DEVELOPMENT OF SEMI-AUTOMATED TOOLS TO MAP CANCER RESEARCH COMMON DATA ELEMENTS TO THE BIOMEDICAL RESEARCH INTEGRATED DOMAIN GROUP MODEL

A DISSERTATION

SUBMITTED TO THE FACULTY OF

THE UNIVERSITY OF MINNESTA

BY

ROBINETTE RENNER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

ADVISOR: GUOQIAN JIANG, M.D., PH.D.

CO-ADVISOR: CHAD MYERS, PH.D

March 2020

Acknowledgements

As with any dissertation, my work would not have been possible without the support of my committee, research colleagues, professional colleagues, friends, and family. I want to thank my committee for their invaluable feedback on my writing and during my presentation rehearsals. I would especially like to thank my advisor, Dr. Jiang, for his help during all aspects of my research. My co-advisor, Dr. Myers, provided helpful mentorship, and his proposal writing class gave me some much-needed momentum.

The collaborative efforts of the research team at the University of South Alabama made the Artificial Neural Network project possible. I would especially like to thank Shengyu Li and Dr. Ryan Benton. Shengyu did all of the algorithm implementation work, helped answer my questions, and modified the code as needed. Dr. Benton patiently answered my questions during late evening phone calls, and his modifications to the final manuscript resulted in a much stronger paper.

I want to thank my professional colleagues, Dr. Abeer Madbouly and Jane Pollack, for the laughter that made my Ph.D. journey far less painful. Abeer was an excellent mentor who never failed to give sound advice when I stopped by her cube to vent. As my research progressed, she helped celebrate my successes. Not only did Jane willingly share her research with the BRIDG model, but she also commiserated with me about the joy and wonder of working full time while going to graduate school.

Dr. Robert Freimuth is neither a research colleague nor a professional colleague. However, he stepped in and filled a gap during a critical time when I was most at risk of walking away from my doctoral work. After the passing of my first advisor and the

serious illness of my new advisor, I struggled with a lack of guidance and mentorship. Dr. Freimuth kept me on track and ensured that I was doing quality work.

Several people walked with me along this journey but sadly did not make it to the finish line. First and foremost was my dad. He gave me the greatest gift, his analytical mind. He supported me through my academic endeavors from my first science fair project (Microbial Growth in Aircraft Fuel Tanks) to my first publication. He was so proud of that publication and had my mother read it to him several times. I wish he were here to read the final dissertation, though I am sure that my mom is glad that she does not have to read out loud a hundred-page manuscript.

Dr. Frank McKinney was my first thesis advisor more than twenty-five years ago. Not only did he make me think differently about Great Blue Herons, but he also became a surrogate grandfather who changed the trajectory of my life. When he passed, his wife, Meryl, stepped into his shoes. During our many lunches together, she was always encouraging me to reach further. When I went back to school for a second Masters degree, she was dismayed that I was not pursuing a Ph.D. I know that they would be so happy that I finally completed the degree that I started so long ago.

Dr. John Carlis was the type of advisor that other advisors should aspire to be. He always put the needs of the student first and took an active role in the research and writing process. So many times, he insisted on seeing my “messy” writing even when I felt that it was not ready to be seen. During our weekly meetings, he challenged my thinking and encouraged me to embrace the “joyful struggle” of academic research. His help, guidance, and friendship encompassed 15 years and two degrees. I only wish that he lived long enough to add this thesis to the long line of other students' theses on his bookshelf.

Although I have joy and sadness about the end of this particular "joyful struggle," I am fortunate to have a strong support team that helped me climb the academic Mount Everest. My mom has been there from the beginning and has always played an active role in my academic work. I have fond childhood memories of sitting on the sofa with her while she read my textbooks and quizzed me on the content. No matter how challenging the subject, she managed to find just the right questions to ask. As I worked through my doctorate, she was always interested in my research and asked me how things were going. She never once complained when I grumpily responded: "No, I am not finished yet." I am so happy that she is able to share my joy of finally being finished with school.

My fabulous bonus daughters, Carmen [1] and Ava [2], have endured the past 5 ½ years without complaint. Too many times, their dad warned them that I was working against a deadline and would not be good company. They never took offense and sometimes would quietly sneak into my office to bring me fresh baked cookies. What I hope that they take away from my graduate school experience is that it is never too late to pursue your dreams and that with a lot of hard work, you can accomplish nearly anything.

The most critical factor in my success has been the love and support of my husband, Jim. There never was any doubt in his mind that I should return to school, and he rode the roller coaster ride of graduate school while maintaining a sense of humor. From the high points of getting articles accepted for publication to the low point when Dr. Carlis passed, he was there, a constant, steady influence that allowed me to continue even when it seemed impossible. If that was not enough, he also read every article and critiqued every presentation, lending both his expertise in data architecture and his wordsmithing skills. I am so lucky to have him in my life.

The Joyful Struggle

In memory of

Dr. Frank McKinney

and

Dr. John Carlis

Abstract

While using data standards can facilitate research by making it easier to share data, manually mapping to data standards creates an obstacle to their adoption. Semi-automated mapping strategies can reduce the manual mapping burden. This research addresses the mapping dilemma by applying well-established and emerging techniques to a real-world use case. First, machine learning approaches were used and evaluated to map Common Data Elements (CDEs) from the National Cancer Institute's (NCI) cancer Data Standards Registry and Repository to the Biomedical Research Integrated Domain Group (BRIDG) model. Second, a graph database that incorporates the CDEs, BRIDG Model, and the NCI Thesaurus was developed and evaluated. A shortest path algorithm was then used to predict mappings from CDEs to classes in the BRIDG model. Finally, analysis was conducted to: determine the strengths and weaknesses of each approach; highlight data quality issues; and determine when either approach or a combination of the approaches provides the optimal results. The results indicate that an artificial neural network-based mapping tool is able to predict CDE to BRIDG class mappings with between 34 - 94% accuracy but is limited by the availability of training data. The results also show that a graph database can be used to map CDEs to BRIDG classes but is limited by the subjective nature of the mapping process. An optimal mapping tool combines machine learning and graph database techniques with the knowledge and experience of a human subject matter expert.

Table of Contents

List of Tables ix

List of Figures x

Acronyms xi

Dissertation Organization 1

Chapter 1: Problem Statement 2

Chapter 2: Research Approach 16

Chapter 3: Semi-Automated Mapping Via an Artificial Neural Network 27

Chapter 4: Semi-Automated Mapping Via a Graph Database 64

Chapter 5: Conclusion 86

References 99

List of Tables

Table 1. Similarity of testing data to training data 43

Table 2. Algorithm parameters 52

Table 3. Accuracy with different training validation 53

Table 4. Example of attribute weights 58

Table 5. Summary of the number of key nodes in the database. 77

Table 6. Example of discrepancies between manually mapped and predicted BRIDG classes. 80

List of Figures

Figure 1. Relationship of BRIDG to Other Standards and Models 18

Figure 2. CDE structure 37

Figure 3. Example of CDE structure...38

Figure 4. Oncology Subset of the BRIDG Model...40

Figure 5. Example of word similarity matrix..45

Figure 6. Neural network structure 47

Figure 7. Training flow chart 49

Figure 8. Best match rates per testing set 55

Figure 9. UML representation of a section of the BRIDG model and its corresponding graph model representation. 71

Figure 10. Data Element Concept section of the ISO 11179 metamodel and its corresponding graph model representation. 72

Figure 11. Representation of the graph-based data model...76

Figure 12. Example path with a distance of 4...79

Acronyms

Acronym	Description
AGNIS	A Growable Network Information System
ANN	Artificial neural network
BRIDG	Biomedical Research Integrated Domain Group
caDSR	cancer Data Standards Registry and Repository
CBER	Center for Biologics Evaluation and Research
CDASH	Clinical Data Acquisition Standards Harmonization
CDE	Common Data Element
CDER	Center for Drug Evaluation and Research
CDISC	Clinical Data Interchange Standards Consortium
CDMH	Common Data Model Harmonization project
CIBMTR	Center for International Blood and Marrow Transplant Research
CSV	Comma Separated Values
DDL	Data Definition Language
DEC	Data Element Concept
EMR	Electronic Medical Record
FDA	Federal Drug Administration
FHIR	Fast Healthcare Interoperability Resources
HCT	Hematopoietic Cell Transplantation
HL7	Health Level 7
HRSA	Health Resources and Services Administration
i2b2	Informatics for Integrating Biology & the Bedside
ICD	International Classification of Disease
LOINC	Logical Observation Identifiers Names and Codes
MedDRA	Medical Dictionary for Regulatory Activities
NCI	National Cancer Institute
NMDP	National Marrow Donor Program
OAANN	Ontology Alignment by Artificial Neural Network
OMOP	Observational Medical Outcomes Partnership
OWL	Web Ontology Language
SCTOD	Stem Cell Therapeutic Outcomes Database
SDTM	Study Data Tabulation Model
SNOMED CT	Systematized Nomenclature of Medicine, Clinical Terms
UML	Unified Modeling Language
UMLS	Unified Medical Language System
VD	Value Domain

Dissertation Organization

This dissertation consists of three manuscripts that have either been published or are in press, an introduction to the research approach, and a conclusion. References for all chapters can be found at the end of the manuscript.

Chapter one is a perspective paper that presents a real-world use case that illustrates the negative impact that the manual mapping of cancer study data standards has on the exchange of clinical information for research purposes. Chapter two continues this introductory material with further exploration of the manual mapping problem along with two potential strategies for developing a semi-automated mapping tool. Chapter three is a research paper that presents collaborative work to develop a semi-automated mapping tool using an artificial neural network algorithm. This work was conducted in collaboration with Shengyu Li from the University of Alabama, and the chapter includes a detailed description of the collaborative efforts. Chapter four is a research paper that presents work to develop a graph database and to use a graph algorithm as a semi-automated mapping approach. Chapter five summarizes the findings of the research efforts and describes next steps.

Chapter 1: Problem Statement

Integration of Hematopoietic Cell Transplantation Outcomes Data: Data Standards Are Not Enough

Robinette Renner, John Carlis, Martin Maiers, J Douglas Rizzo, Colleen O'Neill, Mary Horowitz, Katherine Gee, Dennis Confer

A version of this paper appeared in:

Data Integration in Life Sciences 2015. Lecture Notes in Computer Science, vol 9162.

Authors' contributions

RR wrote the manuscript and led the standards development efforts for the AGNIS project. JC oversaw the writing of the manuscript and edited it in depth. MM was the technical lead on the AGNIS project. DR provided clinical oversight for the AGNIS project. CO, MH, and KG provided administrative oversight of the AGNIS project. DC was the principal investigator for the AGNIS project. All authors read and approved the final version of the manuscript.

Abstract

To complete large-scale clinical research, organizations must share data. Because their database schemas are inherently heterogeneous, they need a standard metadata representation in order to exchange data. The A Growable Network Information System (AGNIS) application facilitates the exchange of hematopoietic cell transplantation outcomes data using data standards. However, adoption rates remain low due to a significant mapping burden. The AGNIS application adoption rates show that developing a data standard is not enough. Tools and resources need to be developed to facilitate utilization of the standard.

Keywords

Data integration, interoperability, data standards, hematopoietic cell transplant outcomes data

Introduction

Data standards are often viewed as the key to interoperability. If everyone speaks the same language, then data can flow freely among all interested parties. Unfortunately, the reality is not that simple. Multiple data standards and heterogeneous database systems preclude such a simplistic view. The reality often involves painstaking, labor-intensive manual mapping of a database system to a particular standard or standards only to find out later that the mappings need to be updated because the standard has changed.

This reality is exemplified by the A Growable Network Information System (AGNIS) application developed by the Center for International Blood and Marrow Transplant Research (CIBMTR). The goal of the AGNIS project is to facilitate the integration of hematopoietic cell transplant (HCT) outcomes data. In this paper we describe the AGNIS project, its current state, the challenges it has encountered, and potential strategies to address the challenges.

Background

The CIBMTR is a research collaboration between the National Marrow Donor Program (NMDP)/Be The Match and the Medical College of Wisconsin. Its mission is to improve the outcomes of HCT and cellular therapies through observational and interventional research. The CIBMTR maintains a registry of outcomes data for more than 390,000 transplant recipients, with data submitted from 350 transplant centers worldwide [3]. This collaboration between the CIBMTR, transplant centers, data managers, clinicians, and researchers provides an invaluable resource for the medical community and the patients they serve.

The CIBMTR also maintains the Stem Cell Therapeutic Outcomes Database (SCTOD). The SCTOD was created as part of the C.W. Bill Young Cell Transplantation Program (Program), established by the Stem Cell Therapeutic and Research Act of 2005. The program collects outcomes data, via several recipient outcomes forms required when either the donor or the recipient is from the United States, to facilitate research that improves patient outcomes and increases the availability of adult volunteer donors and umbilical cord blood units [3]. Summaries of the outcomes data are publicly available via a Health Resources and Services Administration (HRSA) website [4].

To help transplant centers submit this federally required outcomes data, the CIBMTR developed FormsNet, a web-based application allowing real-time data entry and validation. While FormsNet collects high-quality data, for those centers with electronic medical records (EMRs) it requires double-data entry, an expense introducing the possibility of data transcription errors.

Eliminating the double-data entry is crucial to cost-effective sharing of high-quality data. Indeed, Aljurf et al say, "Many centers report data to a myriad of overlapping registries and databases. Integration, interfacing and interoperability are the key ingredients for optimum out-comes and use of these registries" [5]. To improve electronic data integration, the CIBMTR developed the AGNIS (A Growable Network Information System) application, a web service based messaging system enabling the secure transmission of standardized data between disparate database systems [3]. It supports the transmission of all data required by the program.

AGNIS serves strictly as a messaging system. In order for the messages to have meaning, both the transplant center's database and the FormsNet database must speak the same language. A standard language for the representation of the HCT outcomes data is critical to the implementation of AGNIS. The CIBMTR chose the cancer Data Standards Registry and Repository (caDSR), which is maintained by the National Cancer Institute (NCI) Center for Biomedical Informatics and Information Technology (CBIIT) [6], as that standard language because it uses an internationally-recognized metadata framework and provides metadata management tools.

The semantics of a data point in the caDSR are expressed using an internationally recognized framework, ISO / IEC 11179. Its high-level construct is a Common Data Element (CDE), which is comprised of two parts: a Data Element Concept (DEC) and a

Value Domain (VD). The DEC is a contextual representation of the data element -- roughly equivalent to the form question. The VD is a physical representation of the data element which describes how the answer to the question gets stored in the database (data type, maximum length, and list of allowed values if applicable) [7].

The caDSR provides a wealth of web-based metadata development, deployment, and maintenance tools. These tools include the Curation Tool, the CDE Browser, Form Builder, and the Sentinel Tool. The Curation Tool walks the end-user through all steps needed to create well-defined CDEs linked to a common terminology maintained by the NCI Enterprise Vocabulary Services. The CDE Browser allows searching the caDSR CDEs, viewing results via a graphical user interface, and downloading the CDEs in Excel or Extensiblete Markup Language (XML) format. The Form Builder website is CIBMTR's primary means of organizing the CDEs for mapping by the transplant centers. The CDEs are organized in a manner that mimics the CIBMTR's data collection forms. The CDEs are presented in the order in which they appear on a form along with form headers and instructions. Finally, the caDSR's Sentinel Tool supports monitoring changes to either CDEs or Form Builder reports via email alerts. This helps maintain the overall metadata quality by showing changes that may negatively impact content [6, 8].

To supplement caDSR tools, the CIBMTR has developed a custom reporting tool that supports its robust, multi-step review process. It automatically verifies that each CDE complies with established caDSR best practices and does not violate the CIBMTR's metadata business rules. A metadata analyst then generates detailed reports to facilitate a comprehensive CDE review that verifies that each CDE accurately captures the data element semantics and is constructed according to proper ISO / IEC 11179 guidelines. A more general report facilitates the reviewing of each CDE for correct semantics by a clinician. This detailed review process ensures CDE quality.

These CDEs serve as the standard language for data transmission via AGNIS. In order to use AGNIS, each transplant center or vendor must map their data elements to the AGNIS CDEs.

Current State of AGNIS Data Transmission

To date 26 HCT recipient outcomes forms and their associated 6,795 FormsNet database fields have been released in the caDSR. 1,515 CDEs are used to represent the database fields on those forms. These forms represent 71.5% of all completed recipient forms in the FormsNet database.

Four vendors, four domestic transplant centers, and one international registry are using AGNIS to submit data to the CIBMTR. The vendor applications are being used by twenty-six centers. The number of centers utilizing AGNIS represents 13% of all domestic transplant centers. The transplant centers support an average of seven forms; the vendors support an average of 18 forms.

Unfortunately, the utilization of AGNIS has yet to reach its full potential. In 2014 AGNIS was used by US transplant centers to submit 7,700 forms which represents just 4% of all recipient forms and 6% of AGNIS-supported forms collected by the CIBMTR. While the utilization of AGNIS has increased since its initial release in 2009, this valuable resource is still underutilized.

In the remainder of the paper, we explore some of the possible reasons for this underutilization and potential strategies for increasing it.

Challenges in Integrating Data

The mapping of the transplant center's database to the CDEs is by far the largest barrier to the utilization of AGNIS. As Warzel et al say, "Complex metadata requirements, overlapping and competing medical terminology standards, and inconsistent information models presented challenges and obstacles to CDE standardization." [9] Their challenges are similar to those in the AGNIS project, which we group into three broad categories: complexities of the mapping process; changes to clinical practice; and inconsistencies and evolution of the data standard.

Complex Mappings

Mapping from heterogeneous database systems to a data standard is complex. It requires an in-depth knowledge of the database system, the clinical domain, and the business process. The transplant center's EMR may consist of separate databases for information about laboratory results, accounting information, HCT -related information, and others. As a result, the human mapper may be required to search across several systems to obtain an accurate mapping. In addition, the mapping to one CDE may require applying complex business rules and calculations to several database fields. Involving clinicians and business process subject matter experts helps ensure that these complex mappings are semantically accurate.

Changes to the Clinical Domain

The nature of the clinical domain necessitates change. As clinical practice changes, new CDEs are created and old CDEs are removed from the forms to maintain data that is timely and useful. Therefore, the CIBMTR periodically reviews and revises forms to ensure that they are consistent with current medical practice. The latest round of form revisions was released in the FormsNet application in 2013. During this revision cycle, 26 recipient forms were revised. Of these, seven were supported by AGNIS. In the AGNIS supported forms alone, 801 questions were added and 580 questions were deleted. These essential changes burden the transplant centers with updating their mappings.

Inconsistencies in and Evolution of the Standard

Inconsistencies commonly exist within a data standard [10-15]. Unfortunately, the CIBMTR has not been immune to this universal challenge. For example, the CIBMTR employs a form-based data management approach. Historically, each form question was defined independently from other forms. Therefore, semantically identical data points were defined multiple times, resulting in inconsistencies across forms. The CIBMTR has addressed these inconsistencies by linking the questions to a robust data dictionary. In November, 2013, the CIBMTR released forms revised using the new data dictionary structure. They had 62% data dictionary instance reuse versus a previous 4%. The new forms' questions were more consistent, facilitating the transplant center's mapping effort.

In addition, the best practices for CDE development have evolved. In AGNIS version 1.0, the XML message structure allowed a particular CDE to be used only once per form. This created a conflict between clinical practice and functional capability. For example, on a form one commonly collects data for the same data element but at different times. Due to XML messaging structure limits, two or more CDEs would be created for this data element. In AGNIS version 2.0, the XML messaging structure did allow for the repetition of a CDE within a form. Thus more generic CDEs facilitating CDE reuse are possible. In the long-term, this change will ease the transplant center's mapping effort. In the short-term, the change complicates overall mapping efforts because the transplant centers need to update their mappings to reflect the usage of the new CDEs. To minimize the impact of these changes, they are incorporated only when the CIBMTR releases new revisions of the forms and comprehensive change notes that map the old CDEs to the new, semantically equivalent CDEs are included.

Strategies to Resolve Data Integration Problems

While the adoption rate of AGNIS among transplant centers is low now, this is a very exciting time for the AGNIS project. The CIBMTR has created a solid foundation for the integration of HCT outcomes data, and there are several innovative strategies that can facilitate the increased utilization of AGNIS. Since each transplant center is different, there is not a one-size-fits-all strategy as transplant center size, patient volume, and available IT resources will vary. Some strategies to facilitate AGNIS adoption are: the BRIDG project; mapping aids; annotation with other standards; and marketing AGNIS and its potential return on investment.

BRIDG

The first strategy to consider is the Biomedical Research Integrated Domain Group (BRIDG) Model. The BRIDG Model represents a collaboration between the NCI, the Clinical Data Interchange Standards Consortium (CDISC), Health Level Seven International (HL7), and the Food and Drug Administration (FDA). Its goal is "to produce a shared view of the dynamic and static semantics for the domain of protocol-driven research and its associated regulatory artifacts." [16] The model is an implementation agnostic representation of the semantics of clinical research that does not provide a physical database model. The CIBMTR, in collaboration with a team of subject matter experts from BRIDG and MD Anderson Cancer Center, along with Computer Science graduate students from the University of Minnesota and NMDP summer interns, mapped the HCT CDEs to the BRIDG Model. This mapping served as the foundation for the development of a BRIDG-compliant physical database for the HCT domain. The first version of the physical database is in the initial stages of review by the broader HCT community.

