sharing and reporting the results of clinical trials. *Jama* 2015; **313**(4): 355-6.
9. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *Journal of biomedical informatics* 2013; **46**(5): 830-6.
10. NIH. NIH Advisory Committee to the Director: NLM Working Group: Final Report on mission, organization, programmatic priorities and strategic vision for NLM (2015-06-11). 2015. https://acd.od.nih.gov/reports/Report-NLM-06112015-ACD.pdf (accessed Nov 22 2016).
11. Smith MD, Saunders RS, Stuckhardt L, McGinnis JM, Institute of M, Committee on the Learning Health Care System in A. Best care at lower cost : the path to continuously learning health care in America. 2013.
12. Johnston A, Jones WS, Hernandez AF. The ADAPTABLE Trial and Aspirin Dosing in Secondary Prevention for Patients with Coronary Artery Disease. *Current cardiology reports* 2016; **18**(8): 81.
13. Dudley JT, Sirota M, Shenoy M, et al. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Science translational medicine* 2011; **3**(96): 96ra76-96ra76.
14. Abhyankar S, Leishear K, Callaghan FM, Demner-Fushman D, McDonald CJ. Lower short- and long-term mortality associated with overweight and obesity in a large cohort study of adult intensive care unit patients. *Critical care (London, England)* 2012; **16**(6): R235.
15. Geifman N, Brinton RD, Kennedy RE, Schneider LS, Butte AJ. Evidence for benefit of statins to modify cognitive decline and risk in Alzheimer's disease. *Alzheimer's research & therapy* 2017; **9**(1): 10.
16. Bertagnolli MM, Sartor O, Chabner BA, et al. Advantages of a Truly Open-Access Data-Sharing Model. *N Engl J Med* 2017; **376**(12): 1178-81.
17. Huser V, DeFalco F, Schuemie M, et al. Multi-site Evaluation of a Data Quality Tool for Patient-Level Clinical Datasets. *eGEMs (Wash DC)* 2016.
18. Huser V, Sastry C, Breymaier M, Idriss A, Cimino JJ. Standardizing data exchange for clinical research protocols and case report forms: An assessment of the suitability of the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM). *Journal of biomedical informatics* 2015.
19. Hripcsak G, Ryan PB, Duke JD, et al. Characterizing treatment pathways at scale using the OHDSI network. *Proceedings of the National Academy of Sciences of the United States of America* 2016; **113**(27): 7329-36.
20. Huser V. Process Mining of Growing Adoption of Genomic Precision Medicine Testing Using Commercial Claims and Encounters Database. *Proc AMIA Symp* 2015.
21. Dhombres F, Huser V, Rodriguez L, Bodenreider O. Assessing the potential risk in drug prescriptions during pregnancy. AMIA Annu Symp Proc 2016 *Proc AMIA Symp* 2016.
22. Huser V. Using an average patient record archetype concept to compare five big data healthcare datasets with claims and electronic health record data. *NIH Research Festival* 2015.
23. Huser V. Analysis of drug use by dose form in large healthcare databases: Data granularity issues and CDM considerations. *2016 OHDSI Symposium* 2016.
24. Cimino JJ, Ayres EJ, Remennik L, et al. The National Institutes of Health's Biomedical Translational Research Information System (BTRIS): design, contents, functionality and experience to date. *Journal of biomedical informatics* 2014; **52**: 11-27.
25. Huser V, Fung KW, Cimino JJ. Natural Language Processing of Free-text Problem List Sections in Structured Clinical Documents: a Case Study at NIH Clinical Center. *AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science* 2014.
26. Rodriguez L, Huser V, Bodenreider O, Cimino J. Automatic coding of Free-Text Medication Data recorded by Research Coordinators *AMIA Annu Symp Proc* 2014.
27. Huser V, Cimino JJ. Don't take your EHR to heaven, donate it to science: legal and research policies for EHR post mortem. *J Am Med Inform Assoc* 2014; **21**(1): 8-12.
