Evaluating Different Approaches to Simplifying Data Access for Clinical Users

Denise Lin

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington
2018

Committee:

Adam Wilcox
John Gennari

Program Authorized to Offer Degree:
Biomedical Informatics and Medical Education

University of Washington

Abstract

Evaluating Different Approaches to Simplifying Data Access for Clinical Users

Denise Lin

Chair of the Supervisory Committee:

Adam Wilcox

Department of Biomedical Informatics and Medical Education

Researchers have difficulty in accessing health care data for multiple different reasons. Although some technologies, like i2b2, have been developed and evaluated to overcome these difficulties, limitations and challenges remain. In addition, there are limited comparisons among query tools, such that users do not have an understanding of which tool works best in which situation. Studies that evaluate and compare such technologies to both guide users and improve tools are needed.

To evaluate and compare between two self-service query tools – LEAF and I2b2, and one common data model – OMOP, I selected different representative query questions that are commonly asked by researchers based on externally-defined query categories; quality measurement, based on observational EHR research studies, and representative queries made by users to the analytics team in our organization. Most of the query questions included four main concepts: the diagnosis, patient age, length of stay, and measurement period. I used the

three different query tools to answer all query questions. I then analyzed the results to determine which is the best approach to increase data access using the two main determinants in Technology Acceptance Model (TAM): perceived usefulness (PU) and perceived ease of use(PEOU). LEAF, developed by the University of Washington, returned as the best performer among the three query tools due to its flexibility, perceived ease of use, and perceived usefulness. Researchers can easily explore its customized features without needing a programming background. The development of these technologies could reduce the challenges for data access in health care.

Table of Contents

Background & Significance

1.1 Introduction

Accessing healthcare data for researchers is demanding and few tools are developed to simplify data access. Although self-service query tools are still developing, access to data is challenging for various reasons. The secondary use of healthcare data is one of the most important uses in clinical care. However, there are huge barriers to get access to the data they need, such as restricted permissions, limited comparison among query tools, and lack of integration of information systems across multiple institutions.

1.2 Secondary Use of Data

Secondary use of data is important. The secondary analysis of existing data has become an increasingly popular method of enhancing the overall efficiency of health research enterprise. A study shows the secondary use of health information has significant implications for basic and clinical research, public health surveillance and management, quality improvement, and safety-monitoring[1]. Another study demonstrated the importance of the secondary use of health data in Ireland[2]. This resource makes economic and ethical sense to use this data as much as possible to improve the effectiveness and efficiency of the health services. Other benefits may include researchers finding a large amount of data online instead of collecting primary data. The increasing availability of such data online encourages the creative use and cross-linking of information from different data sources[3].

Electronic Health Record data is one of the secondary uses to improve patient quality. A research study stated the use of electronic health record data can accurately identify whether

patients need lung cancer screening[4]. The purpose of this study design is to compare the sensitivity, specificity, and positive and negative predictive value of an EHR query to patient self-report, to identify patients who are in need of lung cancer screening. They invited 200 current or former smokers between age 55-80 in a large, community-based health care system. 24 surveys were included in the analysis[4]. In this study, researchers found out EHR data had a 66.7% positive predictive value and 81.8% negative predictive value for identifying patients eligible for lung cancer screening[4]. Although the accuracy of EHR data today will be useful to clinicians to initiate conversations between patients about lung cancer screening information. The study shows the importance of accessing health care data to improve patient quality. Sharing is couple with secondary use because in order for data to be used by others, they need to be accessible[5]. Data is demanding for researchers. The demands for scientific data arise primarily from two areas. One demand comes from the scientific questions that researchers attempt to answer[5]. The second type of demand for scientific data is comprised of a broad range of social influences[6]. Although the two demands are listed separately, they are often intertwined. For example, when data requests come from researchers in other fields, different cultural norms and expectations can complicate sharing[7]. Data sharing is and it is an important example of secondary data use. There is no doubt that data sharing and provision of secondary data access can have a profoundly beneficial impact on progress in biomedicine and the health science[8]. The National Academy of Science stated new knowledge could be transformed into socially beneficial goods and services only by sharing research data[9]. Researchers can use the research information when it is accessible to create products and services for human needs. A report by the Research Information Network of the United Kingdom examined data sharing and

stated the global importance and relevance of data accessibility in research[9]. Data sharing can help medical professionals and researchers review information from patients, cross reference similar medical conditions and other factors across a vast data set, and draw conclusions based on these findings[10]. They can then allow more data to be analyzed, tested and evaluated which treatments are most effective across a large number of patients. The purpose of this research study is to evaluate the secondary use of healthcare data by using different query tools to answer representative questions.

1.3 Benefits and Challenges with Secondary Use of Data

Many studies show the importance of secondary use of data, but accessing data is still a challenge. A systematic review of secondary use in public health shows the use of data has become essential for decision making at the local, national, and global level[11]. Although society is recognizing the benefits of data reuse– transparency and cooperation, research, cost-efficiency and preventing redundancies, this can still be challenging in reality. A global policy framework or guidelines have not yet been developed for most types of data, which leads to one of the challenges of secondary use of data[11]. Another study published in *Journal of the American Medical Informatics Association* shows the importance of the secondary use of health data which applies to personal health information (PHI) for different purposes of the outside of direct health care delivery[12]. Secondary use of health data can enhance individuals' health care experiences, expand knowledge about diseases and treatments, strengthen understanding of health care systems' effectiveness and efficiency. For example, individual patients are able to access their own electronic health records and view health information. It can improve patient quality care, and lead to more efficient, more personalized care outside of clinical settings[12].

However, complex ethical, political, technical, and social issues surround the secondary use of health data. The same study conducted a panel, it was a first step in promoting dialogue among researchers about the opportunities and challenges related to the secondary use of health data[12]. 36 panel members were involved. The panel focused on secondary uses of person-specific health data and four main perspectives were viewed for secondary use of health data. The panel enumerated major issues associated with secondary uses of health data and some major findings and recommendations were record[12]. These recommendations provide guidance that should shape a national framework for secondary use of health data. Despite the significant benefits of secondary use of healthcare data, data availability and institutional differences in practice limit researchers' use of secondary data.

Navigating local data can be difficult because limited tools exist and tools that overcome navigating data are not always available. For example, Individual researchers require more time to verify, analyze, and derive their data and conclusions to publish their results. Manually checking data would make mistakes due to human error, if new tools can automatically rapidly share and assess data quality, it would improve the secondary use of data for researchers. However, these tools have their own set of challenges. Although self-service query tools and common data query tools are developed to address secondary use of data, these tools may not be able to answer all data request questions. Also, it is challenging for researchers to access secondary use of data by writing SQL queries or other methods containing programming.

1.4 Barriers to Secondary Use of Data – Privacy and Confidentiality

Privacy and confidentially create another barrier to secondary use of health data in the medical field of research data. Currently, people are trying to make medical research data available to the public while ensuring the data are de-identified before using. Identifiers can be removed prior to data use to the public to protect patient information, however, challenges remain for ongoing concern and investigation[4]. Reuse of clinical data has been widely spread and is crucial for health care quality, management, reduced costs, population health management, and effective clinical research[13]. Mark Weiner et al. published the *Reuse of Clinical Data for Research and Quality Improvement* stated the health information technology (HIT) today provides an extraordinary coverage for the electronic health records, data use capabilities, improving research and health care quality[14]. Enabling access to high quality, patient-level health information is one approach where data can be transferred and readily accessed to answer questions. However, HIT raises personal privacy and intellectual property concerns, and the goal of information widespread to enable integrated health information access from various health systems remains elusive. Another study found in 2004 by Jane and Schur referred to the research study on 'New Approaches to Confidentiality Protection' by Abowd and Lane[15], they mentioned one approach to new health care access is to design synthetic public use data files that add systematic noise to the microdata. This technique measures data quality and reduces risk since the synthetic data record does not expose the actual data record, and identity disclosure is impossible. The drawback of synthetic data is the reduced of data utility[15]. Very few people are trained in this technology which causes users having difficulties to use the dataset correctly.

Privacy and confidentially act under the HIPAA laws. A study stated although the HIPPA guideline protects health information, still, there is a potential lack of protection of personal health information which is not explicitly covered by HIPAA legislation or regulations[12]. For instance, HIPPA requires the de-identification of data and data use agreements, there are still possibilities for re-identification of patients.

Getting approval due to privacy and confidentially becomes a challenge for secondary use of health data. Approvals may sometimes require a long time, it would affect researchers regarding their future work. The tools that I used in this research study combines de-identified data and data that are not de-identified. An Institutional Review Board (IRB) needed to be filled out before I can get access for approval.

1.5 Data Sharing Limits Secondary Use of Data

Data sharing is especially difficult among health care settings which limits the secondary use of data for researchers. Although the acceptance of willingness to engage in data sharing is increasing, there is also increased perceived risk associated with data sharing, and specific barriers to data sharing persist. Limited systematic framework or global operational guidelines have been created for data sharing, barriers at different levels have limited data sharing[11]. The largest discrepancy in current practices of data sharing is between what people believe should be done with data and what is actually being done[16]. Despite an overall belief that scientific data should be available for use beyond their original purpose[17], scientists are often protective of their data and may not readily engage in sharing practices[18,19]. A study was conducted by DataONE team members to capture researchers perceptions about data sharing and reuse by answering a series of designed questions. The result showed researchers were more concerned

about the possible risks associated with sharing data, and the potential for misuse and misinterpretation[20]. The above study stated individual factors can have an effect on data sharing, privacy and confidentiality barriers across multiple institutions is another factor[20]. Although there are some policies and recommendations related to data sharing, policies for data sharing have tended to vary by different institutions. Another study published by Kohane et al found some problems in data sharing among institutions including shared data presentation, standardized vocabulary, data selection, and confidentiality across multiple institutions[21]. The study also stated sharing data across institutions must first have a shared policy for data security, authentication, and disciplinary action.

Data navigation is even harder across multiple institutions. Although researchers could overcome these issues at one institution, they need to overcome more challenges at other institutions for data sharing.

1.6 Data Navigation tools to increase Secondary Use of Data

Data navigation is an important concept that has been attempted to be addressed. The purpose of self-service query tools is to enable users to perform their day-to-day analytics tasks themselves to get more involved in the more critical analysis process[22]. People need technologies to gain access to data. Accessing EHR data for researchers is a major challenge in using self-service query tools (SSQT) to meet the need of diverse users. Hruby, Ancker, and Weng, authors of the article “Use of self-service query tools varies by experience and research knowledge” reported user experiences across SSQT and how diverse users interact with SSQTs for future effective query tool designs[22]. The study conducted eight semi-structured interviews and user observations at four academic institutions. Users include physicians, clinical

researchers, and EHR data analysts. They were asked to write a query to solve their real-world information using “think aloud” protocols. The studies were all videotaped to capture the actions and thoughts of different users. After completing the query, an exit interview was performed, and a user-action schema was developed and pruned for a video annotation. For remaining quality control stabilized, user actions were annotated with this schema by a single annotator in each video[22]. Users were divided into two groups: experts and novices, based on beyond 2 years of experience with research and SSQT. Results showed in four user actions: browse, enter, review, and select. It is found both experts and novices had similar frequency distributions among actions. One finding showed experts prefer to use the action “Enter ‘Search Criteria’” compared to novices. The study also found most SSQT experts performed a more organized flow of user actions; they intended to add data elements instead of removing them after building the query[22]. Comparing to expert users, novice users implied to develop their queries by adding or removing data. The final results presented a similar pattern in both experts and novices using the SSQT. Self-service query tools (SSQT) have been developed to meet this goal among diverse users. Based on the above study, self-service query tools are suitable for both expert and novice users when building the same queries[22]. There were minimal differences between expert and novice researchers’ user-action pattern which means self-service query tools are developing rapidly today.

I2b2 is one of the self-service query tools that exist to simplify data access. I2b2 is referred to as Informatics for Integrating Biology and the Beside[23]. The primary purpose of i2b2 is to develop the science and the engineering required to enable the clinical investigators of academic medical centers to conduct clinical research that is informed by state-of-the-art genomic and

biomedical informatics[23]. I2b2 is a tool which combines clinical research data with basic sciences research data to develop a new informatics framework. It is a light-acceptance product of CTSA in Harvard University developed by medical informaticians which continuously spread the use of this product this day. I2b2 is an open source tool that is designed to overcome significant obstacles to translating the discoveries of the genomic era into safer, more effective and more personalized health care[23].