The next step of the project is to develop an integration engine that will allow for bidirectional transmission of data between the BRIDG-compliant physical database and the FormsNet database. The integration engine will contain all of the mappings and business rules needed to transmit data between the BRIDG-compliant database and the FormsNet database. For those transplant centers that either do not have a database system or are looking to replace their existing one, the combination of the integration engine and the physical database can facilitate the development of electronic data submission to the CIBMTR by significantly reducing their mapping efforts. In addition, they can extend the database model to capture information specific to their needs.

For those transplant centers who already have an integrated database system, the BRIDG-compliant physical database will not significantly reduce their mapping efforts. To utilize the integration engine, they will still need to map their database to the BRIDG-compliant database. While there may be fewer attributes to which to map, the mapping effort will not be eliminated entirely.

Mapping Aids

Since the mapping of database fields to either the CDEs or the BRIDG-compliant physical database will never be eliminated entirely, innovative solutions are needed to reduce the mapping burden. Interesting research is being conducted into ways to reduce this burden. For example, Lin et al have developed a process that combines CDEs, ontologies, and natural language processing to develop a query tool that matches form question texts to a list of potential CDEs. While the tool that they have developed is still a prototype, initial tests yielded a 90% accuracy rate along with favorable feedback from the testers. [17]

The work of Lin et al is not the only work being done is this field. As another example, the MAPONTO tool uses the OWL ontology language and SQL DDL declarations to map an existing database to an ontology [18]. An evaluation of this tool found that it was "able to infer the semantics of many relational tables occurring in practice in terms of an independently developed ontology" which resulted in "significant saving in terms of human labors." [19]

While this is just a small sampling of the work that is being done in the field of ontology mapping, it indicates that there is significant work upon which a mapping aid for AGNIS can be built.

Annotation of Other Standards

There are a wealth of standards including controlled vocabularies used by electronic medical records such as LOINC, SNOMED CT, and ICD which liberate one from the curse of free text. Ideally, there would be one standard used for the capture of all clinical information. Until that time, annotating the CDEs used by AGNIS with the appropriate code from the most commonly used standards may help facilitate the mapping of a transplant center's database to the CDEs. The annotation of multiple standards would serve as a type of Rosetta Stone. The transplant center may not understand all of the standards, but knowing part may help them understand the whole. The effort needed to annotate the CDEs with other standards would be significant. Prior to beginning the annotation project, it would be critical to survey the AGNIS end-users so that work could begin on the highest-impact standards first.

Communicating and Marketing the Standard

To date there has not been a strong effort to market the AGNIS application. The primary means of marketing it has been through the CIBMTR websites, AGNIS user groups, conferences, 27 visits to transplant centers, and word of mouth. Despite these efforts, some transplant centers are not aware of either the AGNIS application or the data elements defined in the caDSR.

In addition, clearly communicating the return on investment that can potentially be realized with an AGNIS implementation can help facilitate its adoption. The CIBMTR has produced documentation about the estimated costs associated with the mapping process. Unfortunately, the documentation does not account for the long-term cost savings that the mapping effort could provide [18]. Surveying the current users of AGNIS to obtain an estimated cost savings that their project has realized would provide potential users with a clearer understanding of the benefits of AGNIS and help them advocate for an AGNIS implementation project.

Conclusion

The AGNIS project is an excellent case study of the universal challenges in data standards adoption. The complexities of the mapping process, the changing nature of the clinical domain, and inconsistencies within the standard itself hinders the widespread adoption of data standards. While progress has been made since the 2008 update on the NIH Roadmap for Reengineering Clinical Research [20], of which AGNIS is a part, additional work remains.

The move from an underlying model that is forms-based to the BRIDG model with its focus on a structure that is common to all protocol driven research has the potential to lower the barrier to adoption. Transplantation is an area within biomedical research where data sharing has been essential. It stands to reason that the value of a shared domain model and semantics for protocol driven research will be realized in due time in other fields, whether the motivation is financial, regulatory or scientific. Clear, quantitative communication of the benefits of standards adoption along with tools and resources to facilitate its adoption will go a long way towards bridging the gaps in data standard utilization. After all, a standard is worthless unless it is being used.

Acknowledgements

Support for this project was partially provided by contract ADB. No. N01-HC-45215 / HHSN268200425215C from the National Heart, Lung, and Blood Institute and the National Marrow Donor Program.

Thanks are extended to all who reviewed the document and provided their thoughtful feedback. In addition, Kirt Schaper, Bridget Wakaruk, and Tony Wirth helped calculate the metrics gathered here.

Chapter 2: Research Approach

Robinette Renner

Abstract

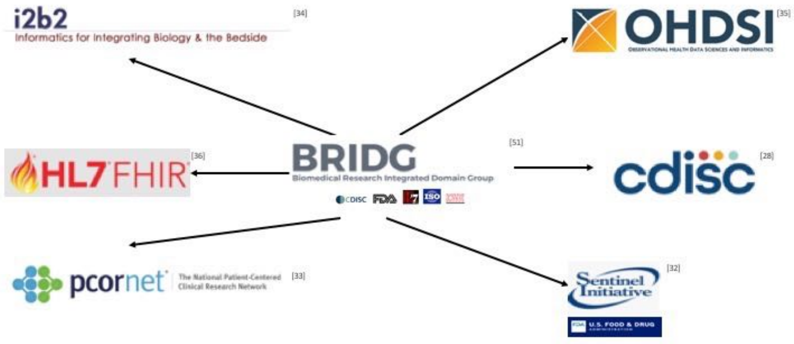
While using data standards can facilitate research by making it easier to share data, manually mapping to data standards creates an obstacle to their adoption. Semi-automated mapping strategies can reduce the manual mapping burden. The research described in this manuscript uses the specific use case of CDE to BRIDG class mappings to address the mapping issue. In particular, it leverages existing ontology alignment work using machine learning approaches. It also explores the emerging work using graph databases to facilitate terminology mapping and maintenance.

Manual Mapping Problem

As mentioned in Chapter 1, the problems inherent in manual mapping have resulted in the relatively low adoption rates of the AGNIS application. As a result, fewer transplantation centers are adopting the electronic data exchange tool that could reduce the burden of reporting to the CIBMTR. To ease the mapping burden, CIBMTR mapped their Common Data Elements (CDEs) to the Biomedical Research Integrated Domain Group (BRIDG) model. This mapping effort has proven beneficial. First, Thomas Klumpp at Thomas Jefferson University used the resulting physical model as the foundation for a research system that facilitates clinical decision making and quality assurance [21]. Second, not only has BRIDG been mapped to a variety of CDISC standards, such as CDASH [22], it also plays a pivotal role in the FDA's Common Data Model Harmonization project [23]. This project mapped the BRIDG model to FHIR [24] and the following data models: Sentinel [25], Patient-Centered Outcomes Research Network (PCORNET) Common Data Model [26], Informatics for Integrating Biology & the Bedside (i2b2) [27], and Observational Medical Outcomes Partnership (OMOP) [28]. As Figure 1 illustrates, BRIDG can now be used to facilitate the mapping to a variety of other standards and models enabling universal query definitions that can be automatically and faithfully translated to the queries based on different data models.

Figure 1. Relationship of BRIDG to Other Standards and Models

Mapping to the BRIDG model can make it easier to map to other clinical data models.


i2b2
Informatics for Integrating Biology & the Bedside
[34]
OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS
[35]
HL7 FHIR
[36]
BRIDG
Biomedical Research Integrated Domain Group
[51]
cdisc
[28]
pcornet
The National Patient-Centered Clinical Research Network
[33]
Sentinel Initiative
U.S. FOOD & DRUG ADMINISTRATION
[32]

Unfortunately, mapping CDEs to the BRIDG model is a laborious process. The CIBMTR's initial effort to map approximately 1,300 CDEs to the BRIDG model took a team of six individuals approximately one year. The team consisted of three technical experts from the CIBMTR, a transplantation physician, an electronic data exchange expert from a transplantation center, and a data architect from BRIDG. The CIBMTR has an additional 2,000 CDEs to map. In addition, clinical practice is continually changing. As a result, new mappings will be needed over time, and the existing mappings will need to be updated as the CDEs and BRIDG model change. Simply mapping CDEs to the BRIDG model is not sufficient to reduce the manual mapping burden.

The CIBMTR's experience with manual mapping is not unique. Manual mapping is a labor-intensive process that can result in complex, context-specific mapping with limit reuse potential [29]. In addition, there is a risk that semantic information can be lost during the mapping process [29, 30]. To further complicate matters, these mappings need to be updated to reflect changes in the underlying terminologies [29]. The temptation is to develop a fully automated mapping tool. Unfortunately, mapping is a subjective process, and fully automated tools still require review by a human subject matter expert [30-32]. A semi-automated approach that combines multiple computational techniques and human subject matter expert review may help address the issues with manual mapping while maintaining mapping quality [33]. Two such computational techniques are machine learning and graph databases.

Machine Learning Techniques

Since both CDEs and ontologies are associations of attributes/terms through relationships, mapping CDEs to the BRIDG model is similar to ontology alignment. Also, the CDEs mapped to specific BRIDG classes tend to have structural similarities. The machine learning techniques used in ontology alignment algorithms detect such similarities. As a result, mapping CDEs to the BRIDG model is a natural extension of the previous work in ontology alignment.

Mao et al. developed the PRIOR+ algorithm [34] to map two ontologies. PRIOR+ consists of a Similarity Generator, Similarity Aggregator, and a Constraint Satisfaction Solver. The Similarity Generator determines similarity based on edit distance, parent-child relationships, and structural similarity of classes. The Similarity Aggregator combines these similarities by assigning weights to each type of similarity. The Constraint Satisfaction Solver uses an interactive activation and competition neural network to adjust for constraints in the ontologies.

Noy and Musen [35] used a rules-based approach to develop the semi-automatic matching algorithm PROMPT. This approach first identifies potential entity matches, and then the user acknowledges or declines the merged entity pairs manually. Anchor-PROMPT [36] is an upgraded version of PROMPT. With Anchor-PROMPT, the user inputs pairs of related terms from two different terminologies that serve as “anchors.” The algorithm then traverses the paths of the relationships in each ontology to find similar concepts.

Doan et al. [31] used a learning-based approach to develop GLUE. GLUE uses a Metalearner that combines the results of a Content Learner and Name Learner to match ontologies. The Content Learner determines similarity based on long textual content of instances such as descriptions. The Name Learner determines similarity based on the name of an input instance. It works best with specific names. Both learners use a Native Bayes learning technique.

Chortaras, Stamou, and Stafylopatis’s [37] learning-based approach views an ontology as a graph which represents ontology concepts as nodes and the relationships between concepts as edges. This results in concepts with relationships that resemble trees. They then used a recursive neural network classifier to determine the similarities between such trees.

Rubiolo et al. used an artificial neural network-based learning approach to align ontologies [38]. Their uses both schema information and training instances to map the ontologies. An interesting aspect of their approach is that the use the WordNet database to assign an eight-digit numeric code to the terms using in the training instances. WordNet is a lexical database that groups related words and includes relationships such as parent-child relationships [39]. The use of WordNet allowed Rubiolo et al. to account for synonyms. They also leveraged the reuse of codes between the English WordNet and Spanish WordNet to minimize the impact that the use of different languages might have [38].

The Ontology Alignment by Artificial Neural Network (OAANN) algorithm developed by Huang et al. [40, 41] uses both rules-based and learning-based approaches to align concepts from two real-world biological ontologies successfully. It develops mapping rules based on the structure of the ontologies. Based on this structure, a 3-dimension vector is created, which represents the concept's name, properties, and relationships. An artificial neural network then uses concept instances to learn and assign weights to the three aspects of a concept.

The use of machine learning to align ontologies is a well-established and stable field [42]. A search of Scopus [43], a database of peer-reviewed literature, for articles containing the keywords "ontology alignment" or "ontology mapping" shows that the number of articles peaked in 2008 with 208 articles and has decreased steadily since then. For 2018, the Scopus database had just 123 articles. Shvaiko et al. [42] acknowledge that progress in this field has slowed down. Since this is a relatively stable field, I will apply the ontology mapping techniques to a new use case instead of improving existing techniques.

Graph Database Techniques

The CDEs and BRIDG model share a common underlying foundation. The CDEs within the caDSR are based on the ISO 11179 metamodel for metadata repositories [44]. This metamodel breaks a data point into reusable structures consisting of conceptual and representational components. Each of the components is created using concepts from the

NCI Thesaurus. BRIDG classes and attributes have definitions based on NCI Thesaurus concepts. As a result, the NCI Thesaurus provides a web of relationships that can connect CDEs to BRIDG classes. Graph databases represent highly interconnected data as a collection of nodes (entities) and edges (relationships). They are particularly effective in using relationships to predict associations between two entities [45] and are a potential technique for mapping CDEs to the BRIDG model. Graph databases have been used data integration [46], semantic annotation [47], clinical concept representation [48-51], CDE mapping [52], and BRIDG model representation [53].

Graph databases have been leveraged for data integration, quality assurance, and semantic annotation. Alaqhatani and Heckel used a graph database for data integration and quality assurance tasks [46]. In their methodology, they first mapped two relational databases and represented the relationships using a triple graph grammar. The UML representations of each relational database and the triple graph grammar statements were used to create a graph database that contained both the relational databases and the relationships between them. The instance data from each relational database was imported into the graph database, and the data integration was then performed at the model level.

In 2014, Johnson et al. [47] developed a graph database that stored metadata about tumor data models. They then captured the semantics of each model by annotating them with concepts from the NCI Thesaurus. This allowed them to find related models via the rich semantics of the NCI Thesaurus, such as synonyms. A limitation of this database is that it can currently only find models that are directly connected via the NCI Thesaurus concepts. They plan to develop more robust queries that will use the distance between graph nodes to find models that are not directly related.

Clinical concept representation using graph databases is a rich research area. The Unified Medical Language System (UMLS) has been a primary focus. The UMLS represents clinical and biomedical research-related concepts from numerous vocabulary sources [54]. The 2019 AB UMLS integrated more than four million concepts from over 200 sources [55]. Similar concepts from multiple sources are represented as a single concept within the UMLS. This robust source of interrelated concepts lends itself well to graph database research.

In 2016 van Mulligan et al. [48] implemented the UMLS as a Neo4j graph database. They successfully used the database to find PubMed articles that referenced migraine-related biomarkers. They also created a migraine subgraph based on a reference set defined by a subject matter expert [49]. The subgraph approach improved the effectiveness of the search. A limitation of their method was the amount of noise in the database. Given the richness of the relationships, at some point, the database was able to find a relationship between each pair of concepts [48]. Also, missing relationships in the graph database reduced its effectiveness [49].

Xiang et al. [50] used k-neighborhood decentralization to optimize UMLS queries and minimize the impact of the large size of the UMLS. This approach allowed for searching of the entire UMLS, determining the shortest path between concepts, and other analysis.

Campbell et al. [51] used a graph database to integrate partial patient records with a variety of clinical vocabularies, primarily SNOMED CT. They used the graph database to find patients with certain clinical findings. They compared the query results to the results of queries run against the i2b2 database, which was a known source of clinical data represented using SNOMED CT concepts. The query results were identical, and the runtimes were similar. However, due to how SNOMED CT is represented in i2b2, they were only able to test parent-child concept relationships. They were not able to run i2b2 queries involving negation relationships. The work also highlighted data quality issues within SNOMED CT.

Graph databases have also been used to map Common Data Elements (CDEs). In 2018, Ulrich et al. [52] used a graph database to represent 666 cancer-related CDEs from different metadata repositories. While Ulrich has previously used the ISO 11179 metamodel for representing CDEs [56], the graph database in the current research represented CDEs simply as a single node with relationships to data types, options, and sources. He then compared the CDE names using a pattern recognition-based similarity algorithm that used both the five-gram algorithm and the metric Longest Common Subsequence algorithm. If present, CDE options were also compared. If two CDEs were found to be similar, a new relationship was created between the CDEs, indicating that they are similar. While this work was able to map related CDEs, it has several limitations. First, the CDEs are not represented using the ISO 11179 metamodel, which would provide additional relationships between CDEs. Second, it uses pattern recognition instead of semantics. Annotating the CDEs with concepts from an established vocabulary such as the NCI Thesaurus would provide additional relationships that could be leveraged to find related CDEs.

From a BRIDG model perspective, Jane Pollack has converted the BRIDG model into a Neo4j graph database and developed CDE to BRIDG mapping tooling [53]. Her tooling not only updates CDE to BRIDG mappings when the BRIDG model is updated, but it also performs quality assurance testing of existing mappings. However, her approach works strictly with existing mappings. It is not able to predict novel mappings. Combining the previous graph database work involving the UMLS and the BRIDG model to predict CDE to BRIDG mappings is a logical next step.

Research Focus

As the experience of the CIBMTR has shown, there is a need for semi-automated tools to map CDEs to the BRIDG model. While semi-automated mapping tools exist for some use cases, such as mapping two ontologies or mapping CDEs from different repositories, no known tools exist for the specific use case of mapping ISO 11179-compliant CDEs to a domain model. A combination of machine learning and graph database techniques may result in a comprehensive approach to mapping CDEs to the BRIDG model and performing quality assurance testing.

Using a real-world use case, I developed tools using both well-established and emerging techniques to address the long-standing mapping dilemma. I accomplished this by leveraging existing mapping approaches using artificial neural networks and graph databases. The objectives of the project are: to develop mapping tools using each of the two approaches; to determine the strengths and weaknesses of each approach; and to determine when either approach or a combination of the approaches provides the optimal results. The following research aims will create tools to increase the interoperability of CDEs:

Research Aim 1: Develop and evaluate a machine learning algorithm to predict mappings between different standards

- Adapt an existing artificial neural network-based ontology alignment algorithm to map two standards that use different metamodels
- Test the algorithm against different data sets with existing manual mappings

Research Aim 2: Develop and evaluate a graph database to predict mappings between different standards

- Develop a graph database that incorporates cancer Data Standards Registry and Repository (caDSR) CDEs, the BRIDG model, and the NCI Thesaurus
- Test a graph-based algorithm against different data sets with existing manual mappings

Chapter three covers Research Aim 1 in detail. It discusses the work to adapt the Ontology Alignment by Artificial Neural Network algorithm developed by Huang et al. and the results of the experiments. Chapter four covers Research Aim 2 in detail. It describes the work to develop a graph database that incorporates CDEs from the caDSR, the BRIDG model, and the NCI Thesaurus. It also discusses the results of using a shortest path algorithm to predict CDE to BRIDG class mappings. Chapter five compares the two approaches and describes how these approaches can best be used going forward.

Chapter 3: Semi-Automated Mapping Via an Artificial Neural Network

Using an Artificial Neural Network to Map Cancer Common Data Elements to the Biomedical Research Integrated Domain Group Model in a Semi-automated Manner

Robinette Renner, Shengyu Li*, Yulong Huang, Ada Chaeli van der Zijp-Tan, Shaobo Tan, Dongqi Li, Mohan Vamsi Kasukurthi, Ryan Benton, Glen M. Borchert, Jingshan Huang, Guoqian Jiang

*Co-first Author

A version of this paper appeared in:
BMC Medical Informatics and Decision Making 2019 19 (Suppl 7):276

Authors' contributions

GJ and JH designed the overall study, directed the research project, and are responsible for the integrity of this work. RR prepared all use cases. SL implemented the software. RR and SL performed the experiments, interpreted and analyzed the results, and drafted the manuscript. RB, GMB, MVK, YH, ST, DL, and AV participated in the discussion of software design, helped with the software implementation, and assisted in result analysis as well. GMB, RB, YH, ST, DL, and AV (led by GMB and RB) reviewed and revised the manuscript. All authors read and approved the final version of the manuscript.

Description of Research Collaboration

Research Phase	**Task**	**Performed By**
Design	Gathering requirements	Robinette Renner
	Modifying Ontology Alignment by Artificial Neural Network algorithm	Shengyu Li
	Determining input variables	Robinette Renner
	Developing training data sets	Robinette Renner
Algorithm Implementation	Developing algorithm architecture	Shengyu Li
	Coding the application	Shengyu Li
	Verifying the application	Shengyu Li
Algorithm Testing	Developing testing data sets	Robinette Renner
	Developing testing scenarios	Robinette Renner
	Testing algorithm	Robinette Renner
	Analyzing testing results	Robinette Renner

Abstract

Background

The medical community uses a variety of data standards for both clinical and research reporting needs. ISO 11179 Common Data Elements (CDEs) represent one such standard that provides robust data point definitions. Another standard is the Biomedical Research Integrated Domain Group (BRIDG) model, which is a domain analysis model that provides a contextual framework for biomedical and clinical research data. Mapping the CDEs to the BRIDG model is important; in particular, it can facilitate mapping the CDEs to other standards. Unfortunately, manual mapping, which is the current method for creating the CDE mappings, is error-prone and time-consuming; this creates a significant barrier for researchers who utilize CDEs.

