TOWARD COMPLETE KNOWLEDGE OF HEALTHCARE DATASETS: EXTRACTION, MODELING, AND REPRESENTATION

by

Jingyi Shi

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2019

Approved by:

_____
Dr. Yaorong Ge

_____
Dr. Mirsad Hadzikadic

_____
Dr. Lixia Yao

_____
Dr. Jing Yang

ABSTRACT

JINGYI SHI. Toward complete knowledge of healthcare datasets: extraction, modeling, and representation. (Under the direction of DR. YAORONG GE)

In the era of big data, identifying the right dataset for analysis has been a severe challenge in data science. Especially in health data science, datasets are frequently complex and have restricted access, thus requiring sufficient time, energy, and background knowledge for users to understand, select, and begin analysis. The complexity largely toughens the development of health data science, and we believe it is important to make significant efforts to improve dataset identification processes. Recognizing the challenge, we believe that to provide complete knowledge of healthcare datasets would offer a solution that facilitates dataset identification to a great extent. As with a catalog of books in a library where people can find the desired book easily, with a complete knowledge of datasets, users are expected to quickly identify the most relevant and high-quality datasets for their research purposes. Toward this goal, we start with providing both content and quality level knowledge that is sufficiently comprehensive to cover the needs of a certain group of users—health data science novices. Specifically, we systematically examined the needs of the target users, extracted knowledge that was tailored to these needs, established quantifiable measurements for data quality (a Publication-based Popularity Index (PPI) and an Association-based intrinsic Quality Index (AQI)), and developed a healthcare Dataset Information Resource (DIR) framework to efficiently represent knowledge for datasets. The results from user studies indicate that the solution is promising.

This dissertation utilizes the three-article format, which includes six chapters. Aside from the introduction and the conclusion chapters, the middle four chapters represent four publications that contribute to the ultimate goal of complete knowledge, including both system and method developments.

ACKNOWLEDGEMENTS

I would first like to thank my doctoral advisor, Dr. Yaorong Ge. Thank you for your guidance and help over the years and for providing this important research topic for my dissertation. Words are not enough to express my gratitude. I feel so lucky to have you as my mentor.

I would like to acknowledge my dissertation committee members: Dr. Mirsad Hadzikadic, Dr. Lixia Yao, and Dr. Jing Yang. Thank you for playing an essential role in my dissertation process and for your time, suggestions, and support all along.

I would like to thank Dr. Lixia Yao again as one of my most important research collaborators. I am grateful to you for always caring about both my research and my life.

I would like to thank Dr. Lisa Russell-Pinson and Chris Harrington, who guided me a lot in dissertation writing. Thank you for your knowledge and help.

I would like to thank all my colleagues in the Health Informatics Lab for always standing by my side.

In addition, I would like to thank my family, especially my husband, who provided continuous support physically, mentally, and professionally.

Finally, I would like to acknowledge the funding opportunities that sponsored my research by thanking Dr. Ge for providing the financial support. Mainly, I would like to thank the Graduate Assistant Support Plan (GASP) at the University of North Carolina at Charlotte.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Overview

With the development of computing, storage, and network technologies, the era of big data has arrived and has brought both opportunities and challenges. Data science—leveraging statistics, computer science, and domain knowledge to extract insights from data—has benefited scientific research in a variety of fields. This is especially true for health data science, which generates data-driven solutions to complicated real-world health-related problems. An increasing number of data scientists have promoted better patient care and have helped save patients' lives by analyzing healthcare datasets. However, with both the number of and the complexity of healthcare datasets increasing, data scientists have encountered difficulty in acquiring adequate knowledge to assist a solid understanding of these datasets and to select the right dataset for data analysis. Unlike datasets in other disciplines, usually healthcare data are originally collected from a variety of devices, for diverse purposes, and with specific designs. This makes healthcare datasets even more complex than others. Therefore, we believe that identifying the right dataset for analysis has been a severe challenge in health data science and providing complete knowledge of healthcare datasets for data scientists should be a solution to the challenge and an ultimate goal of the health data science community. In progress toward this ultimate goal, data scientists can better take advantage of the existing data to solve problems.