Common data model (CDM) is another tool for data navigation. Common data models are often used in research when researchers need to exchange or share data for particular reasons or uses. CDMs are used in clinical research to standardize and facilitate these exchanges from multiple sources[24] . Common data model differs in both purpose and design. It allows users to generate evidence from a wide variety of sources. The current CDMs have been developed to support secondary use of healthcare data in research. The Observational Medical Outcomes Partnership (OMOP) is one of the commonly used CDMs that allows for the systematic analysis of disparate observational databases. Its concept is to transform data contained within those databases into a common format data model as well as a common representation such as terminologies, vocabularies, coding schemes, and then perform systematic analyses using a library of standard analytic routines that have been written based on the common format[25]. OMOP also accommodates different data domains typically found within observational data (demographics, visits, condition occurrences, drug exposures, procedures, and laboratory data)[26].

1.7 Challenges with Previous Approaches

Despite the existing two tools – i2b2 and OMOP for simplifying the secondary use of health data, these previous approaches still face challenges. According to the research by Hruby et al in "Use of self-service query tools varies by experience and research knowledge," self-service query tools is still a barrier to designing effective query tools due to lack of understanding[22]. In this research, it is showed that the proficient use of self-service query tools requires significant technical experience.

I2b2 needs to overcome two major obstacles[22]. The first obstacle is comprised of the computational challenges of discovery across large, heterogeneous datasets in clinical care and the genome-wide measurements made of the corresponding patients. The lack of knowledge of genomic-level physiology and how to study it became the second obstacle.

According to the research, self-service query tools is still a barrier to designing effective query tools due to lack of understanding[22]. In this research, it is showed that the proficient use of self-service query tools requires significant technical experience. The evaluation of i2b2 for clinical research stated the graphical interface in i2b2 imposes limitations that prevent all query conditions from being satisfied without post-processing in the form of extensive manual intervention, which may not be practical when selecting large research cohorts[27]. Some of the challenges related to modeling, storing and retrieving temporal data from a clinical information system arise from the limitations in how these systems are used to capture the information during clinical care[27]. Another limitation is the diagnoses related information is often buried in textual narratives, which makes it less accessible to automated retrieval methods. The same study found the limitations that exist in the capture of clinical information apply equally when

trying to retrieve it from the clinical system[27]. For example, trying to query other clinical conditions or medications or lab results that followed a given diagnoses from the electronic chart is limited by how effectively the diagnoses information was captured and stored in the chart during the patient's care.

Although OMOP CDMs have various benefits, limitations still exist. A research found out a potential limitation of mapping individual databases to a CDM is that the CDM may not allow some of the relationships or data contained in a local database to be fully represented[31]. It is necessary to map data to a CDM to maintain some of the relationships or data contained in the original data. However, data mapping or transformations can be mismatched or "lossy" when using CDMs as a logical data model for data storage[26]. Unless a CDM is a perfect representation of source data, information loss will occur as a result. Thus, CDM is not recommended for data storage unless it is fully developed or evaluated for that specific content coverage[26]. Some other potential limitations loss of information associated with abstracting above individual source system differences, and differences in supported associations[26]. The transformation of data to CDMs should occur at the latest data processing stream as possible – closest to the analysis to preserve full information content of data for potential secondary use.

Leaf is another tool that has been built to address the problem of the secondary use of health data. LEAF is a self-service clinical data discovery query tool will provide the ability for researchers to individually query a managed, de-identified view to handle data visualization requests developed by University of Washington Medicine Information Technology Services that leverages the Amalga clinical data repository. Leaf can extract and visualize any datasets

into a portable format that researchers can easily query without needing high technical background or support.

Due to the limited comparison between the common model query tools and self-service query tools, people are unclear which tool is most appropriate under specific circumstances. The focus of this research study is to assess these tools and find out the value of them.

1.8 Technology Acceptance Model Framework (TAM)

Technology Acceptance Model is a useful framework to assess tools. Although many models have been proposed to explain and predict the use of a system or tool, the Technology Acceptance Model has been the only one which has captured the most attention of the Information Systems community[28]. The Technology Acceptance Model application in health care settings has been widely spread. It is a theory that has been widely researched outside of health care and it has become an important tool for health IT research. Holden and Karsh reviewed the TAM theory to health care[29]. They reviewed 16 datasets analyzed in over 20 studies of clinicians using health IT for patient care, a model of quantitative relationships between variables. The results found TAM has been widely spread in explaining health care provider's reactions to health IT. Although TAM is well-used in predicting, and explaining clinician end-user acceptance and use of health IT, the remaining challenges indicate a deeper problem with TAM in health care use[29]. TAM requires a need for standardization and a need to continue exploring new theoretically motivated variables and relationships. Researchers have used Technology Acceptance Model to understand the acceptance of different types of information systems.

Another research study found out that researchers wanted to use Technology Acceptance Model (TAM) to find out the acceptance of information technology in Health Information Management (HIM)[30]. The study sampled 187 people from 363 people who were working in the medical records department at Tehran University of Medical Sciences. A researcher-developed questionnaire was applied to users' perception when applying to information technology. Data were analyzed using descriptive statistics and regression analysis by SPSS software. Results demonstrated TAM is a useful framework for HIM to assess user acceptance of information technology. One interesting finding is the result of perceived ease of use (PEOU), and perceived usefulness (PE) were positively correlated with users' attitudes to HIM. PU was more associated than PEOU[30]. The results and findings in this study suggested that user acceptance is a crucial element and should become a major concern for health organizations and health policymakers. This study will use Technology Acceptance Model (TAM) to evaluate between Leaf, i2b2, and OMOP to find out which tool is most appropriately used under certain circumstances. TAM explains the motivation of users by three factors; perceived usefulness, perceived ease of use, and attitude toward use[32]. The two main components were used in this research study: perceived usefulness and perceived ease of use, to assess and study the three query tools – i2b2, Leaf, and OMOP.

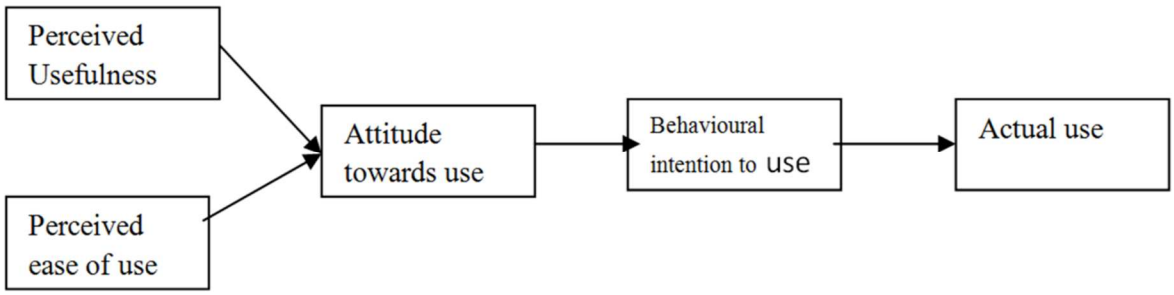

Figure 1: Technology Acceptance Model[33]

Methods

2.1 Overview

The main study of my research is to compare and evaluate between OMOP, LEAF, and i2b2 to find the most appropriate performer to increase data access to health care. The sub study is to determine the 9 types of queries in order to compare between common model tools and self-service tools to determine whether these query tools could answer the questions. I queried 20 questions from different sources and areas to make my results less bias and more varied using i2b2, LEAF, and OMOP. Because there are limited comparisons made between these self-service query tools and common model query tools, researchers do not have an understanding of which tool is most appropriate under certain circumstances. It is also important to evaluate the number of questions these tools could answer by using the representative questions in health care.

I chose Leaf and i2b2 as my two self-service query tools because Leaf is a local tool developed by UW Medicine and i2b2 is a public tool. I would like to evaluate the differences between two self-service query tools such as features and the results. I chose OMOP as my common data model example because it is commonly used across many institutions. For instance, UW Medicine and Seattle Children's Hospital are using it. Also, referring to the research on evaluating common data models, OMOP stands out comparing to Sentinel and PCORnet. OMOP had the highest content coverage and data elements, 76%, compared to Sentinel (37%) and PCORnet (48%)[24]. These four CDMs were chosen from models in use for clinical research data and each model was evaluated based on 11 criteria in six categories: content coverage, integrity, flexibility, ease of querying, standards compatibility, and ease and extent of

implementation[24]. The results showed OMOP is the best match for supporting data sharing from longitudinal EHR-based studies.

2.2 Research Focus

My main focus was to examine whether the three query tools can answer all the questions and what are some differences and similarities comparing these tools.

Initially, I assumed either OMOP or i2b2 would be the most appropriate performer to help researchers increase data access by evaluating the number of questions they answered. The two query tools have been developed and used for several years by various health care industries and they would be mature enough to complete the query questions. I thought LEAF cannot answer all the questions compared to i2b2 because LEAF is a newly developed self-service tool and it may not contain varies specific diseases and features.

Our goal was to test the questions by ourselves and ask different levels of query writers to answer the same problems as us. Initially we want to make a comparison on the results to evaluate if the same questions have different answers by different users. Later, we found that it was too challenging to invite varies levels of people to test 9 types of queries based on the timeline. Ultimately we settled down the research by assessing the query question on my own.

I want to determine the perceived ease of use by comparing various features between Leaf and i2b2. For example, whether i2b2 has the functions of ‘encounter’; ‘emergency’; ‘discharge’ when answering the query questions, and create a comparison table of each feature.

In my sub study, we are trying to determine the 9 types of queries by research representative questions and discussion. Also, we want to find out if the same question can return different results according to different variations based on Kahn’s paper. Among all those 20 questions, I

want to know the number of patients each tool returns after running the query, and I am also interested in whether LEAF and i2b2 are using similar concepts for the same question. I started to query the questions I chose by using the three different query tools in my main study to test perceived ease of use and perceived usefulness using Technology Acceptance Model (TAM).

2.3 Determine the Types of Query Questions

I chose the 9 types of queries based on the representation and essentialness in healthcare. I also chose to answer two questions based on Michael Kahn's paper on "Querying Clinical Databases: How Many Patients?[35]" because I want to know the result difference when the same query question is answered in two ways. This article has not been published yet, but it is the best source from a query writer examining clinical data from complex administrative, clinical and national databases. Michael Kahn's expertise is in data quality assessment is demonstrated in an article published 'The Journal of Electronic Health Data and Methods' named 'A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data[34].' In his paper, he mentioned currently clinical analytics is expanding rapidly with the development of electronic health records, large clinical data warehouses, national quality improvement collaboratives and, the newest addition, distributed research networks[34]. The power of emerging data resources is impressive, but it still leaves miscommunication an issue which results in large differences between data. One example Kahn provided is to determine the number of patients with two related ICD – 9 diagnoses who were seen by a specific provider. The question was asked: "how many patients with neurofibromatosis-1 and scoliosis did provider = 123456 seen recently?" It turned out the same question can return 9 different ways to answer from the electronic medical record database

(Figure 2)[35]. Depending on the creativity of the query writer, each variation in Figure 2 embodied different assumptions and returned closely related results. With Kahn's reference, we decided to choose our two questions (A & I) of queries based on his statement. We chose to query 'A' and 'I' because they have the most significant difference and we would like to evaluate the results based on each variation.