Methods

In this work, we developed a semi-automated algorithm to map CDEs to likely BRIDG classes. First, we extended and improved our previously developed artificial neural network (ANN) alignment algorithm. We then used a collection of 1,284 CDEs with robust mappings to BRIDG classes as the gold standard to train and obtain the appropriate weights of six attributes in CDEs. Afterward, we calculated the similarity between a CDE and each BRIDG class. Finally, the algorithm produces a list of candidate BRIDG classes to which the CDE of interest may belong.

Results

For CDEs semantically similar to those used in training, a match rate of over 90% was achieved. For those partially similar, a match rate of 80% was obtained and for those with drastically different semantics, a match rate of up to 70% was achieved.

Discussion

Our semi-automated mapping process reduces the burden of domain experts. The weights are all significant in six attributes. Experimental results indicate that the availability of training data is more important than the semantic similarity of the testing data to the training data. We address the overfitting problem by selecting CDEs randomly and adjusting the ratio of training and verification samples.

Conclusions

Experimental results on real-world use cases have proven the effectiveness and efficiency of our proposed methodology in mapping CDEs with BRIDG classes, both those CDEs seen before as well as new, unseen CDEs. In addition, it reduces the mapping burden and improves the mapping quality.

Keywords

Common data element, artificial neural network, schema mapping, Biomedical Research Integrated Domain Group (BRIDG) model.

Background

As Andrew Tanenbaum said: “The nice thing about standards is that there are so many to choose from” [57]. While Tanenbaum was talking about digital media standards, the statement applies to clinical data standards as well. Unfortunately, this truism has a corollary: the worst thing about data standards is that they require significant mapping efforts. More often than not, these are resource-intensive manual mappings.

In 2008 Rachel Richesson enumerated the problems with manual mapping, including the significant amount of time needed to develop and maintain the mappings; frequent lack of unambiguous, one-to-one mappings; and the context-specific nature of the mappings that limit their reuse [29]. Unfortunately, ten years later these problems have yet to be fully resolved.

The experience of the Center for International Blood and Marrow Transplant Research (CIBMTR) regarding the implementation of electronic data capture and the adoption of data standards is an excellent case study of the problems with manual mapping. The CIBMTR is a research collaboration between the National Marrow Donor Program (NMDP)/Be The Match and the Medical College of Wisconsin. For more than 45 years, the CIBMTR has been collecting outcomes data and facilitating research in hematopoietic cell transplantation [58]. While the transplantation centers submit most of their data using a Web-based interface, the CIBMTR’s A Growable Network Information System (AGNIS) application [58] allows submissions directly from a transplantation center’s database to the CIBMTR. Common Data Elements (CDEs) from the National Cancer Institute’s (NCI) cancer Data Standards Registry and Repository (caDSR) form the foundation of data transmission via AGNIS. To either send or receive data using AGNIS, a transplantation center must map their internal data points to the CDEs [59].

While some transplantation centers and third-party vendors successfully use AGNIS, difficulties in manually creating and maintaining the mappings have limited its adoption [59]. To reduce the mapping burden, the CIBMTR developed a physical data model based on the Biomedical Research Integrated Domain Group (BRIDG) model [59] and mapped some of their CDEs to this model. In collaboration with multiple organizations such as the NCI and the Clinical Data Interchange Standards Consortium (CDISC) [22], the BRIDG model was developed to “produce a shared view of the dynamic and static semantics of a common domain-of-interest, specifically the domain of protocol-driven research and its associated regulatory artifacts” [60].

In addition to helping create a physical database model, the BRIDG model can facilitate mapping to other standards. For example, the BRIDG model has been harmonized and mapped to CDISC’s Clinical Data Acquisition Standards Harmonization (CDASH) and Study Data Tabulation Model (SDTM) [22]. As a result, the CIBMTR’s mappings to BRIDG can be used to facilitate mapping to CDISC. In 2016 mapping to CDISC became more critical when the Federal Drug Administration (FDA) mandated that most submissions to the FDA Center for Biologics Evaluation and Research (CBER) and Center for Drug Evaluation and Research (CDER) must comply with CDISC standards [61].

Whereas it can make mapping to other standards easier, mapping to BRIDG itself is difficult. Version 5.1 of the BRIDG model contains 320 classes [60]. While the model is subdivided into nine smaller subdomains [22], its size makes the mapping a significant challenge. The CIBMTR has mapped 1,284 CDEs to the BRIDG model. This year-long mapping effort involved six subject matter experts, including a clinician and a BRIDG representative.

The CIBMTR has more than 2,000 CDEs left to map. Based on the previous project, the mapping of the remaining CDEs will take approximately two years, which is unacceptable. One solution to reduce this mapping burden is to develop a semi-automated mapping tool that would recommend candidate matches from which a subject matter expert could select the best mapping.

Semi-automated mapping solutions are an area of active research, especially with ontology alignment. There are structural and conceptual similarities between CDEs and ontologies as both are associations of attributes/terms through relationships. This conceptual view shows the similarity of ontology alignment with the CDE to BRIDG mapping. There are two main methods for mapping ontologies: rule-based and learning-based. For the rule-based approach, a representative method developed by Noy and Musen [35] showcased a semi-automatic approach, PROMPT, based on the SMART algorithm of the same authors [62]. This approach first identifies label matching, then the user acknowledges or declines the merged entity pairs manually. Anchor-PROMPT [36] is an upgraded version of PROMPT that calculates the similarity based on the ontology structure.

For the learning-based approach, GLUE [31] uses machine learning techniques to do the ontology mapping. It uses two base learners, the Content Learner and Name Learner, as inputs to a meta-learner, which forms the final prediction; the meta-learner combines the weighted sum of the outputs of the base learners. The advantage is that it is a suited approach for textual instance descriptions. The disadvantage is that this approach is not applicable to relations or instances.

Other ontology matching algorithms exist. COMA [63] is a platform that combines the result of single matches. A statistical schema matching delivered by He and Chang [64] matches schemas by obtaining the generative model. Rubiolo et al. [38] present an approach based on an artificial neural network (ANN) model within a Knowledge Source Discovery agent. It helps the user to avoid unrelated search results and the possibility of making the wrong decision. Chortaras et al. [37] used a recursive neural network to learn the similarities between ontology concepts and then to combine two ontologies. This method has achieved some promising initial results. PRIOR+ [34] is an ontology mapping approach founded on propagation theory, information retrieval techniques, and artificial intelligence. It provides an estimate of f-measure for ontology constraint.

ANN methods are adopted in many of the algorithms mentioned above because ANNs have high accuracy, strong parallel processing ability, strong distributed storage and learning ability, strong robustness and fault tolerance to noise, full approximation of complex nonlinear relationships, and associative memory functions. Of particular importance, neural networks can extract features not available in many other machine learning methods.

Previously we developed an algorithm, Ontology Alignment by Artificial Neural Network (OAANN) [40, 41] to map two ontologies. It combines the benefits of rule-based and learning-based approaches to learn and adjust weights of concept name, concept properties, and concept relationships between a pair of concepts from two different ontologies. The algorithm applied gradient descent and the targeted design of each attribute similarity. We used this algorithm to successfully align concepts from two real-world biological ontologies with 0.9 Precision and 0.85 Recall, significantly reducing the time that domain experts spend on mapping ontologies.

This manuscript is an extension from our previously published work [19]. Compared with the original paper, most sections were significantly extended in this version. Major extensions (not including minor modifications) are summarized as follows. (1) Background: expanded introduction along with eight new references. (2) Materials and Methods: expanded description of the metamodels, especially the ISO 11179 CDE structure along with one new mapping example, four new figures, and some new references; added information about three new testing data sets along with one new table; and one new example for similarity determination. (3) Results and Discussion: extensive evaluation of our algorithm on real-world test cases; major modifications to the overall testing approach along with three new testing sets; one new flowchart describing the training process; one new sub-section about testing with preexisting mappings; one new chart summarizing the testing results; expanded description along with one new table in the Weights sub-section; and expanded description in the Overfitting sub-section.

Our most significant contribution in this work is to facilitate mapping two different meta-models: ISO 11179 CDEs and the BRIDG domain model. Importantly, as far as we know, no other work semi-automatically maps ISO 11179 CDEs to a domain model.

The rest of this paper is structured as follows. In the Materials section, we discuss key aspects of the ISO 11179 standard, the BRIDG model that form the foundation of the algorithm, and the data sets used for verification and testing. In the Methods section, we describe the details of our alignment algorithm. In the Results section, we report the testing of the algorithm and its application to a set of unaligned CDEs. In the Discussion section, we provide a detailed analysis of our experimental results. Finally, we conclude by discussing critical future work.

Materials

Metamodels and mappings

Our algorithm maps two different metamodels: ISO 11179-based CDEs and the BRIDG domain model. The ISO 11179 standard serves as the metamodel for the National Cancer Institute's cancer Data Standards Registry and Repository (caDSR) [44]. This metamodel breaks a data point into reusable structures consisting of conceptual and representational components. The conceptual component refers to a Data Element Concept (DEC) [65]. The DEC consists of two parts: object class and property, describing concepts and concept characteristics, respectively.

The representational component of a CDE refers to a Value Domain (VD), which describes a set of allowed values in CDEs. The set of allowed values could be constrained to a specific set of permissible values or constrained by a list of requirements such as data type and maximum length [65]. Each VD has a representation term describing the information that VD is capturing. See Figure 2 for an illustration of the structure. An example of a CDE is one used to capture a patient's specific type of Acute Myeloid Leukemia. See Figure 3 for the structure of the CDE Acute Myeloid Leukemia Classification Type [66].

Figure 2. CDE structure

CDE mainly consists of two parts: Data Element Concept and Value Domain.

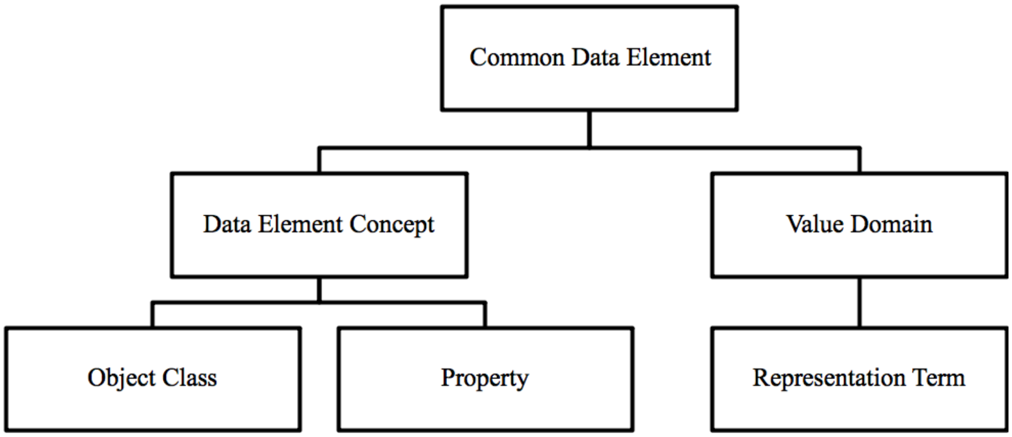
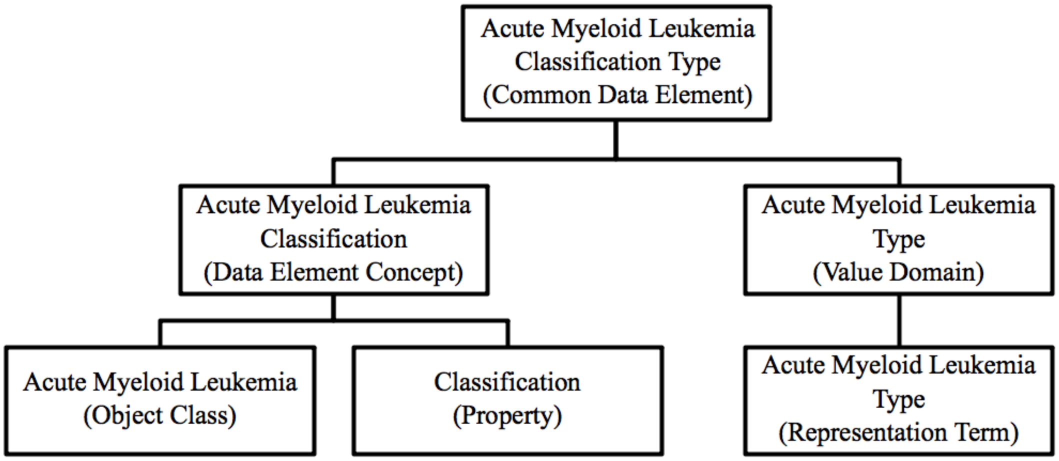


Figure 3. Example of CDE structure

The CDE structure for a data element capturing a patient's specific type of Acute Myeloid Leukemia.


Acute Myeloid Leukemia Classification Type (Common Data Element)
Acute Myeloid Leukemia Classification (Data Element Concept)
Acute Myeloid Leukemia Type (Value Domain)
Acute Myeloid Leukemia (Object Class)
Classification (Property)
Acute Myeloid Leukemia Type (Representation Term)

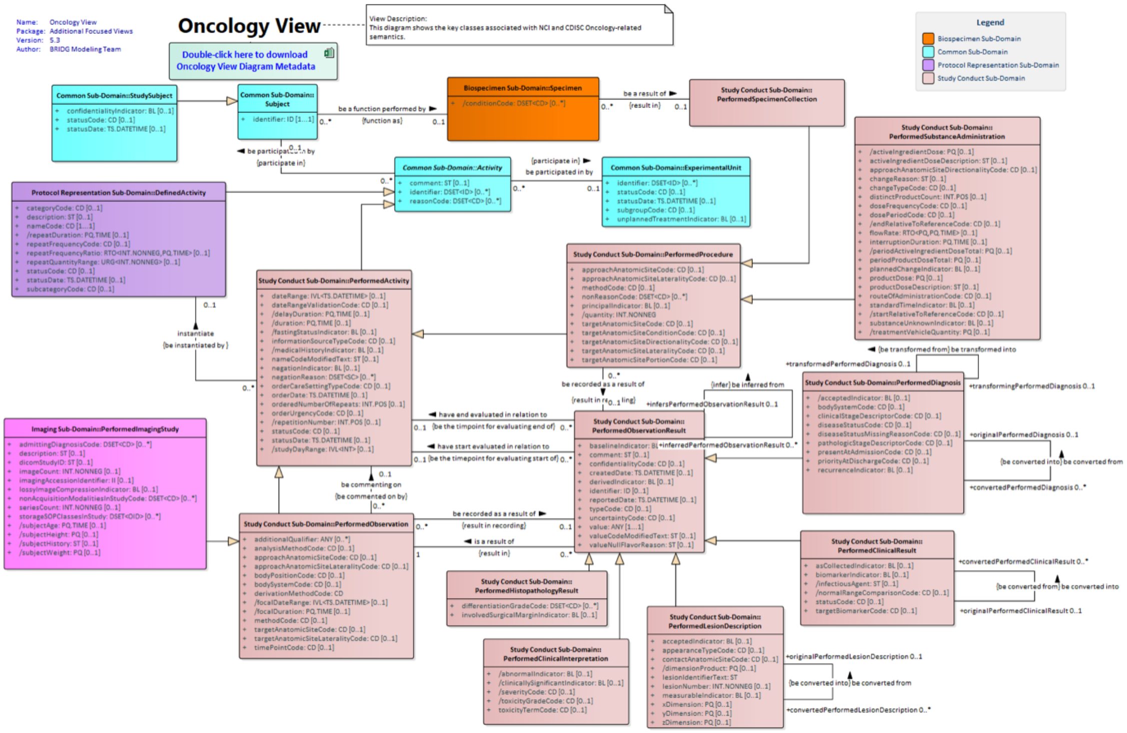
The algorithm uses the six CDE attributes that capture the core semantics: CDE Long Name, Object Class, Property, Value Domain Long Name, Representation Term, and Question Text. Since the CDE Long Name should be created by concatenating the Object Class, Property, and Representation Term [67], it is one of the key attributes. For example, the CDE represented in Figure 3 has an Object Class of "Acute Myeloid Leukemia"; a Property of "Classification"; and a Representation Term of "Acute Myeloid Leukemia Type." Combined they produce a CDE Long Name of "Acute Myeloid Leukemia Classification Type." The question text presents the semantics of the data element in everyday language instead of the formal syntax of the CDE. Therefore, it was included as an input variable. For example, the CDE in Figure 3 has a Question Text of "What was the classification of the acute myelogenous leukemia?" The Value Domain Long Name was included to ensure the complete representation of the CDE's semantics. CDE attributes such as allowed values, data type, and maximum length were not included in the analysis because they do not contribute to the semantic meaning of the CDE.

Within the caDSR, the constructs within a CDE and its attributes are associated with concepts from the NCI Thesaurus, a controlled vocabulary maintained by the Enterprise Vocabulary Service [44]. These concepts provide the concept name, definition, synonyms, and relationships to other concepts. Our algorithm currently leverages the concept name only.

In contrast to the ISO 11179, the BRIDG model represents data points using Unified Modeling Language (UML) classes, attributes, and relationships [68]. In UML, classes represent a classification of an object and attributes represent an object's property [69]. In relation to the ISO 11179 metamodel, they correspond to the object class and property respectively [65]. To facilitate viewing of the BRIDG model, it has been subdivided into topic-specific views. Figure 4 shows the Oncology view of the model.

Figure 4. Oncology Subset of the BRIDG Model[60]