Discovering the right dataset is essential to obtain the right insights. To select a right dataset and start data analysis, it requires significant time, energy, and fundamental knowledge to locate candidate datasets among numerous healthcare datasets, to learn the datasets, to realize their quality toward successful analysis, and to select

relevant features as well as proper analytical methods. Challenges are even more pronounced for beginning learners of data science. Becoming aware of these difficulties, the data science community has made efforts to integrate information from a wide range of dataset resources to facilitate dataset identification. Google released a Dataset Search [1] engine in 2018, which provides organized descriptions of datasets for all domains. For health-related data, a few platforms have been developed, including HealthData.gov [2] for government healthcare datasets, data.CDC.gov [3] for governmental disease control and prevention data, and DataMed.org [4] by the bioCADDIE project [5] for biomedical datasets and repositories. However, most of the current integration systems contain shallow information for healthcare datasets instead of knowledge, and none of them targets a specific user population. Without a knowledge-level representation and user-oriented design, the existing systems can hardly ease the difficult dataset identification process and answer sophisticated questions. Meanwhile, knowledge about the quality of healthcare datasets is rarely discussed and delivered in the context of dataset identification, even though it is critically essential toward successful data analysis. Groups of researchers have studied characters of Electronic Health Record (EHR) data quality for decades. For example, EHR data quality has been categorized into completeness, correctness, concordance, plausibility, and currency by [6] in 2013, harmonized into conformance, completeness, and plausibility by [7] in 2016, and classified into completeness, consistency, credibility, and timeliness by [8] in 2018. However, these EHR data quality studies have not reached a consensus on dimensions and definitions, while other types of healthcare data are seldom considered. Moreover, current studies discuss quality at a conceptual level but lack quantified measurements that are ready-to-use to assist a straight-forward dataset quality evaluation.

Recognizing these shortcomings, we are endeavoring to push toward the goal of complete knowledge of healthcare datasets by filling gaps in the field. Therefore,

we aim to provide knowledge about datasets that is sufficiently comprehensive to a targeted sub-group of data scientists. Particularly, we focus on the special knowledge, in both content and quality levels, needed by health data science novices and on developing effective methods to extract, to model, and to represent this knowledge of healthcare datasets.

This dissertation is structured in the three-article format, which includes six chapters. Aside from this introduction and the conclusion chapters, the middle four chapters represent four publications, including both system and method developments. Chapter 2 (paper 1) describes the development of a healthcare Dataset Information Resource (DIR) framework, which represents tailored knowledge for novices based on Semantic Web technologies [9] and enables the ability to answer sophisticated questions. Chapter 3 (paper 2) illustrates a quality measurement in an explicit perspective, that is a Publication-based Popularity Index (PPI) that can quantifiably evaluate the overall usefulness of a dataset. Chapter 4 (paper 3) proposes an entropic feature selection method that is designed specifically for health data challenges, and Chapter 5 (paper 4) describes another quality measurement (an Association-based intrinsic Quality Index (AQI)) in an implicit perspective based on the proposed feature selection method.

1.2 Overview of Paper 1: Development of the Healthcare Dataset Information Resource (DIR)

It is important for data scientists to have a good understanding of the availability of relevant datasets as well as the content, structure, and existing analyses of these datasets. While a number of efforts are underway to integrate the large amount and variety of datasets, there is a lack of information resources that focus on the specific learning needs of some targeted audiences. To address this gap, we have developed a semantic DIR framework to specifically address the challenges of entry-level data scientists in learning to identify, understand, and analyze major datasets with an

initial focus on healthcare. The DIR does not contain actual data from the datasets but aims to provide comprehensive knowledge about the datasets and their analyses.