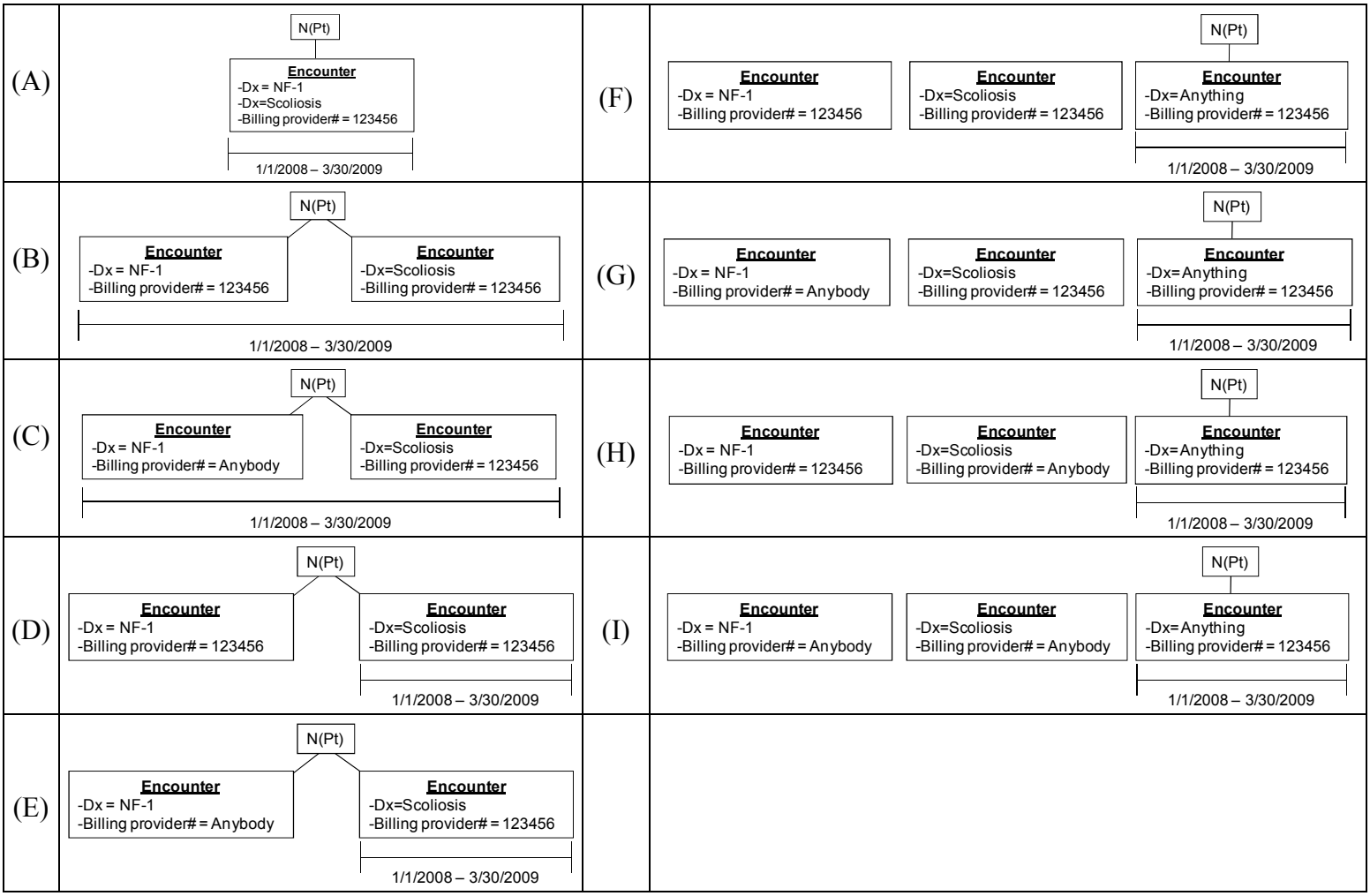

Figure 2: Kahn's 9 types of queries answering the same question[35]

The data requests that I chose from are also based on real-world examples, including the three of the data queries from the analytics team in University of Washington. The 9 types of queries

are divided into 3 different categories: quality measurement; based on observational research studies; and ask the analytics team.

2.4 Actions to Get Data and Writing Queries

We chose these 3 categories of queries because they are most representative and essential in health care; also, we want to diversify our choices to cover more information. For quality measurement, I included 30-day re-admission, hospital prevention, and health care professionals. We decided to randomly choose three questions for both hospital preventions and quality measure for professionals and clinicians by using a true random number generator to generate the avoid personal bias. We focused on VTE, ED, and discharge in hospital prevention because these three elements are most representative in hospitals (Table 1) most represented metrics used in hospitals[36].

I was thinking about querying measure description at first for hospital prevention and professionals and clinicians prevention[37] (Table 2). However, after meeting with Nicholas Dobbins, one of the developers of LEAF, he suggested that it is difficult to perform calculation for measure description such as median and percentage in LEAF. Ultimately, I decided to only query the denominator statement containing more specific features like age, length of stay, diagnoses, and measurement period.

For 30-day re-admission, we included the association between frequent elderly attendees to the emergency department with increased 30-day mortality[38]. We want to determine factors associated with 30-day readmission in elderly patients admitted for community-acquired pneumonia (CAP)[39]. We also want to examine whether or not a potentially modifiable factor

such as physical activity affects the lowering risk of 30-day readmission in patients with chronic obstructive pulmonary disease (COPD) (Table 3)[40].

I also chose three queries questions based on observational studies in with EHR data. I found 10 observational studies on electronic health records based on research and selected the three most interesting articles for my queries. I focused on type 2 diabetes including HbA1c comparing between body mass index and insulin initiators. The query is based on Patients with T2DM with more than one office visits between 6 and 18 months before the index date, and with ≥1 glycosylated hemoglobin (HbA_{1c}) result in the 6-month preindex (baseline) period were included between November 2014 and February 2016.[41]. The second choice I made the study examining survival across older adults with different chronic multimorbidity patterns (CMPs)[42]. I chose this query because 'multimorbidity' refers to as patients with three or more chronic diseases, it is interesting to find out the returned results with each tool. I chose to query the diagnoses 'acromegaly' because it is a very rare disease in the US. I am curious if these query tools can find patients with this specific disease. (Table 4)[43].

The last three queries we wanted to use a different source instead of the external sources. We asked the UW analytics team to provide us with three research questions that are representative. We received the questions from the Director of Research IT. We then found that the questions he gave us were not actually a query question, but rather topics about how health care works and is structured. We modified the questions into queries that might be answered by self-service query tools (Table 5).

I also included three questions from an evaluation of i2b2 to make my research and analysis more generalized (Table 6)[27].

I used i2b2 and LEAF to query the 20 questions at the same time to see the similarities and differences because these two are both self-service tools. For hospital prevention questions, I included three columns: measurement description, numerator, and denominator. For professionals and clinicians prevention, I included only the denominator. I chose to query the denominator because the measurement description and numerator require calculations which is more challenging in self-service tools to execute. There were some challenges with i2b2 when I was querying the questions. Unlike LEAF, the data in demo version of i2b2 is not current. I tested the date range starting from 2017 until 2010; I could only find data before 2010. Therefore, the date range for i2b2 and LEAF of the same query is different, but they both can answer the questions and the date difference does not affect the capabilities of i2b2.

I started writing in OMOP after finish running queries in both i2b2 and LEAF because i2b2 and LEAF are similar, it is easier to compare the features and data when querying at the same time. The four main tables I used in OMOP are person, visit occurrence, condition occurrence, and concept. I generally used these four tables because I need patient information such as 'patient_id'; the visit occurrence type is included to determine whether a patient is admitted into a hospital as an inpatient, an outpatient or admitted through the emergency department. Condition occurrence table is also needed to join the three tables together to find the unique value to count the number of patients; concept table is also relevant for users to find out the specific diagnosis in the query questions.

I created a comparison table to indicate the number of patients executing from each tool to see the difference or gap in those numbers answering the same questions. I am also interested in finding out if i2b2, LEAF, and OMOP can answer all the questions, and will any of the results

return no data. After obtaining the results, I generated two tables to make a comparison between the three query tools. A master list was created containing all 20 questions with additional three columns behind for the number of patients in LEAF, i2b2, and OMOP. Another table was built to compare the features of LEAF and i2b2 since they both are self-service query tools. I did not include OMOP in the second table because it is a common data model which does not have the same features as self-service tools. I want to compare the number gaps between each result, the flexibility, perceived ease of use, and perceived usefulness to determine which query tool is most appropriately used at specific circumstances.

Table 1 presents the three electronic clinical quality measures for hospital preventions including Venous thromboembolism (VTE), emergency department (ED), and discharge. I started to query measure description at first, however, the measure description includes calculations such as proportion and median. From the research I found, the current version of i2b2 could not do calculations[27]. I focused on the numerator and denominator statement, I used LEAF first to query the numerator statement in one column, and denominator statement in the second column. However, this plan did not work out, no patient was found if the numerator and denominator is combined. Alternatively, I chose to query the denominator because the denominator contains more specific information such as patient age, length of stay, and measurement period.

Hospital Prevention	Measure Description	Numerator Statement	Denominator Statement
VTE	This measure assesses the number of patients who received VTE prophylaxis or has documentation why no VTE prophylaxis was given the day of or the day after hospital admission or surgery end date for surgeries that start the day of or the day after hospital admission	Patients who received VTE prophylaxis: the day of or the day after hospital admission the day of or the day after surgery end date for surgeries that end the day of or the day after hospital admission Patients who have documentation of a reason why no VTE prophylaxis was given: between arrival and hospital admission the day of or the day after hospital admission the day of or the day after surgery end date (for surgeries that end the day of or the day after hospital admission)	Patients age 18 and older discharged from hospital inpatient acute care without a diagnosis of venous thromboembolism (VTE) or obstetrics with a length of stay less than or equal to 120 days that ends during the measurement period
ED	Median time (in minutes) from admit decision time to time of departure from the emergency department for emergency department patients admitted to inpatient status	Measure Observations Statement: Time (in minutes) from Decision to Admit to ED facility location departure for patients admitted to the facility from the emergency department	Initial Population Statement: Inpatient Encounters ending during the measurement period with Length of Stay (Discharge Date minus Admission Date) less than or equal to 120 days, and where the decision to admit was made during the preceding emergency department visit at the same physical facility Measure Population Statement: Equals initial population
Discharge	Median elapsed time from emergency department arrival to emergency room departure for patients discharged from the emergency department	Measure Observations Statement: Median elapsed time (in minutes) from emergency department arrival to emergency room departure for patients discharged from the emergency department	Initial Population Statement: Emergency department encounters during the measurement period Measure Population Statement: Equals initial population

Table 1: Query Questions for Hospital Prevention[36]

Professionals & Clinicians	Denominator Statement
	Patients 18-75 years of age with diabetes with a visit during the measurement period
	All patients aged 18 years and older before the start of the measurement period with at least one eligible encounter during the measurement period
	Patients 18-85 years of age who had a diagnosis of essential hypertension within the first six months of the measurement period or any time prior to the measurement period

Table 2: Query Questions for Professionals and Clinicians Prevention[37]

Table 2 shows the three randomly chosen query questions from electronic clinical quality measures for professionals and clinicians. I did not include the measure description and the numerator columns because from the example above for hospital prevention, I could not query the measure description due to calculations. The denominator has more precise information to accurately find the number of patients.

30-day Readmission	Title	Query Sentence
	Association between the elderly frequent attender to the emergency department and 30-day mortality: A retrospective study over 10 years.	Patients aged 65 years and older, with 3 or more visits within a calendar year were identified.
	Factors associated with 30-day readmission after hospitalization for community-acquired pneumonia in older patients: a cross-sectional study in seven Spanish regions.	Patients aged ≥65 years admitted through the emergency department with a diagnosis compatible with CAP.
	Associations between physical activity and 30-day readmission risk in chronic obstructive pulmonary disease.	Patients discharged between January 1, 2011 and December 31, 2012, aged 40 years or older, on a bronchodilator or steroid inhaler.

Table 3: Query Questions for 30-day Readmission[38,39,40]

Table 3 contains three query questions for 30-day readmission. I chose these three query questions based on the papers I read and I found these three statements most related to my research topic. Some of the papers are representative, but they are not query questions. It includes the title of the papers and the query sentences from each chosen paper. Also, I chose different diagnoses and different contents to make the questions more diverse. For example, both first and second query has patient age 65 years and older, but they have different contents where the first question includes 3 or more visits, and the second contains diagnoses with CAP.

EHR Data Based on Research	Title	Query Sentence
	Characteristics Associated with the choice of first injectable therapy among US patients with type 2 diabetes.	Patients with T2DM, ≥1 office visit between 6 and 18 months before the index date, and with ≥1 glycosylated hemoglobin (HbA_{1c}) result in the 6-month preindex (baseline) period were included between November 2014 and February 2016.
	Survival in relation to multimorbidity patterns in older adults in primary care in Barcelona, Spain (2010-2014): a longitudinal study based on electronic health records.	Patients aged ≥65 years with chronic multimorbidity patterns (CMPs) identified.
	Use of Electronic Health Records to characterize a rare disease in the use USA: treatment, comorbidities and follow-up trends among patients with a confirmed diagnosis of acromegaly.	Patients aged 65 years and younger admitted as Inpatient or outpatient between 2008-2013 with diagnosis of acromegaly

Table 4: Query Questions for Observational EHR Research data[41,42,43]

Table 4 includes questions from observational electronic health record data based on my research. I found a lot of EHR data related to diabetes, but I only chose one to avoid the bias from this diagnoses. I chose to query the diagnoses chronic multimorbidity patterns (CMPs) because this diagnoses defines as patients who have multiple (3 or more) chronic conditions, such as diabetes, kidney disease, heart failure, hypertension, depression, cancer, or others. I chose the diagnoses 'acromegaly' because it is a rare disease in the USA and I am curious if these query tools could find this diagnoses.

Research Questions Adapted from Analytics Team	Query Sentence
	Is there an interaction between hormonal contraceptives and cystic fibrosis medications that may affect the efficacy of the contraceptive method or the cystic fibrosis medication. Query: patient with a diagnosis of cystic fibrosis who have been prescribed with cystic fibrosis medications and also prescribed oral contraceptives.
	Prove or disprove the hypothesis that intraoperative cardiac arrest has significant effects on mortality, graft survival, and perioperative morbidity. Query: patient with intraoperative cardiac arrest (LOS)
	Prove or disprove the hypothesis that 10% of inpatient records for UWMC and HMC Medicine services contain progress notes that contain clinically misleading content resulting from copying and pasting. Query: UWMC & HMC Inpatient records and diagnosis and procedures

Table 5: Query Questions from the Analytics Team

Table 5 indicates the questions the analytics team sent to us. I divided the query sentences into two parts because the questions they sent was not actually query questions. The first part is the original questions the analytics team sent. For part two, my professor and I changed the questions into query questions where it starts with the word “Query.” The three questions did not include any patient age compared to the above three tables.