The portion of the BRIDG model representing cancer-related semantics


Name: Oncology View
Package: Additional Focused Views
Version: 5.3
Author: BRIDG Modeling Team
Oncology View
View Description:
This diagram shows the key classes associated with NCI and CDISC Oncology-related semantics.
Double-click here to download Oncology View Diagram Metadata
Legend
🟧 Biospecimen Sub-Domain
🟦 Common Sub-Domain
🟪 Protocol Representation Sub-Domain
▨ Study Conduct Sub-Domain
Common Sub-Domain::StudySubject
+ confidentialityIndicator: BL [0..1]
+ statusCode: CD [0..1]
+ statusDate: TS.DATETIME [0..1]
Common Sub-Domain:: Subject
+ identifier: ID [1...1]
be a function performed by
{function as}
0..*
0..1
Biospecimen Sub-Domain::Specimen
+ /conditionCode: DSET<CD> [0..*]
be a result of
{result in}
0..*
0..1
Study Conduct Sub-Domain:: PerformedSpecimenCollection
be participated in by
0..1
{participate in}
Common Sub-Domain::Activity
+ comment: ST [0..1]
+ identifier: DSET<ID> [0..*]
+ reasonCode: DSET<CD> [0..*]
0..*
{participate in}
be participated in by
0..*
0..1
Common Sub-Domain::ExperimentalUnit
+ identifier: DSET<ID> [0..*]
+ statusCode: CD [0..1]
+ statusDate: TS.DATETIME [0..1]
+ subgroupCode: CD [0..1]
+ unplannedTreatmentIndicator: BL [0..1]
Protocol Representation Sub-Domain::DefinedActivity
+ categoryCode: CD [0..1]
+ description: ST [0..1]
+ nameCode: CD [1...1]
+ /repeatDuration: PQ.TIME [0..1]
+ repeatFrequencyCode: CD [0..1]
+ repeatFrequencyRatio: RTO<INT.NONNEG,PQ.TIME> [0..1]
+ repeatQuantityRange: URG<INT.NONNEG> [0..1]
+ statusCode: CD [0..1]
+ statusDate: TS.DATETIME [0..1]
+ subcategoryCode: CD [0..1]
0..1
instantiate
{be instantiated by }
0..*
Study Conduct Sub-Domain::PerformedSubstanceAdministration
+ /activeIngredientDose: PQ [0..1]
+ activeIngredientDoseDescription: ST [0..1]
+ approachAnatomicSiteDirectionalityCode: CD [0..1]
+ changeReason: ST [0..1]
+ changeTypeCode: CD [0..1]
+ distinctProductCount: INT.POS [0..1]
+ doseFrequencyCode: CD [0..1]
+ dosePeriodCode: CD [0..1]
+ /endRelativeToReferenceCode: CD [0..1]
+ flowRate: RTO<PQ,PQ.TIME> [0..1]
+ interruptionDuration: PQ.TIME [0..1]
+ /periodActiveIngredientDoseTotal: PQ [0..1]
+ periodProductDoseTotal: PQ [0..1]
+ plannedChangeIndicator: BL [0..1]
+ productDose: PQ [0..1]
+ productDoseDescription: ST [0..1]
+ routeOfAdministrationCode: CD [0..1]
+ standardTimeIndicator: BL [0..1]
+ /startRelativeToReferenceCode: CD [0..1]
+ substanceUnknownIndicator: BL [0..1]
+ /treatmentVehicleQuantity: PQ [0..1]
Study Conduct Sub-Domain::PerformedProcedure
+ approachAnatomicSiteCode: CD [0..1]
+ approachAnatomicSiteLateralityCode: CD [0..1]
+ methodCode: CD [0..1]
+ nonReasonCode: DSET<CD> [0..*]
+ principalIndicator: BL [0..1]
+ /quantity: INT.NONNEG
+ targetAnatomicSiteCode: CD [0..1]
+ targetAnatomicSiteConditionCode: CD [0..1]
+ targetAnatomicSiteDirectionalityCode: CD [0..1]
+ targetAnatomicSiteLateralityCode: CD [0..1]
+ targetAnatomicSitePortionCode: CD [0..1]
Study Conduct Sub-Domain::PerformedActivity
+ dateRange: IVL<TS.DATETIME> [0..1]
+ dateRangeValidationCode: CD [0..1]
+ /delayDuration: PQ.TIME [0..1]
+ /duration: PQ.TIME [0..1]
+ /fastingStatusIndicator: BL [0..1]
+ informationSourceTypeCode: CD [0..1]
+ /medicalHistoryIndicator: BL [0..1]
+ nameCodeModifiedText: ST [0..1]
+ negationIndicator: BL [0..1]
+ negationReason: DSET<SC> [0..*]
+ orderCareSettingTypeCode: CD [0..1]
+ orderDate: TS.DATETIME [0..1]
+ orderedNumberOfRepeats: INT.POS [0..1]
+ orderUrgencyCode: CD [0..1]
+ /repetitionNumber: INT.POS [0..1]
+ statusCode: CD [0..1]
+ statusDate: TS.DATETIME [0..1]
+ /studyDayRange: IVL<INT> [0..1]
{be transformed from} be transformed into
+transformedPerformedDiagnosis 0..1
+transformingPerformedDiagnosis 0..1
0..*
be recorded as a result of
{result in re 0..1 ling}
{infer} be inferred from
+infersPerformedObservationResult 0..1
have end evaluated in relation to
0..1
{be the timpoint for evaluating end of}
0..*
have start evaluated in relation to
0..1
{be the timepoint for evaluating start of}
0..*
Study Conduct Sub-Domain:: PerformedObservationResult
+ baselineIndicator: BL
+inferredPerformedObservationResult 0..*
+ comment: ST [0..1]
+ confidentialityCode: CD [0..1]
+ createdDate: TS.DATETIME [0..1]
+ derivedIndicator: BL [0..1]
+ identifier: ID [0..1]
+ reportedDate: TS.DATETIME [0..1]
+ typeCode: CD [0..1]
+ uncertaintyCode: CD [0..1]
+ value: ANY [1...1]
+ valueCodeModifiedText: ST [0..1]
+ valueNullFlavorReason: ST [0..1]
Study Conduct Sub-Domain::PerformedDiagnosis
+ /acceptedIndicator: BL [0..1]
+ bodySystemCode: CD [0..1]
+ clinicalStageDescriptorCode: CD [0..1]
+ diseaseStatusCode: CD [0..1]
+ diseaseStatusMissingReasonCode: CD [0..1]
+ pathologicStageDescriptorCode: CD [0..1]
+ presentAtAdmissionCode: CD [0..1]
+ priorityAtDischargeCode: CD [0..1]
+ recurrenceIndicator: BL [0..1]
+originalPerformedDiagnosis 0..1
{be converted into} be converted from
+convertedPerformedDiagnosis 0..*
Imaging Sub-Domain::PerformedImagingStudy
+ admittingDiagnosisCode: DSET<CD> [0..*]
+ description: ST [0..1]
+ dicomStudyID: ST [0..1]
+ imageCount: INT.NONNEG [0..1]
+ imagingAccessionIdentifier: II [0..1]
+ lossyImageCompressionIndicator: BL [0..1]
+ nonAcquisitionModalitiesInStudyCode: DSET<CD> [0..*]
+ seriesCount: INT.NONNEG [0..1]
+ storageSOPClassesInStudy: DSET<OID> [0..*]
+ /subjectAge: PQ.TIME [0..1]
+ /subjectHeight: PQ [0..1]
+ /subjectHistory: ST [0..1]
+ /subjectWeight: PQ [0..1]
0..1
be commenting on
{be commented on by}
0..*
Study Conduct Sub-Domain::PerformedObservation
+ additionalQualifier: ANY [0..*]
+ analysisMethodCode: CD [0..1]
+ approachAnatomicSiteCode: CD [0..1]
+ approachAnatomicSiteLateralityCode: CD [0..1]
+ bodyPositionCode: CD [0..1]
+ bodySystemCode: CD [0..1]
+ derivationMethodCode: CD
+ /focalDateRange: IVL<TS.DATETIME> [0..1]
+ /focalDuration: PQ.TIME [0..1]
+ methodCode: CD [0..1]
+ targetAnatomicSiteCode: CD [0..1]
+ targetAnatomicSiteLateralityCode: CD [0..1]
+ timePointCode: CD [0..1]
be recorded as a result of
0..*
{result in recording}
0..1
is a result of
1
{result in}
0..*
Study Conduct Sub-Domain:: PerformedHistopathologyResult
+ differentiationGradeCode: DSET<CD> [0..*]
+ involvedSurgicalMarginIndicator: BL [0..1]
Study Conduct Sub-Domain:: PerformedClinicalResult
+ asCollectedIndicator: BL [0..1]
+ biomarkerIndicator: BL [0..1]
+ /infectiousAgent: ST [0..1]
+ /normalRangeComparisonCode: CD [0..1]
+ statusCode: CD [0..1]
+ targetBiomarkerCode: CD [0..1]
+convertedPerformedClinicalResult 0..*
{be converted from} be converted into
+originalPerformedClinicalResult 0..1
Study Conduct Sub-Domain:: PerformedClinicalInterpretation
+ /abnormalIndicator: BL [0..1]
+ /clinicallySignificantIndicator: BL [0..1]
+ /severityCode: CD [0..1]
+ /toxicityGradeCode: CD [0..1]
+ toxicityTermCode: CD [0..1]
Study Conduct Sub-Domain:: PerformedLesionDescription
+ acceptedIndicator: BL [0..1]
+ appearanceTypeCode: CD [0..1]
+ contactAnatomicSiteCode: CD [0..1]
+ /dimensionProduct: PQ [0..1]
+ lesionIdentifierText: ST
+ lesionNumber: INT.NONNEG [0..1]
+ measurableIndicator: BL [0..1]
+ xDimension: PQ [0..1]
+ yDimension: PQ [0..1]
+ zDimension: PQ [0..1]
+originalPerformedLesionDescription 0..1
{be converted into} be converted from
+convertedPerformedLesionDescription 0..*

The CDE for Acute Myeloid Leukemia Classification Type [66] is mapped to the BRIDG class "PerformedDiagnosis" and attribute "value." 25 CDEs use a similar object class, property, and representation term structure to represent the specific disease classification. Of those 25, 17 have been manually mapped to the BRIDG model. All are mapped to the class "PerformedDiagnosis" and the attribute "value." One can easily see how an algorithm can predict the mapping for the eight remaining CDEs. While this is a simple example, the algorithm leverages similar, though

potentially subtle, patterns to map CDEs to BRIDG classes.

Data sets

We used two different types of data sets: training sets and testing sets. The training set consisted of 1,232 CDEs mapped to the BRIDG model. To create the base training set, we first examined the 1,284 CDEs mapped to the BRIDG model by CIBMTR. These CDEs can be considered a "gold standard" because a robust team of experts, including a clinician and BRIDG representative, performed the mapping. Of the 1,284 CDEs, 1,232 are actively used and can be downloaded from the caDSR's search site, the CDE Browser [70]; the remaining 52 CDEs were retired. Given that the 52 CDEs are retired (and not easily accessible to the broader community), it was decided to exclude them from the training set.

The 1,232 active CDEs have been grouped into 57 BRIDG classes. We refined the training set to exclude those CDEs mapped to BRIDG classes associated with less than 10 CDEs. CDEs mapped to such BRIDG classes do not provide enough training examples and will interfere with the training results. The final training set consisted of 1,134 CDEs mapped to 19 BRIDG classes. We divided the data into two groups: training and verification. We tested the effectiveness of the algorithm using different training to verification ratios: 90% training and 10% verification; 75% training and 25% verification.

For testing purposes, we compared the effectiveness of the algorithm against three different sets of CDEs that had been previously mapped to BRIDG. With these testing sets, the correct BRIDG class is known and can be compared to the prediction of the algorithm. Also, each testing set has different degrees of similarity to the training data set. Since the bulk of the semantic meaning of a CDE is contained within the DEC [67] [65], we determined the degree of semantic similarity by calculating the percent of NCI Thesaurus concepts in the DECs of the testing set that also occurs in the DECs of the training set.

The first testing set consists of the previously mentioned 52 retired CDEs. Even though the retired CDEs are not actively used, their mappings to BRIDG are still valid. Furthermore, the same team mapped the retired CDEs as well as the training data. Hence, this represents the purest test case, as there are no subjective mapping differences that may lead to unexpected results. The second testing set consists of 220 CDEs that the CIBMTR mapped in 2017. These CDEs are associated with a new therapeutic domain and were mapped by a different individual and not the original mapping team. This represents a more difficult case, due to domain changes and changes in the team. However, to this point, all the mappings were done in the same organization. The third set contains CDEs created by the NCI's curation team to represent CDISC's CDASH variables. Since the BRIDG release documentation contains mappings to the CDASH variables, mappings from these CDEs to the BRIDG model can be determined. Since the CDEs in this testing set were created and mapped by a different organization than the training set, this represents an excellent opportunity to see how the algorithm performs with a different organization's content. As of December 2018, the NCI had created CDEs for more than 600 of the CDASH variables. However, not all of the CDEs had all of the attributes required by the algorithm nor were BRIDG mappings clear for all CDEs. We were able to use 186 CDEs for testing purposes.

The BRIDG model is periodically updated to reflect the changing clinical domain. The CDEs in the training set were mapped to version 3.0 of the BRIDG model, which was the current version in 2012. In 2017, version 5.0 of the model was released [60]. Since the new

CDEs were mapped to BRIDG version 5.0, we tested only those CDEs mapped to BRIDG classes present in BRIDG 3.0. Table 1 summarizes the characteristics of each testing set.

Table 1. Similarity of testing data to training data

Testing Set	Number of CDEs	Semantic Similarity
Similar	52	86.54%
Moderately Different	220	68.64%
Different	186	4.52%

Methods

Purpose and Overview of Our Method

Our method for mapping ISO 11179 CDEs to the BRIDG model is an expansion of an existing algorithm for aligning ontologies, Ontology Alignment by Artificial Neural Network (OAANN) [40, 41]. The ANN algorithm consists of training and verification phases. The goal of the training phase is to determine the best weights of the six attributes to classify the CDEs using the similarity of the CDE with each BRIDG class mapping. It outputs the top ten most probable BRIDG classes. Domain experts can use these recommendations to facilitate their mapping efforts. The verification phase verifies the accuracy of the mapping without changing the model.

Data Preparation

We determine the attribute similarity of two CDEs by first comparing the similarities between each of their six attributes. We determine the attribute similarity by first creating a matrix that calculates the similarity of each word in the attribute's phrase. According to our previous research [17], we calculate string similarity using the following equation:

$$s_{word} = 1 - \frac{d}{l}$$

The edit distance, a commonly used measure for measuring word difference, is denoted as *d*. The length of the longer string is *l*. For example, the edit distance *d* is two between word "what" and word "was". The "h" in "what" is deleted in the first step. Then, "t" is substituted by "s" in "wat". The above two steps successfully changed the word "what" into the word "was" by doing the minimum number of single-character edits, i.e. insertions, deletions or substitutions. The length of the longer word "what" is 4. Thus, the similarity between "what" and "was" is 0.5. We obtain the maximum similarity from the matrix and put the similarity into a list, $L_{word\_similarity}$. We then delete the column and the row where the maximum similarity exists. We repeat this process until the matrix is empty and get the final list of $L_{word\_similarity}$. A set of different word similarity threshold was chosen, i.e., from 0.6 to 0.9. These thresholds were applied both during the training and verification process. Once determined, the threshold does not change during the training and verification phases. We compare every similarity l_{n} from $L_{word\_similarity}$ with the threshold we set. Two words match if l_{n} is greater or equals to the word similarity threshold. Finally, the attribute similarity $s_i \; (i = 1, 2 \ldots 6)$ equals the number of words matched divided by the number of similarities in the word similarity list. Continuing the example from the Materials section, we compare the similarity of the CDE Long Names of the CDE Acute Myeloid Leukemia Classification Type [66] and the CDE Chronic Myelogenous Leukemia Classification Type [71]. Their question text is "What was the classification of the acute myelogenous leukemia?” and “What was the classification of the chronic myelogenous leukemia?” respectively. We build the word similarity matrix in Figure 5.

Figure 5. Example of word similarity matrix

The following is the similarity matrix of the question text corresponding to the CDE Acute Myeloid Leukemia Classification Type and the CDE Chronic Myelogenous Leukemia Classification Type. Their corresponding question text is “What was the classification of the acute myelogenous leukemia?” and “What was the classification of the chronic myelogenous leukemia?” respectively. After calculating the similarity between every word and generating the word similarity matrix, we build the word similarity list by sorting and obtaining the maximum similarity from the matrix. The maximum similarity is represented by grey background. Note that after obtaining the maximum similarity, the similarities of this column and this row will be ignored, meaning that they will no longer participate in the sorting.

	what	was	the	classification	of	the	chronic	myelogenous	leukemia
what	1.00	0.50	0.25	0.14	0.00	0.25	0.14	0.00	0.00
was	0.50	1.00	0.00	0.14	0.00	0.00	0.00	0.09	0.00
the	0.25	0.00	1.00	0.07	0.00	1.00	0.14	0.09	0.13
classification	0.14	0.14	0.07	1.00	0.07	0.07	0.21	0.00	0.21
of	0.00	0.00	0.00	0.07	1.00	0.00	0.14	0.09	0.00
the	0.25	0.00	1.00	0.07	0.00	1.00	0.14	0.09	0.13
acute	0.14	0.00	0.14	0.21	0.14	0.14	0.00	0.09	0.25
myelogenous	0.00	0.09	0.09	0.00	0.09	0.09	0.09	1.00	0.18
leukemia	0.00	0.00	0.12	0.21	0.00	0.13	0.25	0.18	1.00

We sort the matrix and obtain the first maximum similarity from the first row and the first column and append this similarity into $L_{word\_similarity}$. So $L_{word\_similarity}=\{1\}$. After repeating the process, $L_{word\_similarity}=\{1,1,1,1,1,1,1,1,0\}$. When we set the word similarity threshold to 0.9, the attribute similarity, $s_6 = 8/9$.

After we obtain all attribute similarities, the overall similarity between two CDEs is calculated as the weighted sum of the attribute similarities, s_1, s_2, s_3, s_4, s_5, and s_6:

$$s = \sum_{i=1}^{6}(w_i s_i)$$

where $\sum_{i=1}^{6} w_i = 1$, w_i are initialized into $1/6$ and were adjusted through the weight learning procedure.

We learn the weights for the six attributes during the group classification process. We design the learning problem as follows:

Task, T: Recommend the most likely top ten BRIDG classes for a CDE

Performance measure, P: Accuracy measurements for the 1,232 CDEs already grouped

Training experience, E: a set of classified CDEs by manual matching

Target function, V: a list of class recommendation

Target function representation: $V(b) = \sum_{i=1}^{6} w_i s_i$

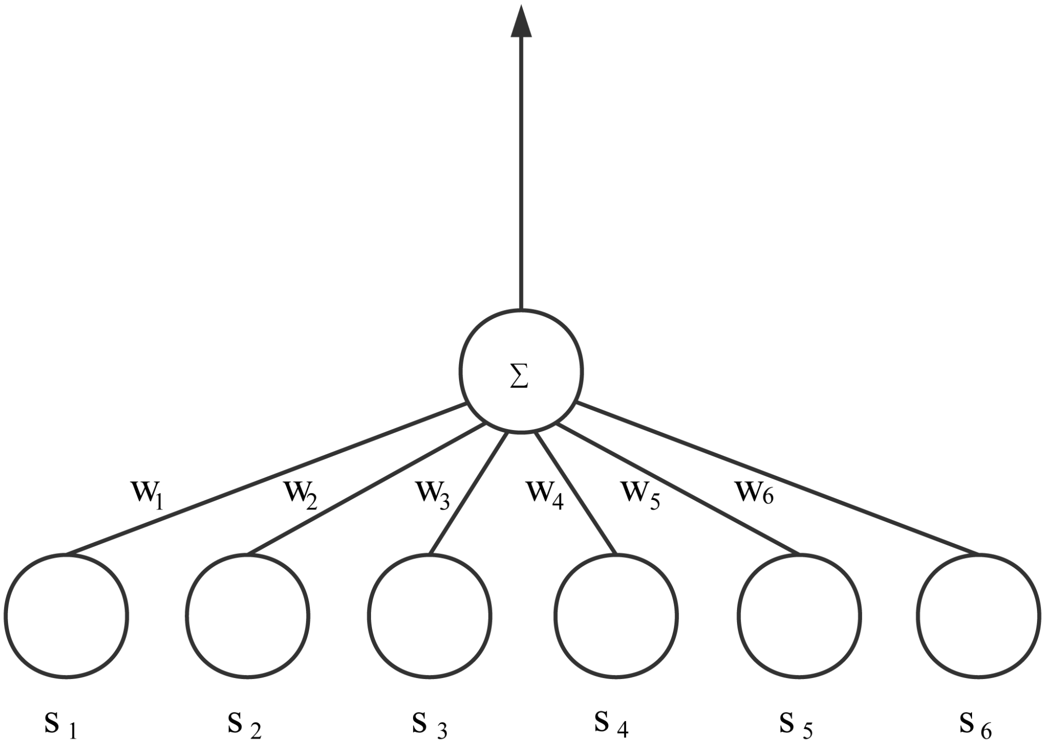
Network Design

We modified our previous network design [17] to use a two-layer 6 × 1 network. Figure 6 illustrates the network's vector inputs of s_1, s_2, s_3, s_4, s_5, and s_6. The neural network output is the overall similarity between two CDEs. Section B of the Method part gives the way to calculate the similarity value s between two CDEs.

Figure 6. Neural network structure

The inputs are the similarities of six attributes. The output is the overall similarity between two CDEs.

$$s_{overall} = w_1 s_1 + w_2 s_2 + w_3 s_3 + w_4 s_4 + w_5 s_5 + w_6 s_6$$



Hypothesis space and our searching strategy

The hypothesis space is a 6-dimensional space consisting of six vectors $w_i (i = 1, 2, ..., 6)$. We use gradient descent as our training rule. We minimize the training error of all training examples, so our task is to find such a vector. Training error E is calculated as

$$E(\vec{w}) \equiv \frac{1}{2} \sum_{d \in D} (t_d - o_d)^2$$

in accordance with [72]. In $E(\vec{w})$, the set of training examples is denoted as D, the target output for training example d as t_d , the output of the network for d as o_d. Two concepts are the base of training error, which are s_{avg} and $s_{avg\_other\_cls}$.

s_{avg}: Let us say that $CDE_{\mathrm{m}\text{-}n}$ belongs to $BRIDG_m$. From $BRIDG_1$, we pick up the first element from $BRIDG_1$ which is $CDE_{\mathrm{m\text{-}1}}$and calculate the overall similarity between the selected $CDE_{\mathrm{m\text{-}1}}$and other CDEs belong to this BRIDG class, separately. Then average these total similarities to get s_{avg}.

$s_{avg\_other\_cls}$: We use the same CDE $CDE_{\mathrm{m\text{-}1}}$, to calculate the similarity of this CDE with other CDEs from classes other than $BRIDG_1$. Then average those similarities to get $s_{avg\_other\_cls}$.