The framework leverages Semantic Web technologies and the W3C Dataset Description Standard [10] for knowledge integration and representation and includes natural language processing (NLP)-based methods to enable knowledge extraction and question answering. The prototype DIR implementation comprises four major components—dataset metadata and related knowledge, search modules, question answering for frequently asked questions, and blogs. Furthermore, the DIR currently includes information on 12 commonly used large and complex healthcare datasets. The initial usage evaluation based on health informatics novices indicates that the DIR is helpful and beginner-friendly. Further development of both content and function levels is underway.

1.3 Overview of Paper 2: the Publication-Based Popularity Index (PPI)

Data are critical in this age of big data and machine learning. Due to their inherent complexity, health-related data are unique in that the datasets are usually acquired for specific purposes and with special designs. As an increasing number of healthcare datasets become available, of which many are public, choosing a quality dataset that is suitable for specific research inquiries is becoming a challenging question for health informatics researchers, especially the learners of this field. On the other hand, from the data provider's perspective, it is important to identify features of datasets that make some datasets more valuable than others in order to improve the design and acquisition of future datasets. To address these questions, we need to develop formal mechanisms to measure the goodness of datasets according to certain criteria.

In this study, we propose one way of measuring the value of healthcare datasets that is based on how often the datasets are used and reported by researchers, which we call the PPI. In this article, we describe the design of the PPI and discuss its properties. We demonstrate the utility of the PPI by ranking 14 representative healthcare

datasets and believe that the PPI can enable an overall ranking of all healthcare datasets; thus, it provides an important dimension to sort search results for dataset integration systems as well as a starting point for identifying and examining the design of the most valuable healthcare datasets so that features of these datasets can inform future designs.

1.4 Overview of Paper 3: the Proposed Entropic Feature Selection Method

Health data are generally complex in type and small in sample size. Such domain-specific challenges make it difficult to capture information reliably and to contribute further to the issue of generalization. To assist the analytics of healthcare datasets, we develop a feature selection method based on the concept of Coverage Adjusted Standardized Mutual Information (CASMI). The main advantages of the proposed method are: 1) it selects features more efficiently with the help of an improved entropy estimator, particularly when the sample size is small, and 2) it automatically learns the number of features to be selected based on the information from sample data. Additionally, the proposed method handles feature redundancy from the perspective of joint-distribution. This method focuses on non-ordinal data, while it works with numerical data with an appropriate binning method. A simulation study comparing the proposed method to six widely cited feature selection methods shows that the proposed method performs better when measured by the Information Recovery Ratio, particularly when the sample size is small. Moreover, the proposed method establishes the foundation of the AQI.

1.5 Overview of Paper 4: the Association-Based Intrinsic Quality Index (AQI)

As the number and source of health-related datasets continue to grow significantly, identification of datasets that are most appropriate for a research question is becoming ever more important for the field of health data analytics. The complexity of health-related data further exacerbates the challenge in dataset identification as it requires

significant efforts to understand a dataset before recognizing its appropriateness and quality to the research purpose. While the appropriateness of a dataset is largely a function of data semantics and research questions, we hypothesize that the quality of the dataset can be assessed by some intrinsic properties of the features in the dataset, and these properties are common across all datasets. Moreover, we believe that a good understanding of the usefulness of features in datasets is important to not only data analysts but also data providers because it will help them improve the design and acquisition of datasets in the future.

In this study, we propose one way of measuring the intrinsic quality of healthcare datasets that is based on the degree of association among attributes (features and outcomes) in a dataset, which we call the AQI. In this article, we describe the design of the AQI and discuss its properties. We demonstrate the utility of the AQI by a user study and results from two pairs of real healthcare datasets. We believe that the AQI can help assess the intrinsic quality of healthcare datasets and thus provide an important metric to assist dataset identification for researchers and a perspective for identifying and examining the design of the most valuable healthcare datasets so that features of these datasets can inform future designs. Furthermore, we argue that the AQI can also help researchers discover research opportunities within a given dataset.