Evaluation of i2b2	Description of data requests in Evaluation of i2b2
	PI needs counts of patients that had Basal cell Carcinomas, Squamous Cell Carcinomas, Squamous Cell Carcinomas in situ, Displastic Nevus, any kind of Nevus, melanomas, etc. Broken down by years 2004 and 2005.
	Female patients from January 2002 onwards with a diagnosis of DVT/PE and age less than 51.
	PI needs a list of patients who have undergone Orthopedic surgery with a therapeutic INR on the day of surgery.

Table 6: Query Questions from the Evaluation of i2b2[27]

Table 6 illustrates the three questions in the evaluation of i2b2 paper. I chose to include these three questions because it is chosen from a valuable source which would make our selection of questions broader. Also, i2b2 has already answered these three questions and recorded in the paper, I wanted to test if the version of i2b2 I use could answer the questions as well.

Results

Result 1: Perceived Usefulness

Performance

After answering the 20 questions by these three query tools, the results in table 7 shows 90% (18 questions) can be answered by both LEAF and i2b2. OMOP data model can answer 19 out of 20 questions (95%). When comparing i2b2 with LEAF and OMOP, it is clear that LEAF and OMOP show more number of patients than in i2b2 due to the use of different data. The data requests contain similar information such as patient age, diagnoses, length of stay, and measurement period. These information are all included in the three query tools which had a high performance of answering these questions. All three query tools shows the timing when searching for the number of patients. When the systems could not find any patients due to long time searching, they will show an error message to notify users and users could make adjustments. Leaf and i2b2 also provide clear categories when searching for the specific information. OMOP is useful when users know how to program; users can create their own SQL queries according to the specific contents to provide more accurate results. Leaf and i2b2 would enhance the efficiency if researchers want to find the number of patients in a fast pace.

Flexibility

Flexibility is considered as perceived usefulness because I want to evaluate the tools based on the capability of the user interface. This is in contrast to flexibility of the user interface which would be a component of perceived ease of use. OMOP has a strong flexibility comparing to i2b2 and LEAF because OMOP is all about writing SQL queries. Each person might have a

different query for the same question according to Michael Kahn's paper. For example, question #19 and #20 are answering the same question from different variations in all three query tools. By comparing the results for these two questions, all three tools returned different results. Below is the screenshots from three different query tools. The results in OMOP show a significant difference when comparing the two diagnoses whether they are in the same encounter or any encounter. Leaf and i2b2 returned a similar result when comparing these two questions.

```
-- SAME VISIT
select count(p.person_id)
from uzOMOP.OMOP.visit_occurrence v, uzOMOP.OMOP.person p,
uzOMOP.OMOP.condition_occurrence co1, uzOMOP.OMOP.condition_occurrence co2
where v.visit_start_date between '2017-01-01' and '2018-01-01'
and v.person_id = p.person_id
and v.visit_occurrence_id = co1.visit_occurrence_id
and v.visit_occurrence_id = co2.visit_occurrence_id
and co1.condition_concept_id in (44828085, 45538758)
and co2.condition_concept_id in (44829870, 45548935)


-- ANY VISIT
select count(p.person_id)
from uzOMOP.OMOP.visit_occurrence v, uzOMOP.OMOP.person p,
uzOMOP.OMOP.condition_occurrence co1, uzOMOP.OMOP.condition_occurrence co2
where v.visit_start_date between '2017-01-01' and '2018-01-01'
and v.person_id = p.person_id
and p.person_id = co1.person_id
and p.person_id = co2.person_id
and co1.condition_concept_id in (44828085, 45538758)
and co2.condition_concept_id in (44829870, 45548935)
```

Figure 3: Screenshot of OMOP for questions #19 and #20

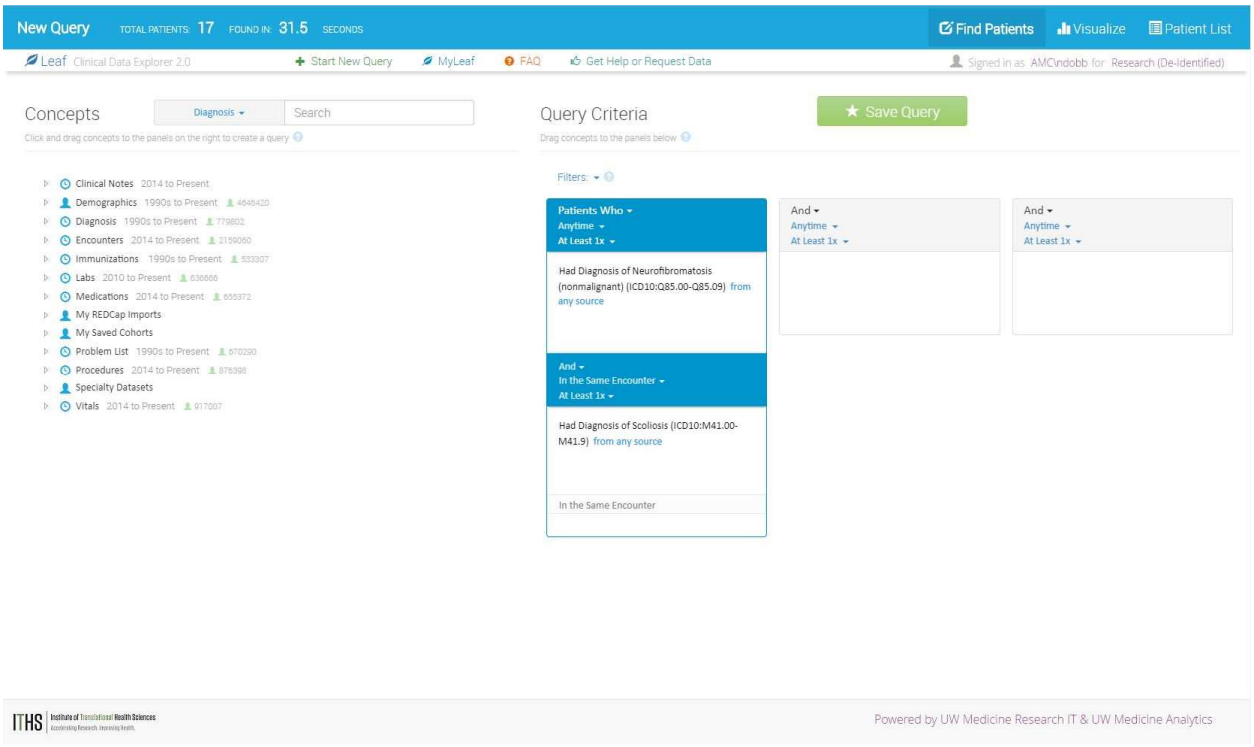

Figure 4: Screenshot of Leaf when diagnoses are in the same encounter (question #19)

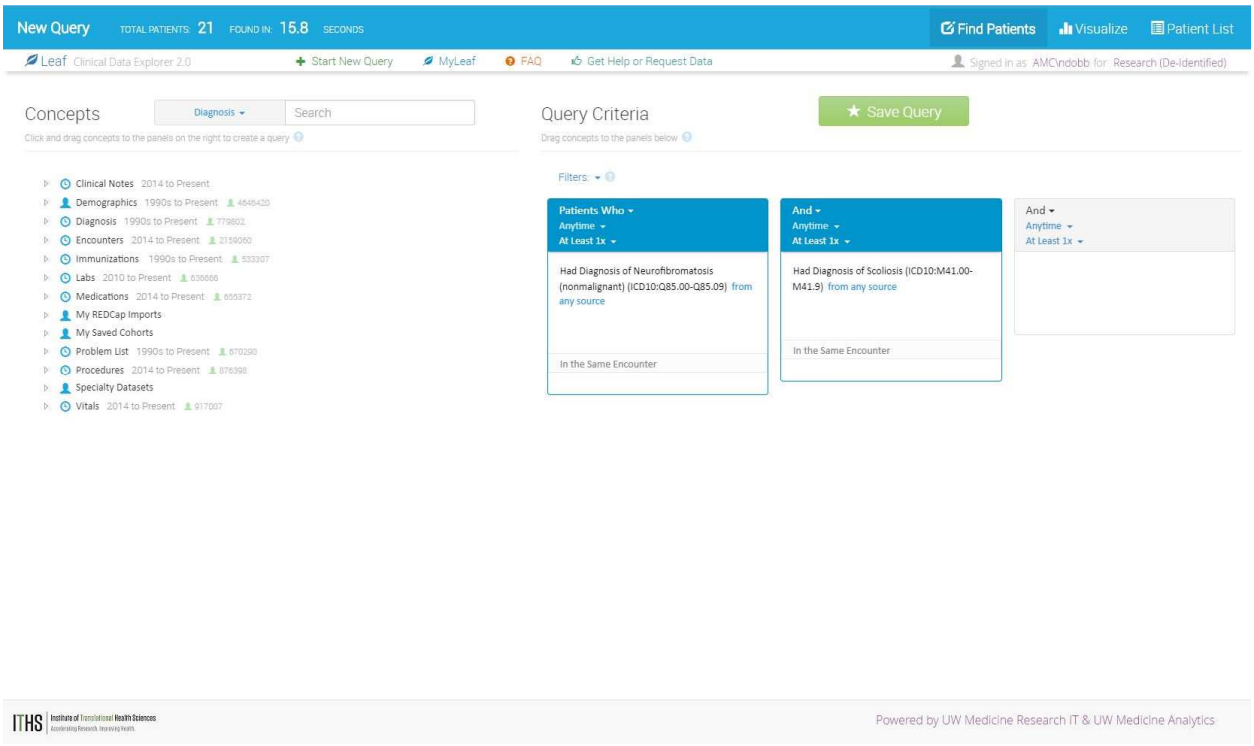

Figure 5: Screenshot of Leaf when diagnoses are in the any encounter (question #20)

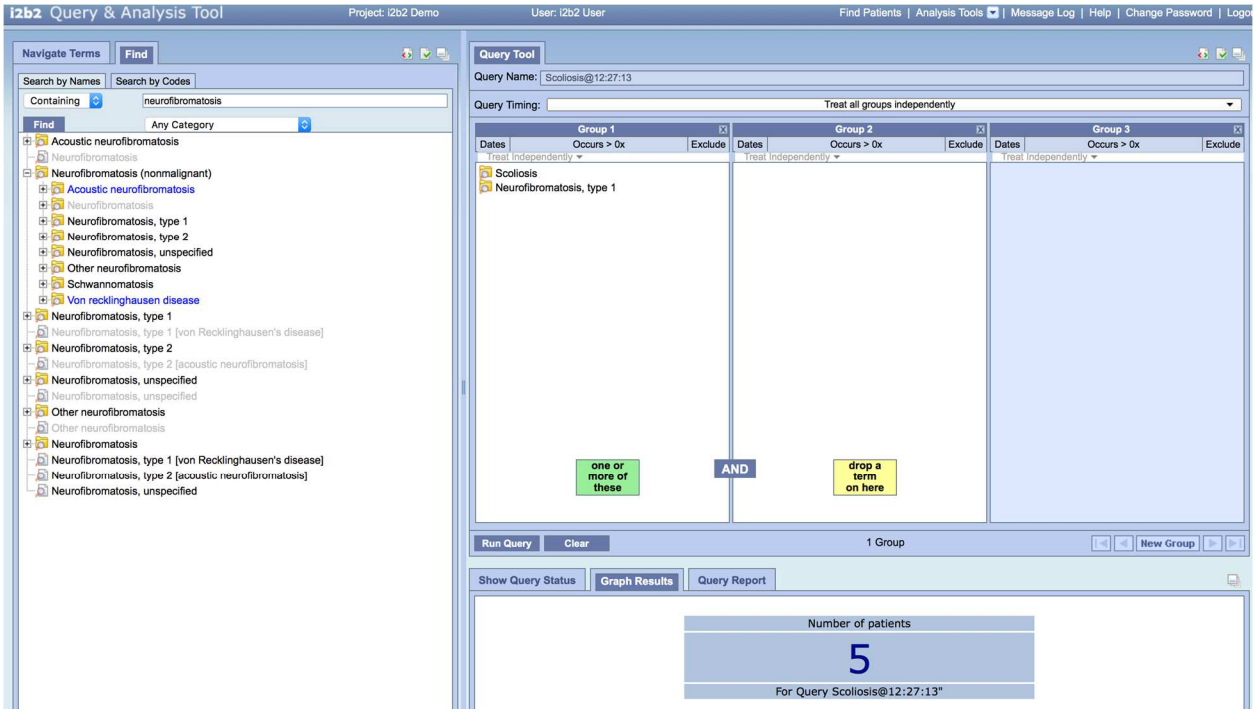

Figure 6: Screenshot of i2b2 when diagnoses are in the same encounter (question #19)