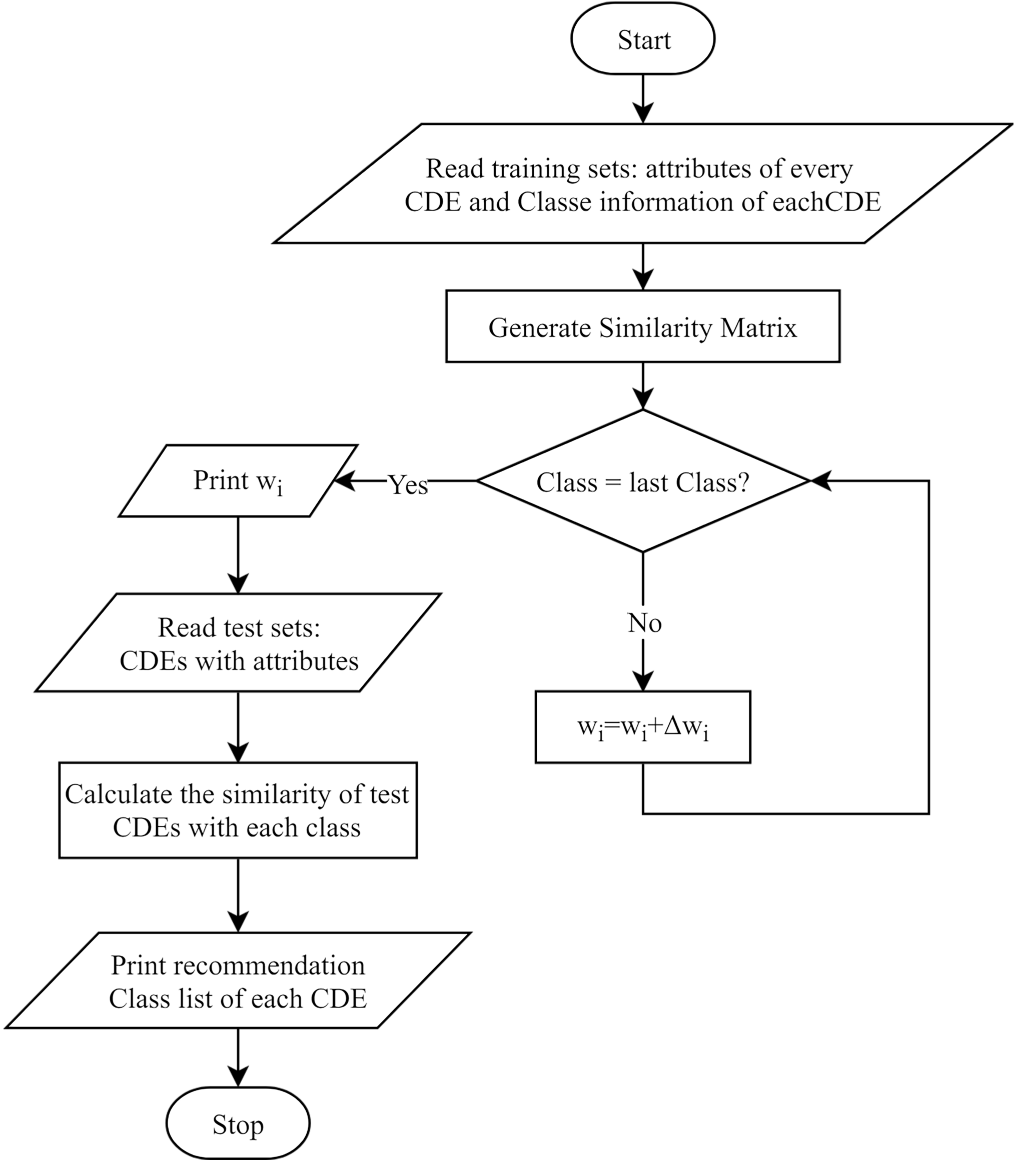
As we apply the gradient descent, we traverse one BRIDG class after another, pick up $CDE_{\mathrm{m\text{-}1}}$ to calculate s_{avg} and $s_{avg\_other\_cls}$, and finally make the adjustment of w_i. The adjustment is represented by Δw_i. The calculation of Δw_i is

$$\Delta w_i \equiv \eta \sum_{d \in D} (t_d - o_d) s_{id}$$

in accordance with [72]. η represents the learning rate of Δw_i. The procedure is shown in Figure 7. Notice that the BRIDG class is trained from less to more according to their CDE amount in each class. As the number of data increases, our accuracy rate generally increases.

Figure 7. Training flow chart

The training flow demonstrates the process of training and recommendation.


Start
Read training sets: attributes of every CDE and Classe information of eachCDE
Generate Similarity Matrix
Class = last Class?
Yes
No
Print w_i
w_i=w_i+Δw_i
Read test sets: CDEs with attributes
Calculate the similarity of test CDEs with each class
Print recommendation Class list of each CDE
Stop

Error Type

In order to determine if the weights need to be updated and how to update the weights, two different formulas are needed. For the former, we need to calculate the training error, which is given in the following formula:

$$E(\vec{w}) \equiv \frac{1}{2}\sum_{d \in D}(s_{\max} - s_{avg})^2$$

We want our s_{avg} to be as large as possible to minimize $E(\vec{w})$, so the larger value $s_{\max}$ should be the target output. As the goal is to minimize error, if the error is greater than zero and the number of updates are below a provided threshold, we update the weights using the following formula:

$$\Delta w_i \equiv \eta \sum_{d \in D}(s_{\max} - s_{avg})s_{id}$$

t_d is represented by $s_{\max}$. $s_{\max}$is the higher value between s_{avg} and $s_{avg\_other\_cls}$.

o_d is represented by s_{avg}, which is the real output.

The weights are updated iteratively until verification accuracy stabilizes. As the iteration time (number of updates) increases, when the other parameters are using the same combination, the verification accuracy will increase and then gradually become stable. Experimentally, it was determined that when the iteration time is larger than 50, the error ($E(\vec{w})$) will only fluctuate up and down by 1%. Since the results stabilized at this point, the maximum number of iterations was set to 50. Additionally, the learning rate η is fixed to 0.05. The value was chosen after performing some exploratory experimentation that indicated this value resulted in stable, repeatable results compared to higher values, and equivalent performance with respect to lower values.

Once the formulae were chosen, and the parameters fixed, we could then train the model to discover the six weights. Once the training is completed, we then proceed with the verification and testing results, which are discussed in Results.

Results

Overall Testing Approach

Our testing consisted of two phases. First, we verified the algorithm using a subset of the training data. Then, we tested the algorithm using three testing sets of CDEs for which we had pre-existing mappings. For each testing set, we determined the best match rate and the parameters needed to obtain the best match rate. We defined the match rate as the percent of total CDEs for which the existing BRIDG class mapped to the CDE was present in the list of top ten BRIDG classes returned by the algorithm. Table 2 shows the parameters used for verification and testing along with the optimal values as determined by our testing.

Table 2. Algorithm parameters

Parameter	Description	Values Tested	Optimal Values
Training ratio	Ratio of training to verification data	75% training and 25% verification 90% training and 10% verification	75% training and 25% verification
Training CDEs per BRIDG class	Determines the training list	4 - 10	8
Similarity threshold	Determines the threshold for considering two words to be similar	0.6 – 1.0	0.7 or 0.8

Verification of Training Data

Table 3 shows the accuracy of the algorithm when it returns one to ten potential BRIDG classes for each CDE in the verification data set. When the training-verification ratio changes from 3:1 (75%/25%) to 9:1 (90%/10%), the training performance increases. When the algorithm returns more potential class matches, the accuracy of the algorithm increases and the performance differences between the 3:1 and 9:1 versions of the algorithm decreases. When the algorithm returns 10 potential class matches reaches ten, the accuracy of the algorithm reaches more than 90%. Also, we found that there was a significant increase in accuracy when the number of potential class matches returned increases from 1 to 2. This means the efficiency of the calculations and the cost performance of the results are relatively high. From the point of view from domain experts, returning 10 potential matches is reasonable because of the high accuracy.

Table 3. Accuracy with different training validation

Top n	Accuracy (training set: verification set=3:1) (%)	Accuracy (training set: verification set=9:1) (%)
1	33.99%	41.52%
2	51.96%	63.16%
3	64.71%	73.10%
4	71.90%	80.70%
5	76.14%	83.04%
6	82.03%	85.38%
7	85.29%	87.72%
8	86.93%	90.06%
9	89.22%	91.23%
10	92.16%	94.15%

Testing with Pre-Existing Mappings

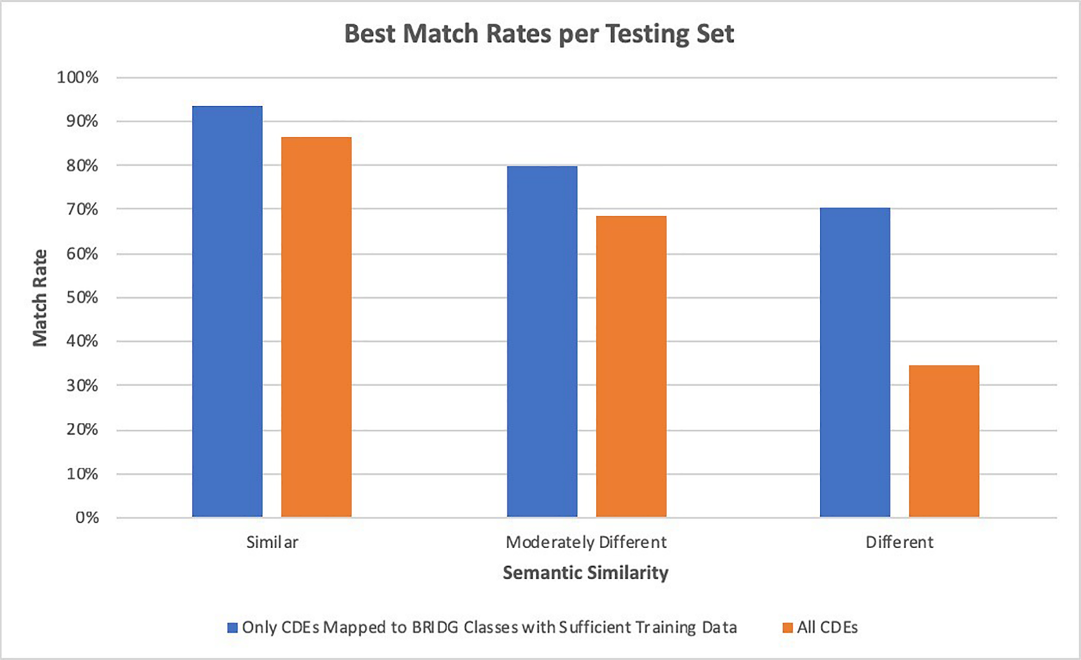
Using the testing data sets described in the Materials section, we evaluated the algorithm to determine which combination of parameters produced the best match rate. Additionally, for each testing set, we ran the algorithm twice. One run utilized all CDEs contained in the testing set and the other run used only those CDEs that had been manually mapped to a BRIDG class for which there was sufficient training data. The goal of the last testing scenario was to determine the accuracy of the algorithm when we knew that it should produce an accurate match.

Overall, for the testing sets, we found that training-verification ratios of 3:1 (75%/25%) and 9:1 (90%/10%), performed equally well. However, the 3:1 version achieved this performance while requiring fewer training CDEs per BRIDG class (8 vs. 10). This makes the algorithm more flexible when testing novel CDEs. A similarity threshold of 0.7 or 0.8 achieved optimal match rates.

Depending upon the semantic similarity of the testing set and the availability of sufficient training examples, the algorithm produced match rates between 34 – 94%. The lowest match rates occurred when testing the algorithm using CDEs that were semantically different than the training set (4.52% similarity). Figure 8 summarizes the testing results.

Figure 8. Best match rates per testing set

Bars refer to the best matching rate for testing sets with different degrees of semantic similarity compared to the training set: similar, moderately different, and different. The blue bars represent the situation in which the testing set contains only CDEs mapped to BRIDG classes with sufficient training data. The orange bars represent the situation that the testing set contains all CDEs.



Discussion

General Result Analysis

Fundamentally, the algorithm performs very well and has the potential to significantly reduce the mapping burden while improving the quality of the mappings. It should be noted that the use of our algorithm represents a semi-automated mapping process. While the algorithm can make suggestions, it will never replace the need for subject matter expert review and approval. While some testing scenarios resulted in lower match rates, the lower match rates are the result of two factors: semantic differences in the CDEs and lack of training data for the BRIDG class mappings. The testing results indicate that the greater the difference in the DEC concepts from the DEC concepts in the training set, the lower the overall match rate: 87% for the most similar data set versus 34% for the most different when all CDEs in the testing set were tested. However, the match rate for all testing sets increased markedly when the algorithm was run against only those CDEs that mapped to a BRIDG class with sufficient training data: 94% for the most similar data set versus 70% for the most different. This indicates that the availability of training data is more important than the semantic similarity of the testing set to the training set.

Increasing the size of the training set addresses both the semantic differences in the CDEs and the lack of BRIDG class instances in the training set. As the algorithm is used, and the appropriate mappings reviewed by a team of experts, the approved mappings can be added to the training set. This will incrementally improve the functioning of the algorithm. However, it is important to note that just adding data to the training set is not sufficient. One must add training data that increases the variability of the training set.

When testing the Semantically Different testing set, we tried expanding the training data to include the Semantically Similar, and Moderately Different testing sets. This did not increase the match rate and, in one scenario, actually decreased the match rate. A closer examination of the expanded training set revealed that of the 38 BRIDG classes represented in the expanded training set, only five of them resulted in new training instances.

The algorithm can also assist with validating existing mappings. For example, the CDE Other Therapeutic Procedure Administered Indicator [73] was manually mapped to the BRIDG class "PerformedDiagnosis." While the algorithm did match to this class, the ranking was ten. A closer review of the potential classes returned for this CDE showed that the second-ranked class "PerformedProcedure" was a better match. Indeed, of the six CDEs in the combined data set that had a match with a rank of ten, half of the manual mappings were potentially incorrect, and the algorithm returned better potential matches.

Weights

The weights are all significant in six attributes although they are slightly different from each other. This means that the attribute suggestion of the domain experts is accurate. We found that the weight for the Question Text is always slightly higher than the other attributes. Regardless of the parameters, from the training result, the average is 0.198, and the standard deviation is 0.010 for the initial 1,232 CDEs. The average is 0.186, and the standard deviation is 0.001 for the 1504 combined CDEs. Since the Question Text is a complete sentence, it is more semantically useful and plays a greater role in classification. The second column of Table 4 contains an example of trained weights for the 1232 initial

CDEs when Top n, iteration times, training-verification ratio, training elements in class, and similarity threshold is set to 10, 50, 9:1, 10, 0.7, separately. The verification of this combination is 159 out of 171, which is 92.98%. The third column of Table 4 contains an example of trained weights for the 1504 CDEs with the same parameters as the second column. From this combination, the verification is 187 out of 225, which is 83.11%.

Table 4. Example of attribute weights

Weight Name	Weight	
	Verification - 1232	Verification – 1504
w_1 (CDE Long Name)	0.174620962	0.174501362
w_2 (Object Class)	0.156449029	0.155271219
w_3 (Property)	0.159148243	0.156401809
w_4 (Value Domain Long Name)	0.159921116	0.164160267
w_5 (Representation Term)	0.160018435	0.164215956
w_6 (Question Text)	0.189842216	0.185449386

Overfitting

One common problem encountered is overfitting, where the neural network picks weights tailored for the training instances versus the problem. We took two steps to address this. First, to ensure the team did not subconsciously introduce bias, CDEs were selected randomly instead of alphabetically.

Second, as previously noted, the ratio of training and verification samples had been modified. As seen in Table 3, the changes in proportion had minimal impact for the larger values of n. For lower values, more training data led to improvements, indicating the learned weights are fitting to class concepts rather than instances. When the new CDEs are introduced, the 3:1 ratio became preferable as the new CDEs contain new classes. The improvement with the lower ratio indicates that the new classes vary from the initial set of classes; this indicates the 9:1 weights were learning the known class representations. A lower ratio allows more "flexible" weights, which handles new classes at the expense of some accuracy. This indicates that overfitting is not happening.

Finally, the results shown in Figure 8 also provide evidence that overfitting problem is not a problem. The match rate for similar concepts is high, over 80% in both cases. Hence, the networks are learning concepts versus instances. The fact that the "moderately different" are again achieving well above random matching is indicative that the patterns the neural network are transferrable. Again, this is evidence that the networks are looking at patterns of behavior rather than particular concepts. The most interesting result is the "different" semantic concepts. When using CDEs that have sufficient training samples, the CDEs that have very different semantic meanings are achieving approximately a 70% match rate; this indicates the patterns discovered by the neural networks are highly transferable. It is only when all CDEs concepts are used in training data (even those with few examples) that we see a large drop. This is indicative that the neural network is (a) unable to extract what are the meaningful patterns for each type of CDE and (b) the low example CDEs are effectively introducing noise. While this leads to low match rates, it also demonstrates that the neural network is attempting to find meaningful, transferable patterns.

Conclusions

The CDEs in the caDSR provide robust data point definitions that help ensure that clinical data adheres to the FAIR data principles: findable, accessible, interoperable, and reusable [74]. Mapping CDEs to the BRIDG model increases their value by providing a contextual framework and by facilitating their mapping to a variety of other data standards such as CDASH and SDTM.

Because manual mappings have many disadvantages including being extremely time-consuming and rather error-prone, there is an urgent need to map CDEs to BRIDG classes in a semi-automated manner. To handle this important challenge, we have developed an ANN-based machine learning algorithm that semi-automates the mappings between CDEs and BRIDG classes, followed by recommending a list of candidate classes to which the CDE of interest may belong. We evaluated our algorithm using a set of real-world use cases, and our experimental results showed that our algorithm has the potential to not only significantly reduce the mapping burden but also greatly improve the quality of the mappings.

In our future work, we plan to make several changes to the algorithm to further improve its effectiveness. Most significantly, we can leverage the wealth of information contained in the NCI Thesaurus ontology. Each CDE's object class, property, and representation terms are created using concepts found within the NCI Thesaurus [44]. The NCI Thesaurus combines a reference terminology with an ontology to create a computable source of semantic information. In addition to providing a consistent naming convention and detailed definitions, the NCI Thesaurus also provides synonyms, semantic types, and relationships between concepts ^{75}. Expanding the algorithm to include this information will provide users with even more robust matching results.

We will also be adapting the algorithm to use the Data Element Concept (DEC) long name. We did not include the DEC long name in the algorithm because the caDSR tooling automatically constructs it by concatenating the Object Class and Property. Therefore, we assumed that Object Class and Property completely represent the semantics. However, 40% of the Data Element Concept Long Names in the training set were not an exact concatenation of the Object Class and Property. A future iteration of the algorithm should include the Data Element Concept Long Name.

There are some potential enhancements to the current training process that will be considered in the future. For instance, currently the training loss is considered as the sum of squared error, which may lead to an unnecessarily large gradient and possibly make the training process unstable under certain circumstances. To address this in future work, we may define the training loss as mean squared error. In addition, the problem may be defined as a multi-label classification problem, in which case binary cross-entropy may serve as a better loss measure. Moreover, using Stochastic Gradient Descent with momentum rather than the conventional gradient descent has been proposed for the future, as it can result in a smoother training process. Finally, we may also test the utility of Early Stopping, which stops the learning process if the error doesn't decrease after a given number of epochs. This may make it more robust to inclusion of additional data, where 50 iterations may no longer be sufficient.

Acknowledgements

We would like to thank the reviewers from the Biomedical and Health Informatics Workshop in the IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2018 for their comments, which were incorporated into the present work.

We also would like to acknowledge the financial support provided by the National Cancer Institute, which is described further in the Funding Section.

Availability of data and materials

All CDEs used for testing and training can be downloaded at: https://cdebrowser.nci.nih.gov/cdebrowserClient/cdeBrowser.html#/search. To find the CDEs, search for the following:

Context: NCIP

Classification Scheme: Artificial Neural Network Algorithm for BRIDG Mapping

Classification Scheme Items:

- Training Set
- Testing Set - Semantically Similar
- Testing Set - Moderately Different
- Testing Set -Semantically Different

Please note that search options must be adjusted to include CDEs that have been retired.

The source code of the artificial neural network algorithm and the XML input files for all experiments can be downloaded at: https://github.com/Abclisy/ANN-in-CDE

The publication cost of this article was funded by The National Cancer Institute (NCI) of The National Institutes of Health (NIH), under the Award Numbers U01CA180982 and U01CA180940. The views contained in this paper are solely the responsibility of the authors and do not represent the official views, either expressed or implied, of the NIH or the U.S. government.

Chapter 4: Semi-Automated Mapping Via a Graph Database

Challenges in Using a Graph Database to Represent and Analyze Mappings of Cancer Study Data Standards

Robinette Renner, MS MHI, Guoqian Jiang, MD PhD

A version of this paper appeared in:
American Medical Informatics Association Annual Informatic Summit 2020 (in press)

Authors' contributions

RR designed the overall study, implemented the graph database, implemented the graph-based algorithm, prepared the use case, performed the experiments, analyzed the results, and drafted the manuscript. GJ directed the research project, discussed the implementation, assisted with the analysis, and reviewed the manuscript.

Abstract

While using data standards can facilitate research by making it easier to share data, manually mapping to data standards creates an obstacle to their adoption. Semi-automated mapping strategies can reduce the manual mapping burden. Machine learning approaches, such as artificial neural networks, can predict mappings between clinical data standards but are limited by the need for training data. We developed a graph database that incorporates the Biomedical Research Integrated Domain Group (BRIDG) model, Common Data Elements (CDEs) from the National Cancer Institute's (NCI) cancer Data Standards Registry and Repository, and the NCI Thesaurus. We then used a shortest path algorithm to predict mappings from CDEs to classes in the BRIDG model. The resulting graph database provides a robust semantic framework for analysis and quality assurance testing. Using the graph database to predict CDE to BRIDG class mappings was limited by the subjective nature of mapping and data quality issues.