28. Huser V, Cimino JJ. Desiderata for healthcare integrated data repositories based on architectural comparison of three public repositories. *AMIA Annu Symp Proc* 2013; **2013**: 648-56.
29. Bayzid S, Huser V, Ghosh J. Conversion of MIMIC to OHDSI CDM. *2016 OHDSI Symposium (Sep 22, 2016)* 2016.

30. Kury FS, Huser V, Cimino JJ. Reproducing a Prospective Clinical Study as a Computational Retrospective Study in MIMIC-II. *AMIA Annu Symp Proc* 2015; **2015**: 804-13.
31. Kury F, Huser V. Converting the data in the U.S. CMS Virtual Research Data Center to the OHDSI Common Data Model version 5. *OHDSI Symposium, October 2015* 2015.
32. Kahn MG, Callahan TJ, Barnard J, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Wash DC)* 2016; **4**(1): 1244.
33. Health A. Data Quality Collaborative. 2015. http://repository.academyhealth.org/dqc/ (accessed May 15 2015).
34. Huser V, Suchard MA. Size comparison of 17 CDM datasets using IRIS tool. *OHDSI Symposium* 2015.
35. Huser V, Kayaalpm M, Dodd ZA, Cimino JJ. Piloting a Deceased Subject Integrated Data Repository and Protecting Privacy of Relatives. *AMIA Annu Symp Proc* 2014; **2014**.
36. Huser V, Miller A, Vawdrey DK. Evaluating the size of deceased patient EHR research data sets: A multi-year trend analysis. *AMIA Annu Symp Proc* 2014; **2014**.
37. Huser V. OHDSI Data Quality study. 2016. http://www.ohdsi.org/web/wiki/doku.php?id=research:dqstudy (accessed Jan 2 2017).
38. OHDSI. ATLAS – A unified interface for the OHDSI tools. 2016. https://www.ohdsi.org/atlas-a-unified-interface-for-the-ohdsi-tools (accessed March 20 2017).
39. Richesson RL, Horvath MM, Rusincovitch SA. Clinical research informatics and electronic health record data. *Yearb Med Inform* 2014; **9**: 215-23.
40. Geifman N, Bollyky J, Bhattacharya S, Butte AJ. Opening clinical trial data: are the voluntary data-sharing portals enough? *BMC Med* 2015; **13**: 280.
41. Huser V, Sincan M, Bloomberg D, Hess R. Standardizing Research Common Data Elements: Initiatives, Exchange Formats and Current Use by Patient-Level Trial Results Databases (didactic panel). ***Proc AMIA Symp*** 2016.
42. REDCap. CDISC ODM Compatibility in REDCap. 2016. https://projectredcap.org/cdisc.php (accessed March 18 2017).
43. Hume S, Aerts J, Sarnikar S, Huser V. Current applications and future directions for the CDISC Operational Data Model standard: A methodological review. *Journal of biomedical informatics* 2016; **60**: 352-62.
44. Huser V, Schaefer P. CDISC Standards in Clinical Research Informatics (tutorial). *Proc AMIA joint summit* 2014.
45. Cimino JJ, Ayres EJ, Remennik L, et al. The National Institutes of Health's Biomedical Translational Research Information System (BTRIS): Design, contents, functionality and experience to date. *Journal of biomedical informatics* 2013.
46. Stratton P, Sinaii N, Segars J, et al. Return of chronic pelvic pain from endometriosis after raloxifene treatment: a randomized controlled trial. *Obstetrics and gynecology* 2008; **111**(1): 88-96.
47. Huser V, Cimino JJ. Using CDISC Standards to Create Formal and Computable Representations of Human Clinical Research Protocols. *NIH Research Festival* 2014.
48. CDISC. Clinical Trial Registry XML (CTR-XML). 2016. https://www.cdisc.org/standards/foundational/ctr-xml (accessed Jan 20 2017).