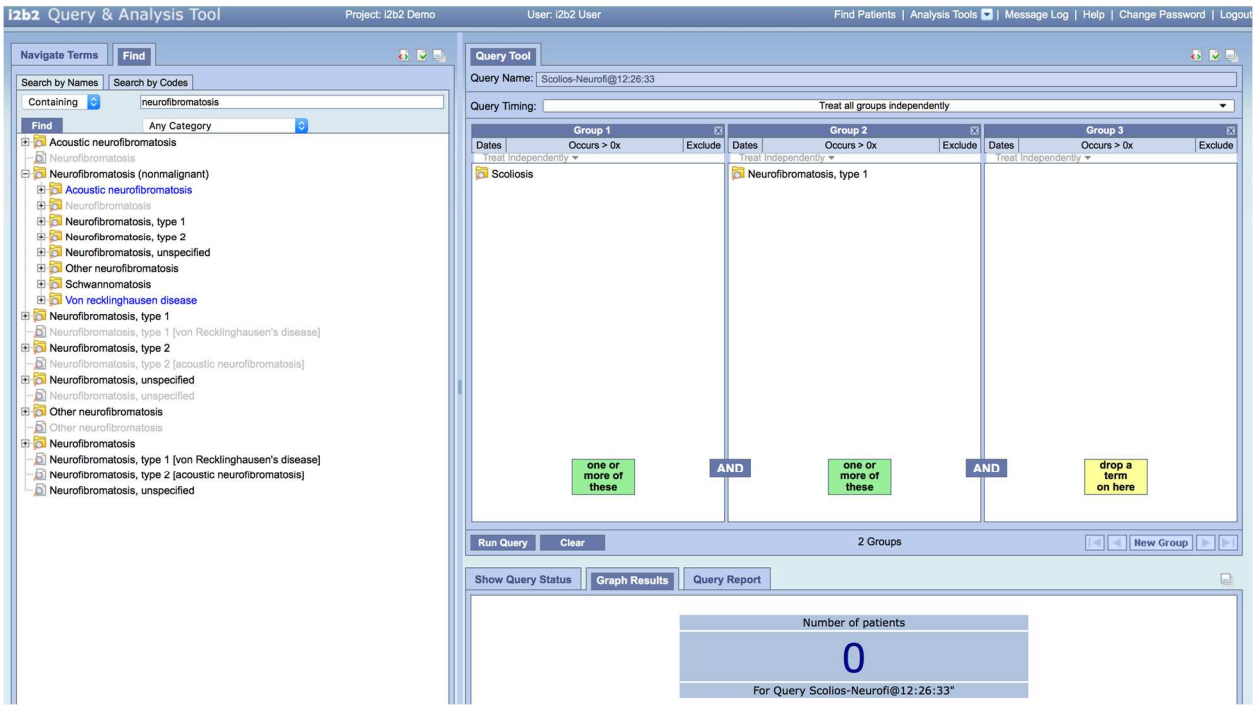

Figure 7: Screenshot of i2b2 when diagnoses are in the any encounter (question #20)

I found Leaf has a greater flexibility because it has the most obvious 'run query' button on the top of the page. The icons in Leaf are easy to understand and clear to see comparing to the SQL queries lines written in OMOP. It is also easy to see the name of the specific diagnoses in Leaf and i2b2 compared to OMOP. In OMOP, it is difficult to see what specific diagnoses were chosen after selecting the 'condition_concept_id' number into SQL queries.

In addition with the resource from Nicholas Dobbins, i2b2 is built within a one large data table; when users are running the query in i2b2, it runs on the whole table to find out the results. As in LEAF, the concepts are built individually and up to date. When users are running a query in LEAF, it runs separately and gets data from different small tables.

Data

The three query tools answered most of the questions. The data differences does not have any impact on the results because each query tool contains different number of patients and the main purpose of this study was only to show the capabilities of these query tools answering the questions.

Among all 20 questions, only question 12 had a similar number of patients shown in the result answered by the three tools. I assume may be it is because of the rare disease – 'acromegaly', few patients are found in all three query tools.

For question number 15, the results are not accurate because both LEAF and i2b2 cannot query notes, so we simplified the query by ignoring the two words "copying" and "pasting." We received this query from our University of Washington analytics team; the query question only contains data from UWMC and HMC. Because i2b2 does not use data from UWMC and HMC, we ignored the two institutions in i2b2. It still returned a result when excluding the institutions.

There are two numbers labeled in red, which are question #16 and #18 in the results of i2b2. I labeled these two results in red because i2b2 does not contain the information I need to query the questions. For example, question #16, the results showed up for Basal cell Carcinomas and Squamous cell Carcinomas, but they appeared in a grey color which I cannot choose to pull over to the query section. It made me realize how important it is to have the number of patients behind each concept when choosing the content we want.

OMOP has one 'N/A' labeled in blue because our OMOP model didn't include the content for question #2.

The two numbers labeled in green from Leaf are question # 13 and #18. Question #13 returned as zero because Leaf did not contain any patients with both cystic fibrosis and oral contraceptives. For question #18, both Leaf i2b2 could not find the correct concept for 'orthopedic surgery'. The concept for this query is a procedure, the search results in Leaf and i2b2 returned as a diagnoses or problem list.

Results	Description of data requests	LEAF Number of Patients	I2b2 Number of Patients	OMOP Number of Patients
1.	Patients age 18 and older discharged from hospital inpatient acute care without a diagnosis of venous thromboembolism (VTE) or obstetrics with a length of stay less than or equal to 120 days that ends during the measurement period	61833	10	795
2.	Initial Population Statement: Inpatient Encounters ending during the measurement period with Length of Stay (Discharge Date minus Admission Date) less than or equal to 120 days, and where the decision to admit was made during the preceding emergency department visit at the same physical facility Measure Population Statement: Equals initial population	20782	20	N/A
3.	Initial Population Statement: Emergency department encounters during the measurement period Measure Population Statement: Equals initial population	75318	20	95419
4.	Patients 18-75 years of age with diabetes with a visit during the measurement period	27743	102	1268
5.	All patients aged 18 years and older before the start of 2017 with at least one eligible encounter during the 2017 (measurement period)	326613	100	360663
6.	Patients 18-85 years of age who had a diagnosis of essential hypertension within the first six months of the measurement period or any time prior to the measurement period	177571	35	29176
7.	Patients aged 65 years and older, with 3 or more visits within a calendar year were identified.	4195	16	32538
8.	Patients aged ≥65 years admitted through the emergency department with a diagnosis compatible with CAP.	7726	25	203
9.	Patients discharged between January 1, 2011 and December 31, 2012, aged	28	41	1018

	40 years or older, on a bronchodilator or steroid inhaler.			
10.	Patients with T2DM, ≥1 office visit between 6 and 18 months before the index date, and with ≥1 glycosylated hemoglobin (HbA_{1c}) result in the 6-month preindex (baseline) period were included between November 2014 and February 2016.	58	1	12913
11.	Patients aged ≥65 years with chronic multimorbidity patterns (CMPs) identified.	16599	13	110
12.	Patients aged 65 years and younger admitted as Inpatient or outpatient between 2008-2013 with diagnosis of acromegaly	121	123	88
13.	Is there an interaction between hormonal contraceptives and cystic fibrosis medications that may affect the efficacy of the contraceptive method or the cystic fibrosis medication. Query: patient with a diagnosis of cystic fibrosis who have been prescribed with cystic fibrosis medications and also prescribed oral contraceptives.	0	11	5563
14.	Prove or disprove the hypothesis that intraoperative cardiac arrest has significant effects on mortality, graft survival, and perioperative morbidity. Query: patient with intraoperative cardiac arrest (LOS).	24	16	145
15.	Prove or disprove the hypothesis that 10% of inpatient records for UWMC and HMC Medicine services contain progress notes that contain clinically misleading content resulting from copying and pasting. Query: UWMC & HMC Inpatient records and diagnosis and procedures.	1143	133	1564
16.	PI needs counts of patients that had Basal cell Carcinomas, Squamous Cell Carcinomas, Squamous Cell	2695	0	241

	Carcinomas in situ, Displastic Nevus, any kind of Nevus, melanomas, etc. Broken down by years 2004 and 2005.			
17.	Female patients from January 2002 onwards with a diagnosis of DVT/PE and age less than 51.	2529	121	250
18.	PI needs a list of patients who have undergone Orthopedic surgery with a therapeutic INR on the day of surgery.	0	0	19282
19.	How many patients with neurofibromatosis type 1 and scoliosis (A).	17	5	0
20.	How many patients with neurofibromatosis type 1 and scoliosis (I).	21	0	2470

Table 7: Master list of all 20 query questions answered by three query tools

Table 7 is a master list of all 20 questions answered by the three query tools including the one question in Kahn's paper. The result differences in this table does not show any relationships between the three tools.

As can be seen in table 7, among 20 questions, i2b2 showed a significant number difference compared to Leaf and OMOP when answering the same questions. One reason is because i2b2 is using different data compared to Leaf and OMOP (both UW data), the other reason might be because of the demo version I used. The date range is not consistent to the current date in the demo version of i2b2, so I adjusted the dates to find the results. I chose only to use the demo version because this study is about evaluating the capabilities of these tools, not the evaluation of data.

Leaf and OMOP are using the same data to answer the same questions, however, the results still showed a huge difference. One reason might be the diagnoses I chose from these two query tools. For example, question #4 and #6 have diagnoses 'diabetes' and 'hypertension'.

When I searched for these two diagnoses, multiple results appeared relating to ‘diabetes’ and ‘hypertension’. The number of patients in each result is different, which made my results difference significant. Although the results contain significant differences, the gaps do not have any impacts between those tools.

Result 2: Perceived Ease of Use

Perceived ease of use is evaluated based on the different features between Leaf and i2b2, whether it is user-friendly for researchers to use. Throughout the exploration of the three query tools, I determined that LEAF has the highest perceived ease of use. Because LEAF is a self-service tool, unlike OMOP, users do not need to write SQL queries. As I mentioned earlier in my paper, not everyone can become a programmer. There are a huge amount of people who need access to data and cannot write SQL queries, OMOP will not be the best choice for those users. Therefore, OMOP has the lowest perceived ease of use. LEAF and i2b2, on the other hand, will make users more able to run queries without writing SQL.

I2b2 and LEAF are similar self-service query tools that are built up with SQL queries behind the back. They are both flexible by customizing date range, the number of occurrences, include/exclude. Between these two tools, I find LEAF has a greater flexibility than i2b2. LEAF can change the number of occurrence for each concept that is pulled into the same encounter in the same column. From table 8, LEAF can also customize age range and length of stay which i2b2 cannot. However, LEAF did not have a term called ‘length of stay’ in visit details at first. Nicholas Dobbins, the developer of LEAF, added after we emailed him about this problem, so it may contain some bias when writing the program for the ‘length of stay’. On the other hand, i2b2 is more developed than LEAF with the length of stay in visit details. Compared to i2b2,

LEAF also has a specific encounter, emergency, and discharge concept where users can specify the query questions.

Features	**LEAF**	**I2b2**
Show patient number	Yes	No
Table	Different small tables	One big table
Encounter concept	Yes	No
Emergency concept	Yes	No
Customize age range	Yes	No
Length of Stay (LOS)	Added when we requested, can customize	Exist, cannot customize
Discharge	Yes	No
Diagnosis	Specific	More specific
Query notes	No	No

Table 8: Comparison between the features of LEAF and I2B2

Table 8 demonstrates the different features between LEAF and i2b2. There 9 different types of features that I discovered while using these two query tools. LEAF shows a better flexibility due to 5 features where i2b2 does not have.

Additionally, LEAF has a function where it can show the number of patients behind each concept, but i2b2 does not. When choosing the concepts in i2b2, it is difficult to find out how many patients are in the concept. If there are zero patients, the result is likely to be zero. It is important to be aware of the number of patients when choosing the concepts. As in LEAF, you can choose the concepts with more patients which it will show a satisfied result.

Concepts can be in the same encounter when a query involves the word “without,” in LEAF. Users can choose the “and not” which represents “without” to exclude from the encounter. In i2b2, you must separate the “without” concepts with other concepts into two different groups

using the “exclude” function. When users are running the query in i2b2, it runs two separate columns together to combine the results.