Introduction

Sharing clinical data can foster innovation[76, 77], reduce the time from bench to bedside[77, 78], improve patient outcomes[77], reduce research costs[77], and increase transparency[77]. Unfortunately, the heterogeneous nature of data stored in local systems creates data silos that make data sharing nearly impossible[79]. While the use of data standards can facilitate the sharing of clinical data[79], sometimes one must support multiple standards. For example, regulatory submissions to the Food and Drug Administration (FDA) must be submitted using Clinical Data Interchange Standards Consortium (CDISC) standards such as the Study Data Tabulation Model (SDTM)[61]. Electronic Medical Record

(EMR) systems, on the other hand, are beginning to use HL7's Fast Healthcare Interoperability Resources for electronic data transmission[80]. If researchers want to consume data from an EMR and use it for studies requiring FDA submission, they must support both standards. This landscape is complicated by the need to annotate clinical terms with concepts from multiple clinical vocabularies such as the use of Medical Dictionary for Regulatory Activities (MedDRA) concepts for adverse event report and Logical Observation Identifiers Names and Codes (LOINC) concepts for reporting laboratory values[81]. Most often mapping between standards is done manually which is time-consuming, expensive, and error-prone[29]. The need to support multiple standards has resulted in an urgent need for tools to map between the standards.

The experience of the Center for International Blood and Marrow Transplant Research (CIBMTR) is an excellent case study of the problems with manual mapping. The CIBMTR collects outcomes data for cellular therapy research[58]. To facilitate data collection, the CIBMTR offers electronic data submission directly from a transplantation center's database[58]. Common Data Elements (CDEs) obtained from the National Cancer Institute's (NCI) cancer Data Standards Registry and Repository (caDSR) provide the foundation for data transmission. These CDEs have been mapped to the Biomedical Research Integrated Domain Group (BRIDG) model[59]. The BRIDG model captures the semantics of data used for clinical research and regulatory submission[60]. The BRIDG model has been harmonized and mapped to a variety of data standards and data models. For example, BRIDG has been mapped to CDISC's Clinical Data Acquisition Standards Harmonization (CDASH) and SDTM standards[22, 82, 83]. The US Food and Drug Administration (FDA) in collaboration with Health Level 7 (HL7) created the Common

Data Model Harmonization (CDMH) project. This project has mapped BRIDG to four clinical data models[23]: Sentinel[25], Patient-Centered Outcomes Research Network (PCORNET) Common Data Model[26], Informatics for Integrating Biology & the Bedside (i2b2)[27], and Observational Medical Outcomes Partnership (OMOP)[28]. The CDMH project also mapped BRIDG to HL7's latest standard, Fast Healthcare Interoperability Resources (FHIR)[24]. The goal of the CDMH project is to provide observational data to researchers, which is a typical clinical use case that demonstrates the value of mappings of clinical study data standards. Unfortunately, while mapping CDEs to the BRIDG model can facilitate the adoption of other clinical data standards, mapping to the BRIDG model itself is a labor-intensive process. A semi-automated mapping strategy could help reduce the mapping burden.

A semi-automated application that maps CDEs to the BRIDG model[84, 85] was created using the Ontology Alignment by Artificial Neural Network (OAANN) developed by Huang et al.[40, 41]. The new algorithm predicts CDE to BRIDG class mappings using key attributes of the ISO 11179 metamodel for CDEs and a robust training set of nearly 1,200 CDEs that have been manually mapped to an appropriate BRIDG class. It returns a list of 10 potential BRIDG classes for a CDE of interest. A subject matter expert then reviews the list and selects the most appropriate BRIDG class. The algorithm was able to predict BRIDG class mappings with up to 94% accuracy. Accuracy was calculated by dividing the number of the predicted mappings that are correct by the number of the manually mappings.

While this approach is effective, it has some limitations. First, the algorithm uses pattern recognition only. The underlying semantics of the CDE and potential synonyms are not considered. Second, if sufficient training data does not exist for a particular BRIDG class, then the algorithm will never map a CDE to it. Graph databases have the potential to overcome these limitations.

Graph databases represent data as a collection of nodes (classes) and edges (relationships). In contrast to relational databases, they provide an efficient way to represent highly interconnected data[45]. Several researchers have used graph databases for a variety of mapping and data integration problems. For example, Alqahtani et al. used a graph database to integrate business data from two heterogeneous data sources[46]. Johnson et al. combined disparate tumor-related models using a Neo4j graph database and annotated the semantic terms with concepts from the NCI Thesaurus[47]. Campbell et al. used a graph database to represent the SNOMED-CT terminology[51]

The work of Ulrich et al. is particularly interesting[52]. They used a graph database to map data elements in disparate metadata repositories and tested the application using more than 600 cancer-related data elements[52]. Their algorithm used the five-gram algorithm and the metric Longest Common Subsequence to determine the similarity of two CDE’s name. If the CDE was associated with a list of allowed values, they supplemented this analysis with a comparison of the allowed values. While this work is promising, it has several limitations. First, their analysis is based only on two CDE attributes: long name and allowed values. It is not clear if the CDEs used in the work of Ulrich et al. were based on the ISO 11179 metamodel. Therefore, those two attributes may have been the only ones available for analysis. Second, their analysis is based strictly on pattern recognition. It does not take into consideration the meaning of the terms (i.e. semantics) in the CDE.

Mapping CDEs to the BRIDG model lends itself well to a graph database approach. The BRIDG model has already been implemented as a Neo4j graph database[53] and other researchers have incorporated the NCI Thesaurus into their graph database work[47]. CDEs based on the ISO 11179 metamodel, such as those found in the caDSR, consist of interrelated components. Such relationships could be represented using a graph model. Also, caDSR CDEs are annotated with concepts from the NCI Thesaurus[67]. BRIDG classes are annotated with definitions that are associated with NCI Thesaurus concepts that explicitly document BRIDG as the definition's source. Therefore, the NCI Thesaurus serves as a common point of reference between CDEs and the BRIDG model. A graph database could leverage this relationship to facilitate mapping CDEs to BRIDG classes.

In this paper, we present our work to develop a graph database that incorporates the BRIDG model, caDSR CDEs, and NCI Thesaurus concepts and to use graph-based algorithms to predict potential BRIDG class matches for the CDEs. Finally, we discuss the feasibility of using a graph-based mapping approach.

This work is significant because it is the first comprehensive representation of the BRIDG model, caDSR CDES, and NCI Thesaurus concepts in a graph database that we are aware of. It also highlights the subjective nature of mapping data standards and the challenges in developing a semi-automated mapping strategy.

Methods

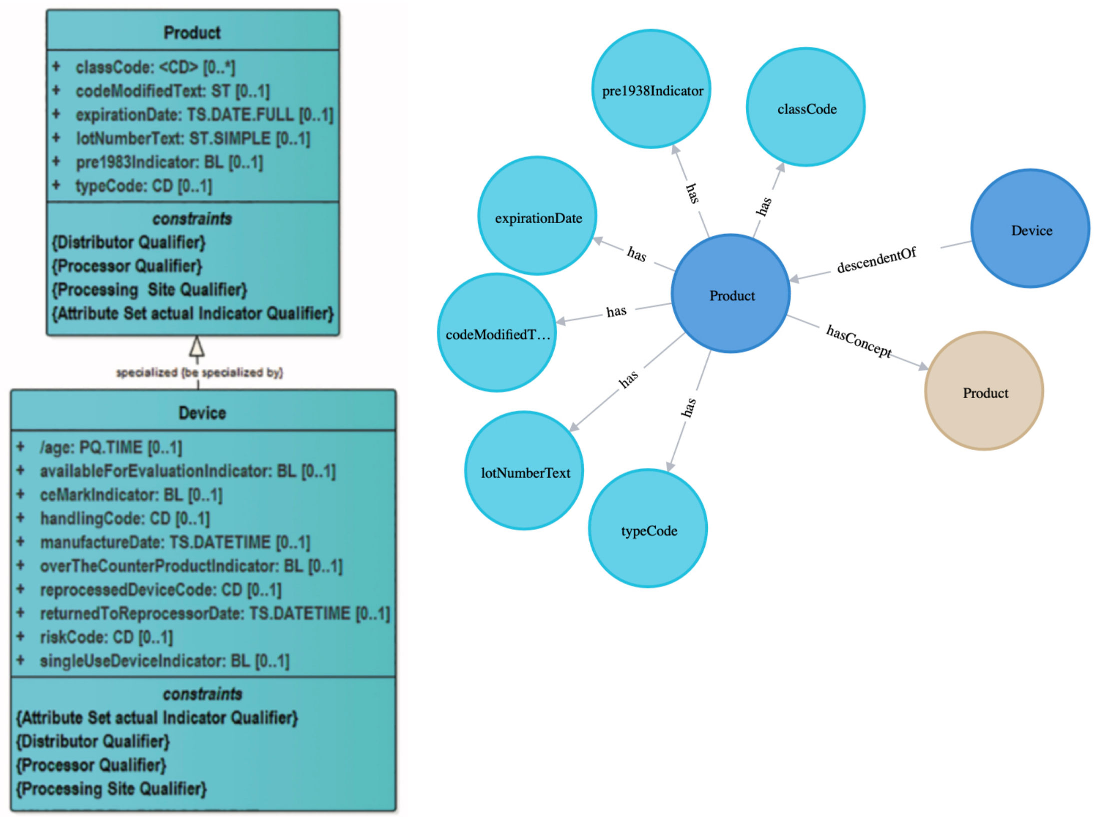
Graph Model Development

<u>BRIDG Model</u>: The foundation of the BRIDG portion of the graph model was the Neo4j implementation of the BRIDG model provided by Jane Pollack[53]. She based her

graph model on the UML model of the BRIDG model available on the BRIDG website[60]. In her graph model, classes and attributes are represented as nodes. Relationships between classes and attributes or classes and classes are represented as edges. Each node and edge has properties that contain additional information. We then supplemented Pollack's model with edges from the BRIDG class node to the appropriate NCI Thesaurus concept node. We determined the appropriate concept by mapping the BRIDG class name to an NCI Thesaurus concept that had both the BRIDG class name listed as a synonym and a matching definition. Since the focus of this work was on matching CDEs to BRIDG classes, BRIDG attributes were not associated with NCI Thesaurus concepts. Figure 9 shows a portion of the BRIDG UML[22] and the corresponding graph model.

Figure 9. UML representation of a section of the BRIDG model[22] (left) and its corresponding graph model representation (right).

Comparison of the UML and graph database representations of a portion of the BRIDG model.

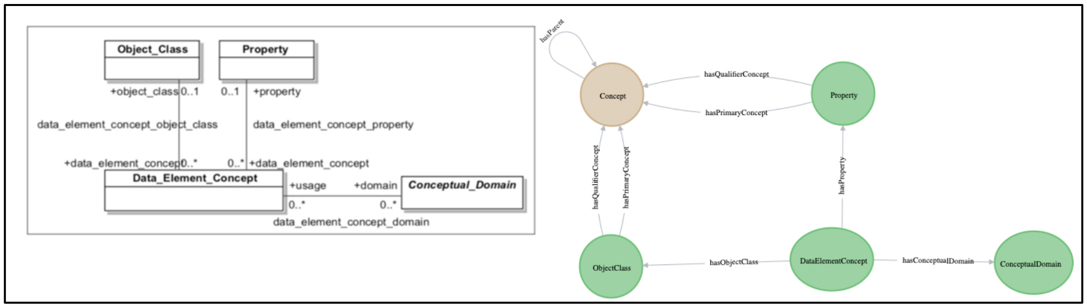

Product
+ classCode: <CD> [0..*]
+ codeModifiedText: ST [0..1]
+ expirationDate: TS.DATE.FULL [0..1]
+ lotNumberText: ST.SIMPLE [0..1]
+ pre1983Indicator: BL [0..1]
+ typeCode: CD [0..1]
constraints
{Distributor Qualifier}
{Processor Qualifier}
{Processing Site Qualifier}
{Attribute Set actual Indicator Qualifier}
specialized (be specialized by)
Device
+ /age: PQ.TIME [0..1]
+ availableForEvaluationIndicator: BL [0..1]
+ ceMarkIndicator: BL [0..1]
+ handlingCode: CD [0..1]
+ manufactureDate: TS.DATETIME [0..1]
+ overTheCounterProductIndicator: BL [0..1]
+ reprocessedDeviceCode: CD [0..1]
+ returnedToReprocessorDate: TS.DATETIME [0..1]
+ riskCode: CD [0..1]
+ singleUseDeviceIndicator: BL [0..1]
constraints
{Attribute Set actual Indicator Qualifier}
{Distributor Qualifier}
{Processor Qualifier}
{Processing Site Qualifier}
pre1938Indicator
classCode
expirationDate
Device
Product
codeModifiedT…
lotNumberText
typeCode
Product
has
has
has
has
has
has
descendentOf
hasConcept

NCI Thesaurus: The concept attributes available within the NCI Thesaurus[86] served as the foundation for the NCI Thesaurus portion of the graph model. In addition to creating a node for Concept, we created nodes for concept attributes such as Semantic Type, Synonym, and Definition. Representing these attributes as nodes allowed for associating multiple instances of an attribute to a concept. Additional properties could also be associated with an attribute. We captured parent-child relationships between concepts as an edge between two concept nodes.

caDSR CDEs: The ISO 11179 metamodel[65, 87] served as the foundation for the caDSR Common Data Element (CDE) portion of the graph model. A CDE consists of two parts: a Data Element Concept (DEC) and a Value Domain. The DEC is the conceptual representation of the CDE and should contain the bulk of the semantic meaning for the CDE[67, 87]. The DEC consists of two parts: the Object Class and the Property. The Object Class is equivalent to a UML class and the Property to a UML attribute[87]. The Object Class and Property are both defined using concepts from the NCI Thesaurus. The Value Domain is associated with a Representation Term that describes the meaning of the data being captured. The Representation Term is also defined using concepts from the NCI Thesaurus[67]. We created nodes for the CDE and its main components, such as the Data Element Concept (DEC) and Value Domain (VD). Each node has properties that contain additional attributes about each component. Edges were created to document the relationships between the various nodes. Also, edges to the associated NCI Thesaurus concepts were added. Figure 10 shows the Data Element Concept portion of the ISO 11179 metamodel[87] and the corresponding section of the graph model.

Figure 10. Data Element Concept section of the ISO 11179 metamodel [87] (left) and its corresponding graph model representation (right).

Comparison of part of the ISO 11179 metamodel and its corresponding graph database representation.



Database Implementation

BRIDG: We populated the BRIDG portion of the Neo4j graph database using the CSV load files and Cypher queries developed by Jane Pollack[53]. To allow for consistent comparison between the graph database mappings and the mappings provided by the Artificial Neural Network algorithm[84 85], we used version 3.2 of the BRIDG model. A CSV load file was created with the relationships between each BRIDG class and the associated NC Thesaurus concepts. A Cypher query was developed to import the CSV load file.

NCI Thesaurus: To populate the NCI Thesaurus section of the Neo4j graph, we first downloaded the Web Ontology Language (OWL) file from the NCI Thesaurus download website[88]. We created a Python script that leveraged the owlready2[89, 90] package to parse the OWL file. The script then generated separate CSV files for each type of NCI Thesaurus node and relationship. For example, separate import files were created for the Concept and Semantic Type nodes. Separate Cypher queries were created that imported each CSV import file.

caDSR CDEs: To populate the CDE portion of the Neo4j graph database, we first downloaded from the CDE Browser website[70] an XML file that contained the information for the 1,689 CDEs used in the ANN mapping project [84 85]. To import the information, we used an approach that was similar to the approach for importing the NCI Thesaurus information. We created a Python script that parsed the XML file and created CSV files for each type of node and edge. We then created a set of Cypher queries that imported each CSV file.

Quality Assurance Testing

BRIDG: To ensure that the Neo4j database was populated correctly, we performed basic quality assurance testing. We tested the BRIDG content by developing a Python script that queried the Neo4j database and created an Excel file with the BRIDG class name and definition. The XlsxWriter Python package[91] was used to create the Excel file, and the Neo4j Bolt Driver Python package[92] was used to query the database. We supplemented the Excel file with the BRIDG class names and definitions obtained from the BRIDG website[60]. The information from both sources was then compared to ensure consistency.

caDSR CDEs: We tested the quality of the CDE information the Neo4j graph database by creating a Python script that queried the Neo4j database [92], the XML source document downloaded from the CDE Browser [70], and generated an Excel file [91] containing the CDE information from both sources. The Excel file was then analyzed to ensure that the CDE information in the Neo4j database was accurate.

NCI Thesaurus: Testing of portions of the NCI Thesaurus content in the Neo4j occurred during the testing of both the BRIDG and CDE content. Both of those testing approaches contained NCI Thesaurus concept attributes such as concept preferred name, concept unique identifier, and definition. Robust quality assurance testing of all concept attributes such as synonyms and semantic types is in progress.

Algorithm Implementation

We leveraged the Shortest Path algorithm in the Neo4j library[93] to determine the shortest path between a CDE and a BRIDG class. We developed a Python script that iteratively ran a Cypher query that found the unweighted shortest path between a CDE and each BRIDG class and returned the distance. Distance was measured as the number of edges between the CDE and the BRIDG class. We tested the algorithm by running it against all of the CDEs used in the ANN mapping project[84, 85]. Since those CDEs have been manually mapped to the BRIDG model by a team of subject matter experts, they represent a gold standard by which we could determine the accuracy of the shortest path algorithm.

The Python script returned an Excel file that contained the CDE public ID, CDE Long Name, the target BRIDG class name, the distance between the CDE and the target BRIDG class, and a text description of the path in terms of the nodes used. The manually mapped BRDIG class was added to the Excel spreadsheet to make it easier to determine the accuracy of the algorithm's prediction. We also developed a Cypher query that visualized specific mapping paths.

Results

Graph Model

Figure 11 shows the graph model of the BRIDG model, caDSR CDEs, and NCI Thesaurus concepts. This model shows how the NCI Thesaurus represents a nexus point connecting BRIDG classes to caDSR CDEs and providing them with. rich semantic annotations. Table 5 summarizes the number of key nodes in the graph database. In addition, the BRIDG classes are associated with 205 distinct NCI Thesaurus concepts, and the CDEs are associated with 1,143.

Figure 11. Representation of the graph-based data model.

Overall model of the graph database with the BRIDG portion in blue, NCI Thesaurus portion in brown, and the caDSR CDEs in green.

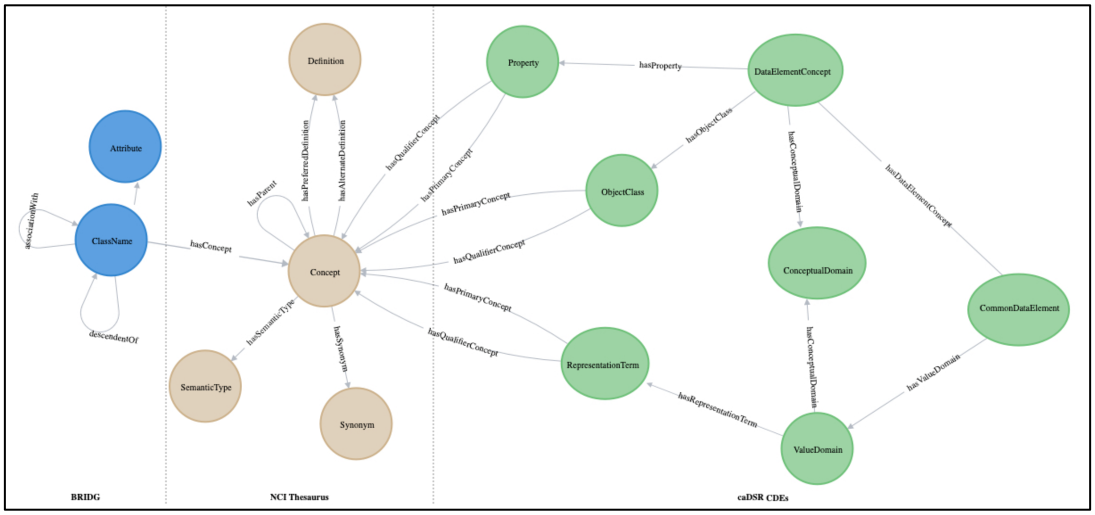


Table 5. Summary of the number of key nodes in the database.

Section	Node	Count
BRIDG	Class	233
	Attribute	639
caDSR	Common Data Element	1689
	Data Element Concept	1193
	Object Class	622
	Property	638
	Value Domain	588
	Representation Term	524
NCI Thesaurus	Concept	147,010
	Semantic Type	128
	Synonym	426,150

Quality Assurance Testing

Quality assurance testing of the BRIDG section of the model verified that BRIDG classes were associated with the correct NCI Thesaurus concept where possible. 12% (28 out of 233) of the BRIDG classes are not associated with an NCI Thesaurus concept. This gap impacted 5 of the 1,689 CDEs that we analyzed.

Quality assurance testing of the caDSR CDEs detected issues when the CDEs were constructed using concepts from the NCI Metathesaurus instead of the NCI Thesaurus. This occurred in 4 CDEs out of the 1,689. caDSR best practice states that CDEs should be constructed using concepts from NCI Thesaurus [87]. Therefore, these CDEs were incorrectly constructed. The graph database implementation can facilitate the detection of such data quality errors.

Shortest Path Analysis

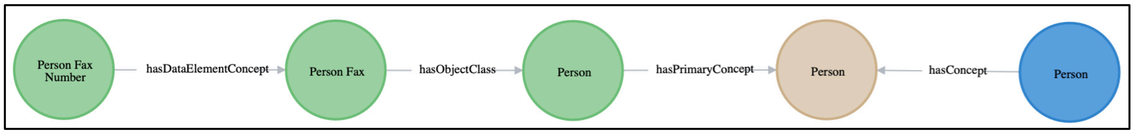
We hypothesized that the path from a CDE to a BRIDG class that had the shortest distance should match the BRIDG class to which the CDE was manually mapped by subject matter experts. We calculated the match rate as the percent of CDEs for which the shortest path led to the manually mapped BRIDG class. The match rate produced by the shortest path algorithm was 16.6% (280 out of 1,689 CDEs). This is much lower than the match rate produced by the ANN algorithm. The ANN algorithm produced a match rate of between 34 - 94%[85]. For the ANN algorithm, the lowest match rate of 34% corresponded to a set of CDEs that were semantically different from the training set and contained many CDEs that were manually mapped to BRIDG classes for which there was insufficient training data. To determine why the algorithm was returning such a low match rate, we looked at those CDEs that had a path distance of 4 and a path that went through the Object Class's primary concept.

Analysis of Paths with a Distance of Four

A path with a distance of 4 indicates that the CDE and the BRIDG class directly share an NCI Thesaurus concept. Since the Object Class is equivalent to a UML class[87], a path from a CDE to a BRIDG class that goes through the Object Class's primary concept and has a distance of 4 should represent a correct match. Figure 12 shows such a path.

Figure 12. Example path with a distance of 4.

Example shortest path with the BRIDG portion in blue, NCI Thesaurus portion in brown, and the caDSR CDEs in green.



64 CDEs had a path to a BRIDG class that had a distance of four and went through the Object Class primary concept. However, the match rate for these CDEs was only 33%. A closer analysis of the incorrectly matched CDEs shows that the mappings were more complex than the shortest path results initially indicate. For the majority of the CDEs analyzed, they were manually mapped to a more specific BRIDG class. For example, the shortest path algorithm associated with the CDE Product Collection Date with the BRIDG class Product. However, the CDE was manually mapped to the BRIDG class PerformedSubstanceExtraction. The manually mapped BRIDG class is correct. A subject matter expert is able to evaluate the meaning of the entire CDE, while the shortest path algorithm focuses on one concept. Table 2 presents some example CDEs along with the manually mapped BRIDG class, predicted BRIDG class, and a comment explaining the discrepancy. Even this analysis is subjective. Another subject matter expert may have a different interpretation of the mappings.

Table 6. Example of discrepancies between manually mapped and predicted BRIDG classes.

CDE Long Name	Manually Mapped BRIDG Class	Predicted BRIDG Class	Comment
Person Fax Number	StudySitePersonnel	Person	Manual mapping is technically incorrect because it introduces semantics not contained within the CDE.
Cellular Therapy Product Unique Identifier	Subject	Product	Manual mapping is technically incorrect because it introduces semantics not contained within the CDE.
Product Batch Number Unique Identifier	ProcessedProduct	Product	Manual mapping is potentially correct; ProcessedProduct has an association with Product.
Person Sex Type	BiologicEntity	Person	Manual mapping is correct; BRIDG is designed to accommodate non-human subjects and CDEs are generally created from a human perspective.
Product Tissue Donor Relationship Type	SubstanceExtraction AdministrationRelationship	Product	Manual mapping is correct; the CDE was mapped to a more specific BRIDG class that better captured the semantics.

Discussion

Reasons for the Low Match Rate

The low match rate returned by the shortest path algorithm was unexpected but brought to light the subjective nature of mapping and data quality issues. Defining metadata and mapping data standards is a subjective process. Each individual involved in the process has use cases and perspectives that influence their work. In many ways, the BRIDG model, the caDSR CDEs, and the NCI Thesaurus are like the parable of the three blind sages trying to describe an elephant. Each is describing the same thing but doing it from a slightly different perspective. For example, the CDEs in the caDSR are primarily created to capture patient-related data such as outcomes data. As a result, they are patient-focused. The BRIDG model was developed to represent both pre-clinical and clinical data[60]. As such, it needs to capture information about non-human research subjects. The CDE Person Birth Date illustrates this situation well. The CDE uses the concept for Person, which is also used by the BRIDG class Person. This resulted in a shortest path with a distance of four. However, the BRIDG model reserves the Person class for human-specific attributes such as occupation. Attributes such as date of birth and gender are in the class BiologicEntity. The path between the CDE Person Date of Birth and the BRIDG class of BiologicEntity has a distance of five. The CDE, the manual mapping to the BRIDG class BiologicEntity, and the associated concepts are fundamentally correct. However, the shortest path algorithm is purely objective. It looks at the underlying data and matches the CDE to the BRIDG class of Person.

A similar difference of perspective is seen with how concepts are defined in the NCI Thesaurus and used by BRIDG and the CDEs. This is exemplified by CDE Surgical Procedure Performed Date, which was manually mapped to the BRIDG class "PerformedProcedure." The mapping distance between the CDE and BRIDG class is seven. The reason for the long mapping distance lies with how the underlying concepts are defined. The CDE uses the concept "Surgical Procedure," which has a semantic type of "Health Care Activity" and a root concept of "Activity." The BRIDG class uses the concept "PerformedProcedure," which has a semantic type of "Research Activity" and a root concept of "Conceptual Entity." The representations of the two concepts in the NCI Thesaurus are different but technically correct. A surgical procedure can be both a health care activity and a research activity. Unfortunately, the different representations have created semantic silos, which the shortest path algorithm struggles to overcome.

The results of the shortest path analysis also revealed potential data quality issues with both the choice of BRIDG classes selected during the manual mapping process and with how the CDEs were defined. For example, the algorithm found a path with a distance of four between the CDE Person Fax Number and the BRIDG class Person. During the manual mapping process, the CDE was mapped to the BRIDG class StudySitePersonnel. The mapping is technically incorrect because it introduces information not contained in the semantic definition of the CDE. Most likely, the individual who performed the manual mapping based the mapping on the context in which the CDE was used and not strictly on the semantics of the CDE. Manually mapping the CDE to StudySitePersonnel limits the reusability of the mapping to other contexts.

The shortest path algorithm also highlighted instances where the CDE was constructed using incorrect concepts. For example, the CDE Disease Involvement Site Other Specify has a path length of seven to the manually mapped BRIDG class of TargetAnatomicSite. Analysis of the CDE revealed that it had incorrectly used the generic NCI Thesaurus concept for Location instead of the correct concept for Anatomic Site. This error made it difficult for the shortest path algorithm to find the path to the correct BRIDG path.

Limitations

There are several limitations to our work. First, we used an unweighted implementation of the shortest path algorithm. According to caDSR best practice, the bulk of a CDE's semantic meaning should be contained in the Data Element Concept[67, 87]. Adjusting the algorithm's parameters so that paths going through the Value Domain have a higher weight, may have produced a better match rate. Second, the algorithm only considers one concept. The complete meaning of a CDE cannot be understood by looking at one concept in isolation. This limits the ability of the algorithm to predict an appropriate BRIDG class. Finally, 12% of the BRIDG classes were not associated with an NCI Thesaurus concept. While this had little impact on this project, to extend this work to other use cases, this gap should be closed.

Next Steps

We plan to enhance to shortest path algorithm implementation to include weighting and to better handle relationships between BRIDG classes. The weighting will prioritize those paths that go through the Data Element Concept which should contain the bulk of the semantic meaning for the CDE[67, 87]. Also, sometimes the manually mapped class was either a descendent of or associated with the predicted BRIDG class. Adjusting the algorithm to include such classes along with the BRIDG class associated with the shortest path may improve the effectiveness of using a graph database to semi-automate the mapping process.

Next, we plan to explore combining the Artificial Neural Network (ANN) algorithm for facilitating the mapping of CDEs to the BRIDG model with the Neo4j graph database implementation of the BRIDG model, caDSR CDEs, and the NCI Thesaurus. Since the ANN algorithm learns from previous CDE to BRIDG mappings, it can handle the subjective nature of the mapping process. The performance of the algorithm diminishes when it has insufficient training data. The performance of the algorithm may improve if it can leverage the underlying semantics of both the CDEs and the BRIDG classes. In particular, performance may improve if the ANN algorithm can incorporate synonyms. In contrast, the Neo4j graph database and shortest path algorithm employ an objective, logical approach to mapping. It looks strictly at the semantic relationships between BRIDG classes, caDSR CDEs, and NCI Thesaurus concepts. It does not consider the context, nor does it learn from previous mappings. Combining the ANN algorithm with the graph database may result in better mapping predictions, especially when there is insufficient training data.

The incorporation of the BRIDG model, caDSR CDEs, and NCI Thesaurus concepts into one graph database creates a framework to perform robust quality assurance testing. Our analysis of the shortest path algorithm results revealed instances of incorrect mappings and poor CDE construction. We plan to develop a collection of Cypher queries to determine potential quality issues.

Another area to explore is expanding the NCI Thesaurus so that it can accommodate multiple perspectives. For example, annotating a concept such as PerformedProcedure as being both a Health Care Activity and a Research Activity will increase its interoperability when used in different contexts and use cases.

Conclusion

The process of manually defining and mapping clinical data standards combines logical, objective reasoning, along with subjective characteristics informed by the individual’s experience and the particular use case. As a result, a more objective mapping approach, such as the graph-based shortest path algorithm, will not be able to replicate the manual mapping results perfectly. An approach, such as an artificial neural network-based algorithm that learns from previous manual mappings, is better able to replicate the manually mapping results but is limited by lack of training data and semantic annotations. Combining the rich semantics contained in a graph database along with the learning capabilities of an artificial neural network may provide for a more robust mapping strategy.

The graph database that incorporates the BRIDG model, caDSR CDEs, and the NCI Thesaurus provides a valuable source of semantic annotations that can be leveraged for a variety of purpose. In particular, the graph database has the potential to facilitate quality assurance testing.

Chapter 5: Conclusion

Robinette Renner

Abstract

The semi-automated mapping tool based on an artificial neural network leveraged pattern recognition to predict CDE to BRIDG class mappings with between 94 – 34% accuracy. It was limited by a lack of training data for all BRIDG classes. It is possible to represent the caDSR CDEs, BRIDG model, and NCI Thesaurus using a graph database and then to use a shortest path algorithm to predict CDE to BRIDG class mappings. The approach is limited by the subjective nature of the mapping process. An optimal mapping strategy guides a subject matter expert through the process by providing artificial neural network and graph database predictions along with graph database visualization capabilities.

Summary

Semi-automated mapping tools can potentially reduce the mapping burden resulting from the need to support multiple clinical data standards. My research has shown that there is no single semi-automated mapping technique that can best address all needs in this mapping endeavor. Instead, any semi-automated mapping tool should combine multiple mapping techniques in order to gain the collective benefits and to reduce their limitations.

The goal of my research was to do the foundational work upon which a future semi-automated mapping tool could be built. The use case was mapping cancer study Common Data Elements (CDEs) from the National Cancer Institute's (NCI) cancer Data Standards Registry and Repository (caDSR) to the Biomedical Research Integrated Domain Group (BRIDG) model. Through the Food and Drug Administration's (FDA) Common Data Model Harmonization project (CDMH) [23], mapping CDEs to the BRIDG model can facilitate the mapping to other common data models such as Sentinel and PCORnet. My research developed and analyzed two semi-automated mapping techniques: an artificial neural network algorithm and a graph database.

The artificial neural network and the graph database semi-automated mapping techniques developed in this research have different strengths and limitations. The artificial neural network's strength is pattern recognition. As a result, it performs best with a diverse set of known, high-quality mappings. Without these known mappings, the artificial neural network is not able to predict CDE to BRIDG class mappings. The graph database's strength is finding logical, but sometimes subtle, relationships between CDEs and BRIDG classes. As a result, it is not able to incorporate the subjective aspects of the mapping process. These strengths are inherent in the design of each approach and are evident in the research findings.

Artificial Neural Network

The artificial neural network algorithm was designed to determine potential CDE to BRIDG class mappings based on how similar a CDE of interest is to training CDEs with known BRIDG class mappings. The design framework consists of two key aspects: pattern recognition and training data. The experimental results show that the success of the algorithm depends on how similar the CDE of interest is to the CDEs in the training set and the availability of training data. As the similarity of the CDE to the training CDEs decreases, so does the accuracy of the algorithm.

Design

Pattern recognition and training data are two design factors that determine those use cases for which an artificial neural network mapping approach will be most effective. The artificial neural network algorithm focuses strictly on pattern recognition. The similarity between two CDEs is calculated using the edit distance needed to transform the words in one CDE's attributes to those of the other CDE. It does not consider other factors such as synonyms or related concepts.

The artificial neural network bases its comparisons on training data comprised of known CDE to BRIDG class mappings. The algorithm replicates these mapping patterns regardless of their accuracy. Therefore, any mistakes made in the past will be perpetuated going forward. Also, the algorithm is not able to produce novel mappings. If no CDEs have ever been mapped to a BRIDG class, it is not able to predict mappings to that class. It is as if the class does not exist. The artificial neural network algorithm depends on a diverse set of known, high-quality mappings.

Findings

The performance of the artificial neural network algorithm was tested from two perspectives: the similarity of the testing CDEs to the training CDEs and the availability of training data. Overall, the artificial neural network algorithm had match rates of 94-34%. This range of match rates highlights those use cases in which the artificial neural network is best suited.

Three testing sets, Semantically Similar, Moderately Different, and Semantically Different, were used to analyze the performance of the algorithm. The percent similarity of the testing sets compared to the training set was 94%, 58%, and 5%. The match rate for the three testing sets was 87%, 69%, and 34% for the Semantically Similar, Moderately Different, and Semantically Different testing sets, respectively. As the similarity of the testing set to the training set decreased, the match rate decreased.

The ability of the artificial neural network algorithm to predict CDE to BRIDG class data is dependent upon the availability of training data. If there is not sufficient training data for a particular BRIDG class, then the algorithm will never predict a mapping to that class. It is as if the BRIDG class does not exist. The experiment results highlight this. The match rate increased when the experiments were run using testing sets that contained only CDEs that mapped to BRIDG classes for which there was sufficient training data. In this scenario, the match rates were 94%, 80%, and 70% for the Semantically Similar, Moderately Different, and Semantically Different testing sets, respectively. The largest increase match rate was for the Semantically Different testing set. For this testing set, the match rate increased from 34% to 70% when it was run using CDEs that mapped to BRIDG classes for which the algorithm had enough information to perform the mapping. The training set used for the experiments had sufficient examples for just 19 of the 233 classes in version 3.2 of the BRIDG model. While the size of the training set will gradually increase as the algorithm is used to map novel CDEs, the need for a diverse set of high-quality training examples limits the scalability of the algorithm.

Graph Database

The graph database mapping technique was designed to find relationships between CDEs and BRIDG classes based on their common usage of concepts found in the NCI Thesaurus. The design of the graph database approach is based on the usage of the NCI Thesaurus as the common foundation upon which CDEs and BRIDG classes are defined. The experimental results show that the performance of the graph database mapping technique depends upon how the NCI Thesaurus concepts, CDEs, and BRIDG classes are defined.

Design

Both CDEs and BRIDG classes are defined using NCI Thesaurus concepts. As a result, the NCI Thesaurus concepts can serve as a common connection point. In addition, CDEs, BRIDG classes, and NCI Thesaurus concepts all consist of numerous internal relationships. For example, BRIDG classes can have children, numerous NCI Thesaurus concepts can have the same semantic type, and CDEs are constructed from components such as Data Element Concepts, which are used by many CDEs. These relationships serve as the key design construct for the graph database mapping technique.

The graph database mapping technique traverses the web of relationships using Neo4j's Shortest Path algorithm. The algorithm finds the shortest path between the CDE of interest and all BRIDG classes. It then returns the BRIDG classes that are the closest to the CDE of interest. The algorithm simply traverses these existing relationships. The relationships themselves are manually created by a subject matter expert when the CDE, BRIDG class, or NCI Thesaurus concept is defined. As a result, the performance of the algorithm is dependent upon how a person defines the underlying components.

Findings

The experiment results show the dependence upon how the subject matter experts have defined the CDEs, BRIDG classes, and NCI Thesaurus. The overall match rate for the graph database mapping technique was only 16.6%. For those instances in which the CDE and the known, manually-mapped BRIDG class directly share a concept, the match rate only increased to 33%. The unexpectedly low match rates made the results more interesting than if the match rates were higher.

Closer analysis of those instances in which the Shortest Path algorithm was not able to correctly predict the appropriate BRIDG classes showed that the reason for the low match rate was not a technical issue with the approach. Instead, it was due to the following: the subjective nature of mapping; different perspectives of the subject matter experts defining the CDEs, BRIDG class, and NCI Thesaurus concepts; and data quality issues. For example, if the NCI Thesaurus concepts, CDEs, and BRIDG classes are defined in a way that results in semantic silos, then the graph database shortest path algorithm is not able to find relevant relationships.

Use Case Analysis

The artificial neural network algorithm uses patterns found in existing mappings to predict future mappings. It does not care if those mappings are correct. It simply replicates the past. As a result, it performs well when there are diverse, high-quality training data that is semantically similar to the CDEs being mapped. The graph database uses relationships between and amongst BRIDG classes, CDEs, and NCI Thesaurus concepts to predict mappings. It depends strictly on how each of those components has been defined and not on how mappings were done in the past. It performs best when the BRIDG classes and CDEs directly share the same concept.

Unfortunately, mapping is rarely a simple process for which one particular mapping technique will be the best. Nor will a fully automated mapping technique be the most effective (no matter how much I want a “magic button” to exist). The best mapping approach will most likely be one that uses multiple techniques to guide a subject matter expert through the mapping process by presenting them with data and visualization tools that enhance their existing knowledge and experience.

Next Steps

This research represents just the start of an interesting area for further exploration. In many ways, it created more questions than it answered. The future work can be broadly categorized as either analysis or code enhancement.

Analysis

The future analysis work consists of additional questions for existing data sets and the analysis of new data sets. First, the existing data set could be subsetted based on the organization that created the content. The data set used to test the graph database approach was not subsetted based on similarity to the training data, as was done for the artificial neural network approach. Since the artificial neural network relied on training data, the testing was conducted on the three separate data sets with different semantic similarity to the training data. The graph database approach, on the other hand, does not need training data. Therefore, the data set used to test the graph database approach consisted of an aggregation of the three testing data sets used in the artificial neural network approach along with the training data. It was assumed that the separation based on similarity to the training data was not relevant to the graph database approach. While that is true, in addition to each data set having different semantic similarity to the training data, the CDEs in the Semantically Different data set were primarily created by curators from the National Cancer Institute. The CDEs in the other data sets were primarily created by curators from the Center for International Blood and Marrow Transplant Research. Different organizations tend to create CDEs for different clinical areas and to follow different best practices. It would be interesting to analyze the effectiveness of the graph database approach using CDEs subsetted based on the organization that created the CDEs.

Second, analysis can be performed on an additional data set. After both the artificial neural network and graph database experiments were completed, another data set was found. In 2010 the NCI created more than 1,800 CDEs to represent the semantics of each BRIDG attribute in version 3.2 of the BRIDG model. These CDEs are annotated with the BRIDG class and attribute to which they map. Since these CDEs were specifically created to represent the BRIDG semantics, they are an excellent test case for both the artificial neural network algorithm and the graph database.

Code Enhancements

The code enhancements represent two main bodies of work: refinement of the technical implementations and the development of a mapping tool that guides the subject matter expert through the mapping process.

Artificial Neural Network

The primary enhancement needed for the artificial neural network algorithm is to incorporate semantics. The current pattern-based algorithm does not consider synonyms or semantic types. Semantic type is particularly important because it allows for the categorization of related concepts. For example, a CDE for white blood cell count value should be mapped to the same BRIDG class as one for platelet count value because both are laboratory procedures. Incorporating semantic information into the artificial neural network algorithm will facilitate analysis.

Graph Database

The graph database work completed to date has been a simple implementation of the standard shortest path algorithm included in the Neo4j database. The work can be improved in several ways such as decreasing the overall runtime, extending the data model to include additional information, and using algorithmic approaches, such as including known CDE to BRIDG class mappings and a weighted shortest path algorithm.