Both LEAF and OMOP need permissions to proceed to future work with data, i2b2 is the only public version where users can download and use it without any restrictions. As mentioned earlier, getting approval for the secondary use of health care data is a challenge due to data privacy and confidentiality. Also, institutions are unclear how far the data is been used by researchers. By skipping the step to get permissions, users can easily access to health care data for research.

Discussion

The results of this research confirm that the three query tools answered most of the 20 questions, but each of the tools still has one or two questions that could not be answered. In my first initiation, the adopted OMOP could answer all query questions leaving LEAF to be the tool that is unable to answer most of the questions.

Comparing with LEAF and OMOP, i2b2 contains fewer data due to the demo version and the use of different data. The i2b2 and OMOP repositories do include a formal terminology layer but differ in how concepts can be hierarchically grouped together. After evaluating between the tools, Leaf has the highest perceived ease of use due to its flexible features and easy user-experience for non-programmers. Researchers could find the number of patients in a short time by pulling useful contents into the same encounter.

Finding 1: Perceived Usefulness – Performance

Our research found out that it is easy to use a common data model to complete complex queries, such as OMOP. Based on the comparison and the research study, OMOP ranked high in the overall evaluation criteria including flexibility and perceived usefulness. However, it has the lowest perceived ease of use due to writing long query sentences and requires trainings for users who are not familiar with SQL. As mentioned earlier, not every researcher can write programs. As for simple query questions, users need to perform long, complex lines of queries to answer the questions. For instance, question #4 is a simple query with patients 18-75 with diabetes in 2017. Users can simply use LEAF or i2b2 with 3-4 drags with minor changes: 1. drag and customize age group, 2. search for diabetes in the diagnosis and drag into the same encounter, 3. change the date range, 4. run the query. When this question applies to OMOP, 14

lines were written in SQL to achieve the result (Figure 9). Three tables were used in this question, switching back and forth between tables and finding the unique value which can represent each table is an essential step, it is important to join all these tables for OMOP to find the unique values. By comparing Figure 8 and Figure 10, it is clear to see i2b2 cannot customize age range like LEAF. If we want patients' ages between 18 – 75 there are two ways in i2b2: 1. include all patients age from 18 to 74 by dragging 5 age groups; 2. Drag the remaining 4 age groups and click 'exclude' in the group 2 .

I also discovered when searching a diagnosis in LEAF or i2b2, it is straightforward for users to use a search bar, but in OMOP, I wrote 5 lines to gain the results for one diagnosis.

From the overall perspective, LEAF is the most appropriate performer among all three query tools according to its flexibility features and high performance. Leaf and i2b2 answered 18 questions, and OMOP answered 19. I found the results surprising when there is only one question difference between Leaf and i2b2 compared to OMOP. From this research study, self-service query tools are able to perform as well as common data query tools.

After comparing the two self-service query tools, I found LEAF has an easier user experience than i2b2. I2b2 was established in 2004 and LEAF was developed in 2016. My first thought is that i2b2 would be a more mature and well-developed tool, but as I was using them both, LEAF impressed me with its flexible temporal criteria including customizing date and age range; choosing admissions source for inpatient; and excluding concepts in the same column. 68% (13) of data requests in the sample contained temporal criteria. However, LEAF did not have 'length of stay' feature at first comparing to i2b2. Users in i2b2 do not have this concern.

Based on the evaluation of i2b2 on predicting asthma exacerbations, it is a user-friendly query and visualization tool for researchers to improve their access to clinical data[27]. I2b2 contains large terminology with enriched laboratory exam and medication concepts, it could answer almost all queries required for this research study. With the free and open source cohort selection tool of i2b2, more researchers would benefit from using i2b2 for clinical research. I found the terminology searching and browsing functionalities very intuitive and easy to use while I was using i2b2. It can build queries by simply drag-and-dropping terminology concepts into different criteria groupings.

Both Leaf and i2b2 use similar methods to determine the number of patients, they both work well with simple standard questions involving diagnoses and a limited number of demographic data elements. But some diagnoses in i2b2 such as 'Squamous Cell Carcinomas' appeared grey after searching. It appeared in grey may be because they do not contain any patients. This is why the number of patients behind each concept is necessary. I could not choose the ones in grey which leads me zero number of patients compared to LEAF. In i2b2, there are some challenges related to terminology encountered which related to the inconsistent use of controlled medical vocabularies when representing laboratory tests and results in the different clinical information system[27].

The results are somewhat surprising because the results in the table show LEAF can answer the same amount of questions as i2b2 and only one question less than OMOP. Also, I was amazed by how flexible LEAF is with its varies features when using the system; LEAF would benefit a lot of non-programmers who need access to health care data.

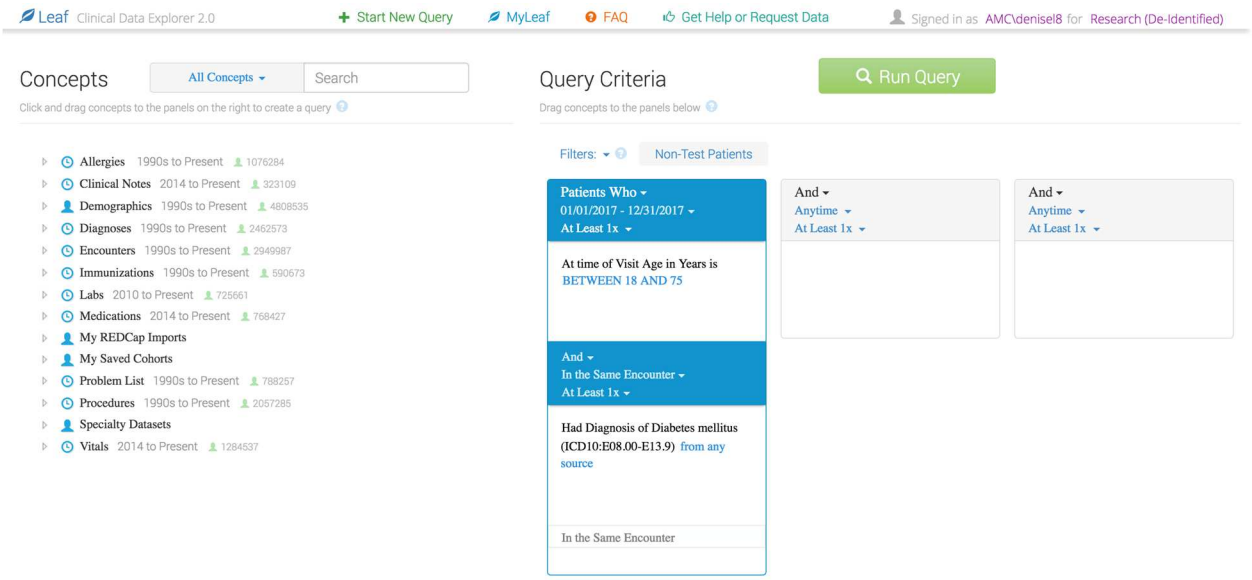

Figure 8: Screenshot of LEAF; Find patients for question #4.

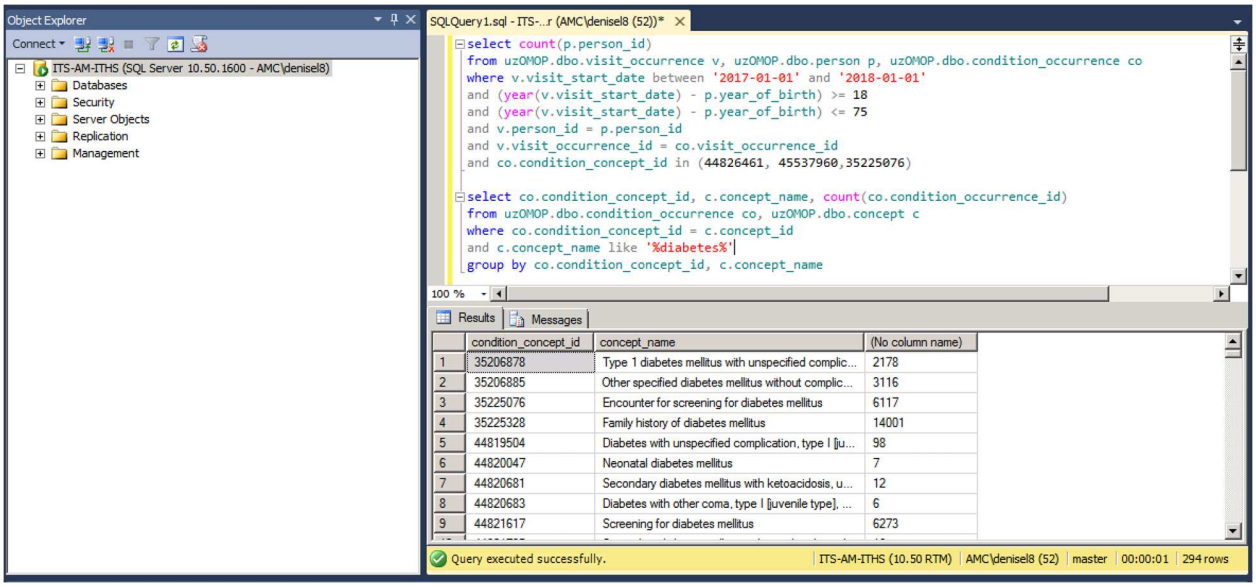

Figure 9: Screenshot of OMOP; Find patients for question #4.

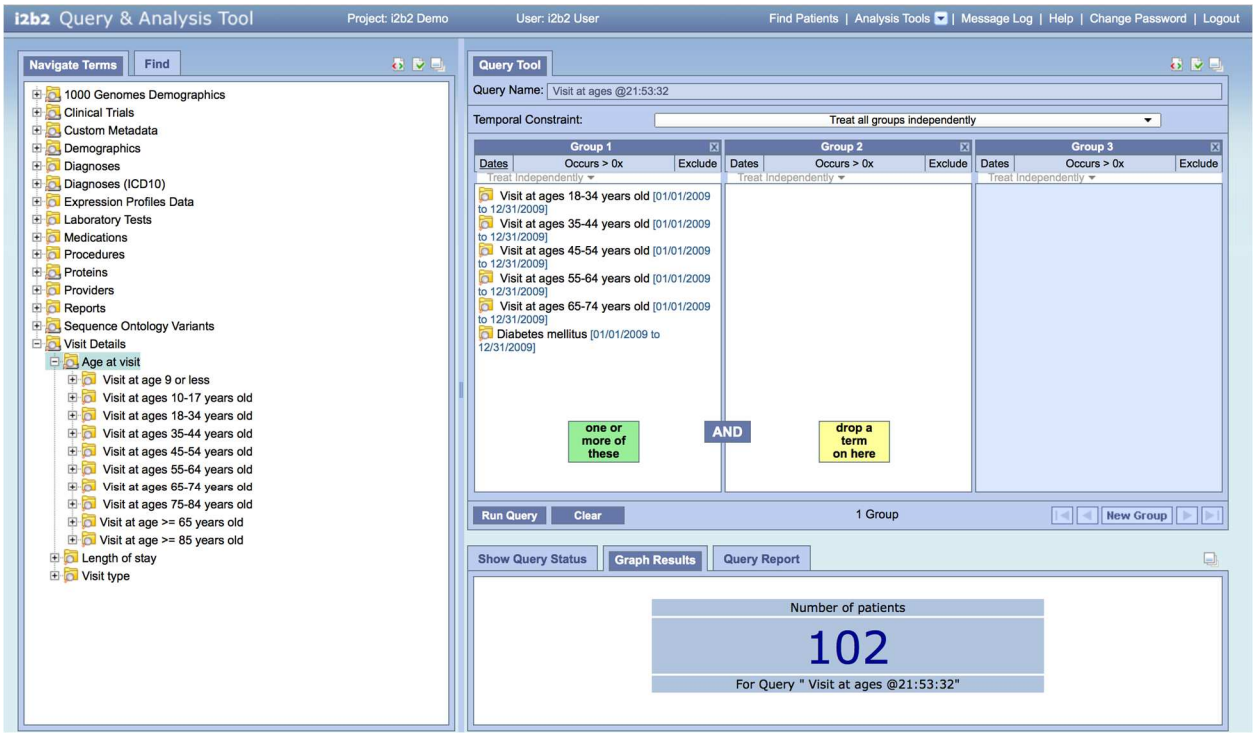

Figure 10: Screenshot of i2b2; Find patients for question #4.

Finding 2: Query Process

I found strengths and limitations of the three different types of query tools under each specific circumstances when writing queries. I made some changes to my questions while I was writing my queries. Before I started to write the queries questions I do not know what information is exactly needed because I was not familiar with the tools. The previous query questions I found did not returned any patients in any of the tools because they were not actually a query question. They did not contain elements such as patient age, diagnosis, and measurement period. Without these elements, it is difficult to use either self-service tools or common data models.

When I started to write queries in OMOP, I faced some challenges. I learned SQL in my bachelor's degree, but I have not used it for a long time. Before writing the SQL queries, I dig

into some research by reading the SQL cookbook and learning SQL online. At first, it was quite difficult for me to write SQL queries because I did not understand the grammar. Also, I am not familiar with the OMOP tables; I read through each table and initiatives to make sure I understand the relationships between the tables I need. With the help of others and getting my memories back, writing SQL became much easier. If researchers choose to query questions in OMOP, they need a lot of trainings from other professionals and it would be time-consuming. Throughout my testing, I found some of the same results according to the evaluation of i2b2 study[27], the study evaluated the usability and functionality of i2b2 using several real world examples of research data requests. i2b2 is a useful tool containing medical terminologies for diagnosis and procedures, medications, and laboratory tests, along with mapping legacy data and generating structured and coded rich metadata. One significant strength of i2b2 is the ability to estimate patient cohort sizes by changing the inclusion/exclusion criteria[27]. In addition, i2b2 also provides terminologies for patient demographics including age, gender, income, language, marital status, race, religion, vital status, and zip codes. The i2b2 data model is valuable when the type of data requests are structured and straightforward without using pre-coordinated concept, which can also be transformed to fit the i2b2 logical data model. However, there is still limitations in i2b2; it has a column called 'modifier_cd' which contains information related to the ranking of modifiers[27]. This column is currently not used by the visual query client tool; the end-users have no way of querying the data using relationships between observations. This limitation requires professional training before use. I2b2 is only allowed to present simple ranking/order relationship, which causes the complex relationships to be lost. Due to the limitation of i2b2, individual clinical observations can only be stored and retrieved at

the same level of an 'encounter' or a 'visit.' For complex relationships, it is necessary to create another pre-coordinated metadata concept outside of i2b2[27].

The inadequacy of provisions for preserving contexts and relationships between individual clinical observations is a shortcoming of the i2b2[27]. This limitation is important when importing data from structured clinical documentation into i2b2, every atomic observation will have to be available as a pre-coordinated concept. The shortcoming is not practical when in an environment with a large-scale EMR implementation. One of the most significant limitations of i2b2 is importing data from structured clinical documentation into i2b2[27]. Along with the current EMR vendors is not yet standardized in the United States, it may limit the researcher's ability to use data in a consistent manner.

There are some temporal criteria such as setting the date range , it was relatively easy to model in the queries using i2b2. Question #5 and #6 as examples, the data requested the time range before the measurement period and during the measurement period (#5); another query requested the date range within the first 6 months of the measurement period or any time prior to the measurement period. In i2b2 it requires multiple runs using two separate datasets and combing together to get the final results, but it would be much easier using a single SQL query in OMOP to express this logic.

According to a journal in AMIA, Overhage et al. concluded OMOP CDM has the possibility of losing data by forcing disparate sources into one common model[31]. When mapping between concepts in different terminologies, various types of complexities may be encountered.

Data query structures are often determined by the structure of the database being queried. Answering the same question may result in different structures when different databases are

used. The types of questions that can also be asked depends on the specific structure of a database according to Kahn's paper[35]. A database that does not represent encounters cannot answer questions about encounters. For example, if a question is asked about an insurance membership database, the three query tools are not able to answer the question[35].

Although LEAF contains the highest flexibility, it is still a local query tool which is not publicly used outside of UW Medicine. OMOP and i2b2 are open-source tools that can be easily download and use. However, the OMOP data I used in this research study requires permissions because the data is from UW Medicine. Because Leaf is a local query tool, one limitation appears as users could only find the information of patients inside UWMC, HMC, and NWHMC when query questions.

Finding 3: Evaluation on Data Elements – Terminology Coverage

In my study, I found OMOP CDM contains the largest medical terminology compared to LEAF and i2b2. The OMOP CDM accommodated the highest percentage of data elements. One example is question #18 which asks for patients who have undergone Orthopedic surgery with a therapeutic INR. In LEAF, I could not find the exact word for 'Orthopedic surgery.' The most equivalent word I found is 'orthopaedic surgical,' however, it was still not the result that I was looking for. I did not choose the similar terminology due to different concepts. The word 'Orthopedic surgery' is a procedure in the query question, Leaf returned the similar word as diagnoses or problem list which do not match the same concept. I2b2 appears with similar results as LEAF and the word 'therapeutic' was not detailed enough in i2b2 and displayed over 200 results. OMOP presented 86 results after searching the word 'therapeutic' and 50 rows for 'orthopedic surgery,' it generated the results in a more specific term by capturing most of the

data elements and domains. The evaluation of common data models also stated OMOP had a 100% match of the data model constraints compared to other models, and 100% of data integration of terminology coverage[24].

OMOP works best with complex queries as it might be challenging for self-service tools to run the queries and writing long SQL queries also may take a long time which decreases efficiency[45]. In OMOP, users can write down each line of queries depending on the questions and their own solution; it is more flexible when writing your own queries and each user might have a different solution to the same question. As stated in a research study, a CDM can be used to minimize variability and enable common interpretation within the context of underlying source data[45]. The same study also confirmed that CDMs are a feasible and useful approach to allow systematic analysis of health care data sources, administrative claims and EMR data.

With the strengths of implementing standardized vocabularies in OMOP, it can improve the integration of various types of data sources.

Finding 4: Calculation Performance

Leaf and i2b2 cannot perform calculations using the current version compared to OMOP CDM. Due to the fact that LEAF and i2b2 can only perform simple criteria queries it is incredibly challenging for these two tools to calculate or aggregate values. There is a column called measurement description back to table 1 and table 2; the measurement description asks users to calculate the percentage and median of patients in both numerator and denominator. OMOP can merely calculate the numbers by writing math formulas in SQL, however, for LEAF and i2b2, there is no such function to execute calculations. The main reason I chose to only query the questions in the denominator is because there is no calculation function in LEAF and i2b2.

Deshmukh et al. evaluated i2b2 on clinical research and discovered i2b2 could not complete query complex data requests containing inclusion or exclusion criteria such as calculations and aggregate data[27]. Aggregating data would address post-processing the observations after obtaining an initial data set. Version 1.3 of i2b2 software strengthened the ability to include conditions based on values of certain findings[27]. This specific function benefits in other types of queries, but it does not directly provide the needs to inclusion or exclusion criteria.

Finding 5: Accessing Data

We faced access challenges at the beginning; I used to have access from the previous quarter for LEAF using the AMC account. However, I did not log in to my account during the summer, my account was deleted. My professor and I asked for permissions at the beginning of spring quarter, and it took two weeks for them to reply. After they replied to our email, I was asked to complete a form. The access took another two weeks, so I was not able to use LEAF for almost a month.

In the meantime, I focused on researching the information I need and find relevance scholars and articles relating to my topic. I also finished CITI training because I was planning to invite different levels of query writers to complete the 9 types of queries I created in order to compare and contrast, it ultimately became a barrier due to many challenges.

Getting familiar with OMOP and i2b2 is another task I achieved before getting access to LEAF. I tested some simple and complex queries in i2b2 and dug deeper into OMOP common model database online. During our meeting at the beginning of March, we finally tested our access to ensure I have got all the access I need for LEAF and OMOP. I2b2 has an online demo version; we do not need access to i2b2. Because it is a demo version, we do not need to worry about data

leakage and data confidentiality. In LEAF, we used de-identification in our research to reduce risk.

We paired up our computers and used Microsoft Remote Desktop, after I downloaded this application we added the connection name we need, and accessed OMOP. Getting access to data is always the hardest step to start a research study. Although data querying for clinical trials is one way to promote the fair and transparent conduct of clinical trials. It would benefit the exploration of additional hypotheses and maximize the use of data for various purposes; some practical issues arise which need to be addressed[46].

One of the main reasons related to my data access is remote access platforms facing technical issues with software limitations and internet requirements. Also, LEAF cannot be opened with Safari, downloading Google Chrome is another step to access LEAF. As for OMOP, an application must be download to add user's name before logging into the system. When accessing to LEAF and OMOP an external VPN is needed. Also, the LEAF developers need to understand the purpose of me using LEAF, and multiple emails need to be sent out to get the approval.

Limitations

Time

Expressing the differences between the queries in English requires very careful articulation of the relationships between encounters, diagnosis, providers and time[35]. With different users querying the same questions, the results may be varied according to query writers' understanding of each question. Initially, we wanted to involve more query writers to answer the 20 questions and compare to my results, are these tools returning the same results or

different. If the results are different what are the differences when we query the questions, did we choose different codes for the same diagnoses? However, due to the time limitation, it is a big challenge to find different levels of query writers to answer all 20 questions. Also, I am using three query tools to determine the results, not every query writer exactly knows how to use all three tools, and it will take them a lot of training to explore the functions. We then narrowed down to find 1 to 2 people to complete only one query question, but we still faced obstacles due to finding the right people and finding the time to train them.

If I had enough time for this research study, I would invite 10 query writers based on different levels and time themselves while completing the query for each question. Users and I could record the time on the timer on each of the query tools; I could compare their queries and time to find out the most accurate query with the least time used. One of our limitations is we did not find others to answer the same query questions and compare the results due to time limitation.

Design

When choosing our query questions from the analytics team, they only sent us the three questions they wanted us to query; it may contain some personal bias while he was selecting the queries for us. It would be best if we can have a list of query questions and randomly choose from the list. The three queries the analytics team gave us are not really query questions, as I started to query, I found the three query tools could not answer any of those three questions. We changed and simplified the three questions into query questions; this may also contain bias from my professor and me.

It is best to choose more questions to answer instead of only answering 20 questions; the results would be more persuasive and precise according to the number of questions been answered. Originally our goal was to answer 50 questions, 10 questions for each type, due to the limited resources and time limit we only examined 20 questions.

Involving other users analysis and opinions on the three tools would improve my evaluation. I would include more users in my study and provide them with a user-experience survey after they answered all 50 questions in three different tools. The survey would consist of features in LEAF and i2b2 according to Table 8; it would also include users thoughts after returning the results, were the results expected, surprised, or mixed. In the end, users would choose which is the best performer and state out the reasons.

Scope

The University of Washington is the only institution we used in this research study; we lack limited resources to develop a more integrated design. Although getting access to other resources might be a challenge, involving more institutes can improve the accuracy of data and increase the amount of resources to retrieve data. We could also compare the results with more query tools, including SQL or other self-service query tools would enhance the efficiency of data and results. Seattle Children's Hospital has its own self-service query tools developed and used in the health informatics department, getting data from other hospitals can also strengthen the final results. Only using the three query tools is another limitation we had in our research study. Self-service query tools and common data models are not only used in health care industries but also a variety of other industries including customer service and business intelligence. Digging deeper into other fields can also help to improve my understanding of

SSQT and CDMs on how it is performed in a non-health care environment and what are some changes comparing to health-related field.

Conclusion

We found LEAF performs the highest flexibility, ease of use, and usefulness among all three query tools by answering 20 query questions chosen from 3 different categories. But each of the query tools has various strengths and limitations according to different circumstances. With the development of self-service query tools, many non-query writers in health care would benefit from this technology by avoiding trainings to save their time.

Self-service query tool - LEAF performs best when query writers need data from UWMC, HMC, and NWMC based on its well-developed and flexible features. Due to many customized features, the user experience of LEAF will increase the efficiency when answering query questions such as answering 'the number of patients' in certain conditions. However, users require permissions to grant access to LEAF by stating the purpose of their use. A VPN is required to log in to LEAF, and a specific web browser (Google Chrome) is needed as well. Compared to LEAF, i2b2 is a free and open source software that welcomes every user to download online for diverse reasons. Open-source tools are important in health care because it ensure unrestricted access to data, users can also deploy the data as widely as they like and alter them to their needs. Although i2b2 has been developed for more than 10 years, some features are not flexible enough to customize queries criteria. This lack of functionality would decrease the efficiency and accuracy of the returning results. These two self-service query tools

work best when users are not familiar with writing SQL queries, LEAF and i2b2 are built on writing SQL behind the system.

OMOP, the common data model is also a flexible query tool for users who can complete well-written SQL queries. OMOP contain the most extensive medical vocabularies and contents as a comparison to LEAF and i2b2; it is a reliable tool for researchers when they are searching for complex diagnoses. OMOP has been well-used in both University of Washington and Seattle Children's Hospital; it is a more comprehensive query tool than LEAF. Because OMOP is based on writing SQL queries, it will maximize the flexibility of each user for the same question. As mentioned earlier, one query question can have 9 or more different answers according to each query writers' understanding of different clinical scenarios and creativity. Although OMOP is also a public source for standardizing the format and content of observational data, retrieving data from OMOP is a challenging step for researchers. Getting access to secondary use of health care data for researchers is still a barrier, accessing data can benefit researchers from gaining more health care information to improve the health care services and expand Big Data usage. But data confidentiality is one of the essential concerns because permissions and restrictions are required for the reuse of health care data. With the development of more open-source self-service query tools, accessing to secondary use of health care data would overcome the current barriers in the future.