First, the run time of the shortest path algorithm could be approved. In order to find the BRIDG class with the shortest path to a CDE, the Python script found the shortest path from all BRIDG classes to a CDE of interest. Iterating through all of the classes takes a considerable amount of time. Running on the shortest path algorithm on 100 CDEs takes nearly four minutes. Reducing the runtime of the algorithm will make it more practical for bulk mapping applications.

Second, for those CDE to BRIDG class mappings that have been verified correct by a subject matter expert, adding relationships from a CDE directly to a BRIDG class could improve the match rate. For example, the CDE for Person Address Postal Code [94], uses the Data Element Concept for Person Address and maps to the BRIDG class "Person." Thirty-nine CDEs use the Data Element Concept Person Address. In theory, all thirty-nine of the CDEs should map to the BRIDG class "Person." In this instance, if the direct CDE to BRIDG class mapping were added to the data model, any CDE using that Data Element Concept would have a shortest path mapping of three.

Third, this work used a simple unweighted implementation of the shortest path algorithm. According to caDSR best practices, the Data Element Concept should contain the bulk of the semantic meaning for the CDE [67, 87]. Also, according to the ISO 11179 metamodel, the Object Class is equivalent to a UML class [87]. Therefore, those shortest paths that go through the Object Class are more likely to be correct than those shortest paths that go through the Representation Term, which is associated with the Value Domain. Using a weighted implementation of the shortest path algorithm, which assigns a lesser weight to shortest paths that go through the Object Class, may yield better results.

Assisted Mapping Tool

Manual mapping is labor-intensive and error-prone [29]. Mapping techniques such as artificial neural networks and graph databases can facilitate the mapping process by revealing patterns and relationships that a subject matter expert may not recognize. However, these mapping techniques cannot be used in isolation. Mapping is inherently a subjective process. A fully automated programmatic solution will not capture the context and nuance needed to produce a correct mapping. A semi-automated mapping tool that leverages the expertise and experience of the subject matter expert while providing them with additional information to guide them along the mapping process will help bridge the gap between a manual mapping approach and a fully automated one.

I propose developing a mapping tool that provides the subject matter expert with additional information and visualization tools that guide them along the mapping process. The mapping workflow would start with the graph database shortest path algorithm. For the CDE of interest, it would display a prioritized list of potential class mappings based on the mapping distance and the nodes traversed. The subject matter expert would be able to select the desired class, visualize the child and associated classes using the graph database, or run the artificial neural network algorithm. After running the artificial neural network algorithm, the subject matter expert would be able to select the desired class or explore the child and associated classes using the graph database.

Conclusion

Mapping tools can reduce the burden of manually mapping between different standards while ensuring the quality of the mappings. However, while mapping tools can assist the subject matter expert performing the mappings, they can never fully replace the subject matter expert. Mapping is inherently a subjective process. In addition, the computational mapping techniques have different strengths and limitations. The subject matter expert can use their knowledge and experience to process the results of the different mapping techniques and then determine the best overall mapping. The goal of mapping tooling should be to facilitate the work of a subject matter expert and not to replace them.

References

1. Jessop M, Weeks M, Austin J. **CARMEN: a practical approach to metadata management**. *Philos Trans R Soc A-Math Phys Eng Sci*. 2010;368(1926):4147-59.

2. John RJL, Potti N, Patel JM, editors. Ava: From Data to Insights Through Conversations. CIDR; 2017.

3. CIBMTR Progress Report 2014. 2014 January 2015.

4. Musen MA, Lewis S, Smith B. **Wrestling with SUMO and bio-ontologies**. *Nat Biotechnol*. 2006;24(1):21; author reply 3.

5. Aljurf M, Rizzo JD, Mohty M, Hussain F, Madrigal A, Pasquini MC, et al. **Challenges and opportunities for HSCT outcome registries: perspective from international HSCT registries experts**. *Bone marrow transplantation*. 2014;49(8):1016-21.

6. Parimalam T, Deepa R, Devi RN, Devi PY. **Detecting Duplicate Records-A Case Study**. 2015.

7. Annex A. Information technology—Metadata registries (MDR)—Part 1: Framework. 2013.

8. Komatsoulis GA, Warzel DB, Hartel FW, Shanbhag K, Chilukuri R, Fragoso G, et al. **caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability**. *Journal of biomedical informatics*. 2008;41(1):106-23.

9. Warzel DB, Andonaydis C, McCurry B, Chilukuri R, Ishmukhamedov S, Covitz P. **Common data element (CDE) management and deployment in clinical trials**.

AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2003:1048.

10. de Coronado S, Wright LW, Fragoso G, Haber MW, Hahn-Dantona EA, Hartel FW, et al. **The NCI Thesaurus quality assurance life cycle**. *Journal of Biomedical Informatics*. 2009;42(3):530-9.

11. Jiang G, Solbrig HR, Chute CG. **Quality evaluation of cancer study Common Data Elements using the UMLS Semantic Network**. *Journal of biomedical informatics*. 2011;44:S78-S85.

12. Adamusiak T, Bodenreider O. **Quality assurance in LOINC using description logic**. *AMIA Annual Symposium Proceedings*. 2012;2012:1099.

13. Elhanan G, Perl Y, Geller J. **A survey of SNOMED CT direct users, 2010: impressions and preferences regarding content and quality**. *Journal of the American Medical Informatics Association*. 2011:amiajnl-2011.

14. Jiang G, Solbrig HR, Chute CG. **Quality evaluation of value sets from cancer study common data elements using the UMLS semantic groups**. *Journal of the American Medical Informatics Association : JAMIA*. 2012;19(e1):e129-36.

15. Mougin F, Bodenreider O. **Auditing the NCI thesaurus with semantic web technologies**. *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium*. 2008:500-4.

16. Tenenbaum JD, Whetzel PL, Anderson K, Borromeo CD, Dinov ID, Gabriel D, et al. **The Biomedical Resource Ontology (BRO) to enable resource discovery in clinical and translational research**. *J Biomed Inform*. 2011;44(1):137-45.

17. Lin C-H, Wu N-Y, Liou D-M. **A multi-technique approach to bridge electronic case report form design and data standard adoption**. *Journal of biomedical informatics*. 2014.

18. An Y, Borgida A, Mylopoulos J. **Refining Semantic Mappings from Relational Tables to Ontologies**. In: Bussler C, Tannen V, Fundulaki I, editors. Semantic Web and Databases. Lecture Notes in Computer Science. 3372: Springer Berlin Heidelberg; 2005. p. 84-90.

19. An Y, Mylopoulos J, Borgida A, editors. Building semantic mappings from databases to ontologies. Proceedings of the National Conference on Artificial Intelligence; 2006: Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.1557.

20. Williams RL, Johnson SB, Greene SM, Larson EB, Green LA, Morris A, et al. **Signposts along the NIH roadmap for reengineering clinical research: lessons from the Clinical Research Networks initiative**. *Archives of internal medicine*. 2008;168(17):1919-25.

21. Klumpp T, Beadle D, DeGregorio N, VanKuren N, Neff J, Bonaccorso C, et al. **Design and Implementation of a Multipurpose Hematopoietic Stem Cell Information System Based on the Biomedical Research Integrated Domain Model**. *Biology of Blood and Marrow Transplantation*. 2019;25(3):S267.

22. Becnel LB, Hastak S, Ver Hoef W, Milius RP, Slack M, Wold D, et al. **BRIDG: a domain information model for translational and clinical protocol-driven research**. *Journal of the American Medical Informatics Association*. 2017;24(5):882-90.

23. **Common Data Models Harmonization FHIR Implementation Guide** [Internet]. HL7 International - Biomedical Research and Regulation Work Group; 2019 [2019-08-14]. Available from: https://build.fhir.org/ig/HL7/cdmh/.

24. **HL7 FHIR** [Internet]. Health Level Seven International; 2018 [2018 Mar 6]. Available from: http://hl7.org/fhir.

25. **FDA's Sentinel Initiative** [Internet]. U.S. Food & Drug Administration; 2018 [2019-08-14]. Available from: https://www.fda.gov/safety/fdas-sentinel-initiative.

26. **The National Patient-Centered Clinical Research Network (PCORnet)** [Internet]. [2019-08-14]. Available from: https://pcornet.org.

27. **Informatics for Integrating Biology and the Bedside (i2b2)** [Internet]. Partners Healthcare; 2019 [2019-08-14]. Available from: https://www.i2b2.org.

28. **Observational Health Data Sciences and Informatics (OHDSI)** [Internet]. 2019 [2019-08-14]. Available from: https://www.ohdsi.org.

29. Richesson RL, Fung KW, Krischer JP. **Heterogeneous but "standard" coding systems for adverse events: Issues in achieving interoperability between apples and oranges**. *Contemporary Clinical Trials*. 2008;29(5):635-45.

30. Richesson RL, Krischer J. **Data standards in clinical research: Gaps, overlaps, challenges and future directions**. *Journal of the American Medical Informatics Association*. 2007;14(6):687-96.

31. Doan A, Madhavan J, Dhamankar R, Domingos P, Halevy A. **Learning to match ontologies on the semantic web**. *The VLDB Journal*. 2003;12(4):303-19.

32. Zheng L, Chen Y, Perl Y, Halper M, Geller J, De Coronado S, editors. Quality Assurance of Concept Roles in the National Cancer Institute thesaurus. 2018 IEEE

International Conference on Bioinformatics and Biomedicine (BIBM); 2018: IEEE.2001-8.

33. Wang Y, Patrick J, Miller G, O'Hallaran J, editors. A computational linguistics motivated mapping of ICPC-2 PLUS to SNOMED CT. BMC medical informatics and decision making; 2008: BioMed Central.S5.

34. Mao M, Peng Y, Spring M. **An adaptive ontology mapping approach with neural network based constraint satisfaction**. *Web Semantics: Science, Services and Agents on the World Wide Web*. 2010;8(1):14-25.

35. Noy NF, Musen MA, editors. Algorithm and tool for automated ontology merging and alignment. Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00) Available as SMI technical report SMI-2000-0831; 2000.

36. Noy NF, Musen MA, editors. Anchor-PROMPT: Using Non-Local Context for Semantic Matching. OIS@ IJCAI; 2001.

37. Chortaras A, Stamou G, Stafylopatis A, editors. Learning ontology alignments using recursive neural networks. International Conference on Artificial Neural Networks; 2005: Springer.811-6.

38. Rubiolo M, Caliusco ML, Stegmayer G, Coronel M, Fabrizi MG. **Knowledge discovery through ontology matching: An approach based on an Artificial Neural Network model**. *Inf Sci*. 2012;194:107-19.

39. University P. **About WordNet**: Princeton University; 2010 [cited 2020 2020-01-05]. Available from: https://wordnet.princeton.edu.

40. Huang J, Dang J, Vidal JM, Huhns MN, editors. Ontology matching using an artificial neural network to learn weights. IJCAI workshop on semantic Web for collaborative knowledge acquisition; 2007.

41. Huang J, Dang J, Huhns MN, Zheng WJ. **Use artificial neural network to align biological ontologies**. *BMC genomics*. 2008;9(2):S16.

42. Shvaiko P, Euzenat J. **Ontology matching: state of the art and future challenges**. *Knowledge and Data Engineering, IEEE Transactions on*. 2013;25(1):158-76.

43. **Scopus** [Internet]. [2019-03-23]. Available from: https://www.scopus.com/search/form.uri?display=basic.

44. Komatsoulis GA, Warzel DB, Hartel FW, Shanbhag K, Chilukuri R, Fragoso G, et al. **caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability**. *Journal of Biomedical Informatics*. 2008;41(1):106-23.

45. Robinson I, Webber J, Eifrem E. **Graph databases**. Sebastopol (CA): O'Reilly Media, Inc.; 2013.

46. Alqahtani A, Heckel R. Model based development of data integration in graph databases using triple graph grammars. In: Mazzara M, Salaun G, Ober I, editors.: Springer Verlag; 2018. p. 399-414.

47. Johnson D, Connor AJ, McKeever S, Wang Z, Deisboeck TS, Quaiser T, et al. **Semantically linking in silico cancer models**. *Cancer informatics*. 2014;13(Suppl 1):133-43.

48. Van Mulligen EM, Vlietstra WJ, Vos R, Kors J, editors. Discovering information from an integrated graph database. Discovering Information from an Integrated Graph Database; 2016.

49. Vlietstra WJ, Zielman R, van Dongen RM, Schultes EA, Wiesman F, Vos R, et al. **Automated extraction of potential migraine biomarkers using a semantic graph**. 2017;71:178-89.

50. Xiang Y, Lu K, James SL, Borlawsky TB, Huang K, Payne PRJJoBI. **k-Neighborhood decentralization: a comprehensive solution to index the UMLS for large scale knowledge discovery**. 2012;45(2):323-36.

51. Campbell WS, Pedersen J, McClay JC, Rao P, Bastola D, Campbell JR. **An alternative database approach for management of SNOMED CT and improved patient data queries**. *Journal of biomedical informatics*. 2015;57:350-7.

52. Ulrich H, Kock-Schoppenhauer AK, Duhm-Harbeck P, Ingenerf J. **Using Graph Tools on Metadata Repositories**. *Stud Health Technol Inform*. 2018;253:55-9.

53. Pollack J. **BRIDGModel2Graph** [Internet]. GitHub; 2019 [2019-08-14]. Available from: https://github.com/nmdp-bioinformatics/BRIDGModel2Graph.

54. Bodenreider O. **The unified medical language system (UMLS): integrating biomedical terminology**. *Nucleic acids research*. 2004;32(suppl 1):D267-D70.

55. **UMLS Statistics - 2019AB Release**: U.S. National Library of Medicine; 2019 [cited 2020 2020-01-05]. Available from: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html.

56. Ulrich H, Kock AK, Duhm-Harbeck P, Habermann JK, Ingenerf J. **Metadata Repository for Improved Data Sharing and Reuse Based on HL7 FHIR**. *Exploring Complexity in Health: an Interdisciplinary Systems Approach*. 2016;228:162-6.

57. Wetherall ASTDJ. **Computer Networks, Fifth Edition**: Prentice Hall; 2011.

58. **CIBMTR Progress Report 2017** [Internet]. The Medical College of Wisconsin, Inc. and the National Marrow Donor Program; 2017 [2019-08-14]. Available from: http://www.cibmtr.org/About/AdminReports/Pages/index.aspx.

59. Renner R, Carlis J, Maiers M, Rizzo JD, O'Neill C, Horowitz M, et al. **Integration of Hematopoietic Cell Transplantation Outcomes Data**. 2015;9162:139-46.

60. **Biomedical Research Integrated Domain Group** [Internet]. [2019-08-14]. Available from: https://bridgmodel.nci.nih.gov.

61. **Study Data Standards: What you need to know** [Internet]. US Food and Drug Administration; 2017 [2019-05-28]. Available from: https://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM511237.pdf.

62. Noy N, SMART MM, editors. Automated Support for Ontology Merging and Alignment. Twelfth Banff Workshop on Knowledge Acquisition, Modeling, and Management-Banff, Alberta, Canada; 1999.

63. Do H-H, Rahm E, editors. COMA: a system for flexible combination of schema matching approaches. Proceedings of the 28th international conference on Very Large Data Bases; 2002: VLDB Endowment.610-21.

64. He B, Chang KC-C, editors. Statistical schema matching across web query interfaces. Proceedings of the 2003 ACM SIGMOD international conference on Management of data; 2003: ACM.217-28.

65. ISO 11179 Specification part 1 version 3. Switzerland: ISO/IEC 2015.

66. **CDE 2682630: Acute Myeloid Leukemia Classification Type** [Available from: https://cdebrowser.nci.nih.gov/cdebrowserClient/cdeBrowser.html#/search?publicId=2682630&version=4.0.

67. **caDSR Training Material - Course 1040 Creating Well-formed Metadata and Metadata Business Rules** [Internet]. National Cancer Institute; [2019-08-14]. Available from: https://wiki.nci.nih.gov/display/COREtraining/1040+Creating+Well-formed+Metadata+and+Metadata+Business+Rules.

68. Ver Hoef ea. **BRIDG 5.1 Model User's Guide** [Internet]. Biomedical Research Integrated Domain Group; [updated 2019-08-27. Available from: https://bridgmodel.nci.nih.gov/download-model/bridg-releases.

69. OMG® Unified Modeling Language® (OMG UML®) Version 2.5.1. 2017.

70. **CDE Browser** [Internet]. National Cancer Institute; [2018-08-31]. Available from: https://cdebrowser.nci.nih.gov/CDEBrowser/.

71. **CDE 2688790: Chronic Myelogenous Leukemia Classification Type** [Available from: https://cdebrowser.nci.nih.gov/cdebrowserClient/cdeBrowser.html#/search?publicId=2688790&version=2.0.

72. Mitchell T. **Machine Learning**: McGraw - Hill Science / Engineering / Mathh; 1997.

73. **CDE 2793029: Other Therapeutic Procedure Administered Indicator** [Available from: https://cdebrowser.nci.nih.gov/cdebrowserClient/cdeBrowser.html#/search?publicId=2793029&version=1.0.

74. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. **The FAIR Guiding Principles for scientific data management and stewardship**. *Scientific data*. 2016;3.

75. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu W-L, Wright LW. **NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information**. *Journal of biomedical informatics*. 2007;40(1):30-43.

76. Tang C, Plasek JM, Bates DW. **Rethinking Data Sharing at the Dawn of a Health Data Economy: A Viewpoint**. *Journal of medical Internet research*. 2018;20(11):e11519.

77. Piwowar HA, Becich MJ, Bilofsky H, Crowley RS. **Towards a data sharing culture: recommendations for leadership from academic health centers**. *PLoS medicine*. 2008;5(9):e183.

78. Federer LM, Lu YL, Joubert DJ, Welsh J, Brandys B. **Biomedical data sharing and reuse: Attitudes and practices of clinical and scientific research staff**. *PloS one*. 2015;10(6):e0129506.

79. Kush R, Goldman M. **Fostering responsible data sharing through standards**. *New England Journal of Medicine*. 2014;370(23):2163-5.

80. **Argonaut Project** [Internet]. Health Level Seven International; 2019 [2019-08-14]. Available from: http://argonautwiki.hl7.org.

81. **Introduction to the Interoperability Standards Advisory** [Internet]. Office of the National Coordinator for Health IT; 2019 [2019-08-14]. Available from: https://www.healthit.gov/isa/.

82. Kibbe W. **Cancer Clinical Research: Enhancing Data Liquidity and Data Altruism**. Oncology Informatics: Elsevier; 2016. p. 41-53.

83. **Clinical Data Interchange Standards Consortium** [Internet]. 2019 [2019-08-14]. Available from: https://www.cdisc.org.

84. Renner R, Li S, Huang Y, Tan S, Li D, v. d. Zijp-Tan A, et al., editors. Mapping Common Data Elements to a Domain Model Using an Artificial Neural Network. 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2018 3-6 Dec. 2018; Madrid.1532-5.

85. Renner RL, Shengyu; Huang,Yulong; van der Zijp-Tan, Ada Chaeli ; Tan, Shaobo; Li, Dongqi ; Kasukurthi, Mohan Vamsi; Benton, Ryan; Borchert, Glen; Huang, Jingshan; Jiang, Guoqian **Using an Artificial Neural Network to Map Cancer Common Data Elements to the Biomedical Research Integrated Domain Group Model in a Semi-automated Manner**. *BMC medical informatics and decision making*. 2019(in press).

86. **NCI Thesaurus** [Internet]. National Cancer Institute; [2019-08-14]. Available from: https://ncit.nci.nih.gov/ncitbrowser/.

87. ISO 11179 Specification part 3 version 3. Switzerlend: ISO/IEC; 2013.

88. **NCI Thesaurus Download** [Internet]. National Cancer Institute Enterprise Vocabulary Services; [2019-08-14]. Available from: https://evs.nci.nih.gov/evs-download/thesaurus-downloads.

89. Lamy J-B. **Owlready2 Documentation** [Internet]. [2019-08-05]. Available from: https://pythonhosted.org/Owlready2/.

90. Lamy J-B. **Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies**. *Artificial intelligence in medicine*. 2017;80:11-28.

91. McNamara J. **XlsxWriter** [Internet]. [Available from: https://xlsxwriter.readthedocs.io/index.html.

92. **Neo4j Bolt Driver 1.7** [Internet]. Neo Technology; 2018 [2019-08-09]. Available from: https://neo4j.com/docs/api/python-driver/current/.

93. **The Neo4j Graph Algorithms User Guide v3.5** [Internet]. Neo4j; 2019 [2019-08-27]. Available from: https://neo4j.com/docs/pdf/neo4j-graph-algorithms-3.5.pdf.

94. **CDE 316: Person Address Postal Code** [Available from: https://cdebrowser.nci.nih.gov/cdebrowserClient/cdeBrowser.html#/search?publicId=316&version=5.0.