NORC data is one particular approach to data access that has been adopted and used by many agencies[48]. This remote access data technology provides researchers with data augmentation and knowledge sharing. It allows researchers to access confidential microdata from their offices by using statistical, technical, and legal protections. Operational protection controls the limit of

researchers' access to information and ensures they are using the data for their specific needs, audit logs, trails, and webcams are also provided to monitor researchers activities[48]. The quality of analysis is increased due to the use of NORC data; it provides permit data archiving, indexing, and curation.

The ideal result of the self-service tool is to test the ability to put up a piece of software of each medical center such that a single query could be written that would go off and find out how big the cohort was that in each of the institutions. The dream was can we do national trial recruitment, can we have visibility to the power that we might be able to bring to the national clinical trials through this technology. In this study, we examined 9 different types of queries by randomly choosing them. It is best to find more types of queries to make our selections more vary. Also, we determined 3 queries for each type, for more additional work choosing more queries would increase the level of accuracy of results.

The result of this study performs the abilities to answer different query questions in both self-service query tools and common data model in health care data. With these query tools developed and widespread, data access to health care for researchers would overcome a big step in the future.

Acknowledgments

I would like to thank my committee members Dr. Adam Wilcox and Dr. John Gennari for their contribution to this work. I also want to give my thanks to Tony Black who helped a lot on getting access and permissions to LEAF with Dr. Adam Wilcox. Nicholas Dobbins and other developers of Leaf for their help with getting familiar, developing and using LEAF by coaching me one-on-one for 2 hours, and for answering multiple LEAF questions arising from my studies. Nicholas Dobbins also helped by adding the 'length of stay' concept in LEAF to make my study easier to query. Thank you all for your patience, understandings, and support to this research study.

List of Figures

Figure Number Page

List of Tables

Table Number Page

Reference

1. Sandhu, E., Weinstein, S., McKethan, A., Jain, S. H. (2012). Secondary Uses of Electronic Health Record Data: Benefits and Barriers. *The Joint Commission Journal on Quality and Patient Safety.* https://doi.org/10.1016/S1553-7250(12)38005-7

2. Harry, C. (n.d.). Secondary use of data – striking a balance. *National Cancer Registry.*

3. Cheng, H. G., Phillips, M. R. (2014). Secondary analysis of existing data: opportunities and implementation. *Shanghai Archives of Psychiatry.* http://doi.org/ 10.11919/j.issn.1002-0829.214171

4. Cole, A. M., Pflugeisen, B., Schwartz, M. R., & Miller, S. C. (2018). Cross sectional study to assess the accuracy of electronic health record data to identify patients in need of lung cancer screening. *BMC Research Notes*. https://doi.org/10.1186/s13104-018-3124-0

5. Boruch, R. F. (1985). Definitions, products, distinctions in data sharing. In S. E. Fienberg, M. E. Martin, & M. L. Straf (Eds.), Sharing research data (pp. 89-122). Washington, DC: National Academy Press.

6. Zimmerman, A.S. (2003). Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists. *University of Michigan Library.*

7. Campbell, E. G., Clarridge, B. R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N. A., & Blumenthal, D. (2002). Data withholding in academic genetics: Evidence from a national survey. JAMA 287(4), 473-480.

8. Burton, P.B., Banner, N., Elliot, M.J., Knoppers, B.M., Banks, J. (2017). Policies and strategies to facilitate secondary use of research data in the health sciences. International Journal of Epidemiology. https://doi.org/10.1093/ije/dyx195

9. U.S.-Nas. (2009). *Ensuring the integrity, accessibility, and stewardship of research data in the digital age*. *Science (New York, N.Y.)*. https://doi.org/10.1126/science.1178927

10. Huff, S. (2017). Open data sharing will improve care, lower costs. Retrieved June 8, 2018, from https://www.athenahealth.com/insight/open-data-sharing-will-improve-care-lower-costs

11. Van Panhuis, W. G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A. J., …

Burke, D. S. (2014). A systematic review of barriers to data sharing in public health. *BMC Public Health*. https://doi.org/10.1186/1471-2458-14-1144

12. Safran, C., Bloomrosen, M., Hammond, We., Labkoff, S., Markel-Fox, S., Tang, P. C., & Detmer, D. E. (2007). Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. *Journal of the American Medical Informatics Association*. https://doi.org/10.1197/jamia.M2273.Introduction

13. Meystre, S. M., Lovis, C., Bürkle, T., Tognola, G., Budrionis, A., & Lehmann, C. U. (2017). Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. *Yearbook of Medical Informatics*. https://doi.org/10.15265/IY-2017-007

14. Weiner, M. G., & Embi, P. J. (2009). Toward reuse of clinical data for research and quality improvement: The end of the beginning? *Annals of Internal Medicine*. https://doi.org/10.7326/0003-4819-151-5-200909010-00141

15. Lane, J., & Schur, C. (2010). Balancing access to health data and privacy: A review of the issues and approaches for the future. *Health Services Research*. https://doi.org/10.1111/j.1475-6773.2010.01141.x

16. Carlson J, Stowell-Bracke M. Data management and sharing from the perspective of graduate students: An examination of culture and practice at the Water Quality Field Station. Libraries and the Academy. 2013;13: 343–361. doi: 10.1353/pla.2013.0034

17. Faniel IM, Zimmerman A. Beyond the data deluge: a research agenda for large-scale data sharing and reuse. International Journal of Digital Curation. 2011;6: 58–69. doi: 10.2218/ijdc.v6i1.172

18. Rodriguez V. Access to data and material for research: putting empirical evidence into perspective. New Genetics and Society. 2009. February 19;28(1):67–86. doi: 10.1080/14636770802670274

19. Tenopir C, Allard S, Douglass K, Aydinoglu A, Wu L, Read E, et al. Data sharing by scientists: Practices and perceptions. PLOS ONE. 2011;6: e21101 doi: 10.1371/journal.pone.0021101

20. Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., … Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0134826

21. Kohane, I. S., van Wingerde, F. J., Fackler, J. C., Cimino, C., Kilbridge, P., Murphy, S., … Szolovits, P. (1996). Sharing electronic medical records across multiple heterogeneous

and competing institutions. *Proceedings : A Conference of the American Medical Informatics Association / ... AMIA Annual Fall Symposium. AMIA Fall Symposium*.

22. Hruby, G. W., Ancker, J., & Weng, C. (2015). Use of Self-Service Query Tools Varies by Experience and Research Knowledge. In *Studies in Health Technology and Informatics*. https://doi.org/10.3233/978-1-61499-564-7-1023

23. Informatics for Integrating Biology & the Bedside. (n.d.). Retrieved May 29, 2018, from https://www.i2b2.org/about/index.html

24. Garza, M., Del Fiol, G., Tenenbaum, J., Walden, A., & Zozus, M. N. (2016). Evaluating common data models for use with a longitudinal community registry. *Journal of Biomedical Informatics*. https://doi.org/10.1016/j.jbi.2016.10.016

25. OMOP Common Data Model. (2018). Retrieved May 29, 2018, from https://www.ohdsi.org/data-standardization/the-common-data-model/

26. Voss, E. A., Makadia, R., Matcho, A., Ma, Q., Knoll, C., Schuemie, M., … Ryan, P. B. (2015). Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *Journal of the American Medical Informatics Association*. https://doi.org/10.1093/jamia/ocu023

27. Deshmukh, V. G., Meystre, S. M., & Mitchell, J. A. (2009). Evaluating the informatics for integrating biology and the bedside system for clinical research. *BMC Medical Research Methodology*. https://doi.org/10.1186/1471-2288-9-70

28. Chuttur, M. (2009). Overview of the Technology Acceptance Model: Origins, Developments and Future Directions. *Association for Information Systems.*

29. Abdekhoda, M., Ahmadi, M., Dehnad, a, & Hosseini, a F. (2014). Information technology acceptance in health information management. *Methods Of Information In Medicine*. https://doi.org/10.3414/me13-01-0079

30. Holden, R. J., & Karsh, B.-T. (2010). The technology acceptance model: its past and its future in health care. *J Biomed Inform*. https://doi.org/10.1016/j.jbi.2009.07.002
31. Overhage, J. M., Ryan, P. B., Reich, C. G., Hartzema, A. G., & Stang, P. E. (2012). Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association : JAMIA*. https://doi.org/10.1136/amiajnl-2011-000376

32. Taherdoost, H. (2017). A review of technology acceptance and adoption models and

theories. *ScienceDirect.* https://doi.org/ 10.1016/j.promfg.2018.03.137

33. Tella, A., & Olasina, G. (2014). Predicting Users' Continuance Intention Toward E-payment System. *International Journal of Information Systems and Social Change*. https://doi.org/10.4018/ijissc.2014010104

34. Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., … Schilling, L. (2016). A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Washington, DC)*. https://doi.org/10.13063/2327-9214.1244

35. Kahn, M. G. (Unpublished). Querying Clinical Databases: How Many Patients?

36. Eligible Hospital / Critical Access Hospital eCQMs. (2018). Retrieved May 29, 2018, from https://ecqi.healthit.gov/eligible-hospital-critical-access-hospital-ecqms

37. Eligible Professional / Eligible Clinician eCQMs. (2018). Retrieved May 29, 2018, from https://ecqi.healthit.gov/eligible-professional-eligible-clinician-ecqms

38. Shen, Y., Chien Tay, Y., Wee Kwan Teo, E., Liu, N., Wei Lam, S., Eng Hock Ong, M., & Author, C. (2018). Association between the elderly frequent attender to the emergency department and 30-day mortality: A retrospective study over 10 years. *World J Emerg Med*. https://doi.org/10.5847/wjem.j.1920–8642.2018.01.003

39. Toledo, D., Soldevila, N., Torner, N., Pérez-Lozano, M. J., Espejo, E., Navarro, G., … On-behalf of the Project FIS PI12/02079 Working Group. (2018). Factors associated with 30-day readmission after hospitalisation for community-acquired pneumonia in older patients: a cross-sectional study in seven Spanish regions. *BMJ Open*. https://doi.org/10.1136/bmjopen-2017-020243

40. Nguyen, H. Q., Chu, L., Amy Liu, I.-L., Lee, J. S., Suh, D., Korotzer, B., … Gould, M. K. (2014). Associations between Physical Activity and 30-Day Readmission Risk in Chronic Obstructive Pulmonary Disease. *Annals of the American Thoracic Society*. https://doi.org/10.1513/AnnalsATS.201401-017OC

41. Yu, M., Mody, R., Landó, L. F., Shui, A., Kallenbach, L., Slipski, L., & de Oliveira, C. P. (2017). Characteristics Associated with the Choice of First Injectable Therapy Among US Patients With Type 2 Diabetes. *Clinical Therapeutics*.

42. Ibarra-Castillo, C., Guisado-Clavero, M., Violan-Fors, C., Pons-Vigués, M., López-Jiménez, T., & Roso-Llorach, A. (2018). Survival in relation to multimorbidity patterns in older adults in primary care in Barcelona, Spain (2010-2014): A longitudinal study based on

electronic health records. *Journal of Epidemiology and Community Health*. https://doi.org/10.1136/jech-2017-209984

43. Silverstein, J., Roe, E., Munir, K., Fox, J., Emir, B., Kouznetsova, M., … King, D. (2018). USE OF ELECTRONIC HEALTH RECORDS TO CHARACTERIZE A RARE DISEASE IN THE USA: TREATMENT, COMORBIDITIES AND FOLLOW-UP TRENDS AMONG PATIENTS WITH A CONFIRMED DIAGNOSIS OF ACROMEGALY. Retrieved May 29, 2018, from https://www.ncbi.nlm.nih.gov/pubmed/29624099

44. Meystre, S. M., Deshmukh, V. G., & Mitchell, J. (2009). A clinical use case to evaluate the i2b2 Hive: predicting asthma exacerbations. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*.

45. Reisinger, S. J., Ryan, P. B., O'Hara, D. J., Powell, G. E., Painter, J. L., Pattishall, E. N., & Morris, J. a. (2010). Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. *Journal of the American Medical Informatics Association : JAMIA*. https://doi.org/10.1136/jamia.2009.002477

46. Mbuagbaw, L., Foster, G., Cheng, J., & Thabane, L. (2017). Challenges to complete and useful data sharing. *Trials*. https://doi.org/10.1186/s13063-017-1816-8

47. Lane, J., & Schur, C. (2010). Balancing access to health data and privacy: A review of the issues and approaches for the future. *Health Services Research*. https://doi.org/10.1111/j.1475-6773.2010.01141.x