CHAPTER 2: DEVELOPMENT OF A HEALTHCARE DATASET INFORMATION RESOURCE (DIR) BASED ON SEMANTIC WEB

2.1 Background

Healthcare data is rapidly growing in the era of big data. An increasing number of researchers are leveraging these datasets to improve the quality of patient care. However, challenges caused by a variety of purposes, designs, and techniques when health data were originally collected boost the complexity and diversity of healthcare datasets. For health data analysis, it requires significant time, energy, and fundamental knowledge to identify, understand, and choose the right datasets. The challenges for students and researchers who have little experience are even more pronounced. A number of online data resources, such as HealthData.gov [1], Data.CDC.gov [2], and Society of General Internal Medicine (SGIM) Research Dataset Compendium [3], integrate basic information for public datasets, which help new investigators choose datasets to a certain extent. However, the simple descriptions in these portals are hardly adequate for them to identify a suitable dataset to delve into. Simple search functions, such as a keywords search, provided by most of the resources cannot handle more complex and less concrete questions that typical novices have, such as finding existing analytical methods that are suitable for analyzing a particular dataset. Meanwhile, proprietary datasets, often having limited information in these portals, are even harder to understand and analyze.

Noticing these shortcomings, emerging research projects are attempting to build structured dataset information resources that address the challenge of dataset discovery and accessibility. For example, the Stanford University School of Medicine established the Center for Expanded Data Annotation and Retrieval (CEDAR) project in

2015 to facilitate researchers' standard use of metadata by developing an authoring-friendly computational ecosystem for metadata development, evaluation, use, and refinement [4]. By 2017, they had developed a CEDAR Workbench, which was an ontology-assisted tool to help scientific experiment metadata authoring [5]. Meanwhile, the University of California at San Diego is leading the development of a data discovery index system, the biomedical and healthCAre Data Discovery Index Ecosystem (bioCADDIE) [6], to index data that are stored elsewhere to facilitate data integration tasks that adopt content standards and high-level schema. A prototype biomedical data search engine, DataMed [7], under the bioCADDIE project, has included metadata extracted from multiple biomedical data repositories, such as the Cambridge Crystallographic Data Centre (CCDC) and U.S. National Center for Biotechnology Information (NCBI)'s BioProject. Similar to what PubMed (a free search engine that comprises more than 28 million citations from multiple literature databases and resources) has done for the biomedical literature, DataMed aims to make a comparable contribution for biomedical data.

However, the current attempts, focusing on integrating and searching datasets and dataset information, often lack consideration of the learning needs of specific target user populations. Particularly, there is no resource specifically designed to address the needs of health informatics students and novice researchers. Their learning curve is considerably steep when they explore datasets using existing resources. We believe the lack of a healthcare dataset information resource that brings information from various resources together to address the unique needs and questions from these learners is an important gap in health informatics development.

To bridge the gap, we have developed the Dataset Information Resource (DIR) framework, specifically aimed at helping entry-level health informatics students and researchers. For these novices, the challenges are different from established researchers. It is not the discovery of datasets that is important. Rather, the importance lies in

the surveying of the landscape of existing datasets and the identification of a proper dataset from the set of common datasets for a given problem. Additionally, the understanding of the dataset and related analytical methods is critically important. The DIR framework does not contain actual data from the datasets. Instead, it is a specialized knowledge base that provides comprehensive knowledge and answers sophisticated questions about noteworthy datasets that address the needs of beginning learners. Besides common information about datasets, such as descriptions, we focus more on knowledge needed by novices, such as analytical methods that datasets can utilize. In this case, novices can quickly obtain a solid understanding through concrete cases. Moreover, we provide dataset blogs in the DIR so that users can easily start data analysis by following sample codes and instructions.

For a flexible, meaningful, and robust knowledge representation, we leveraged Semantic Web [8] technologies. Meanwhile, we incorporated the W3C Dataset Description Profile standard [9] developed by the Semantic Web Health Care and Life Sciences (HCLS) interest group to ensure that the metadata delivered are well defined and organized. The current DIR prototype focuses on 12 representative datasets in healthcare, including both public and proprietary datasets. The prototype is published and accessible via https://cci-hit.uncc.edu/dir/.

2.2 Methods

The DIR framework is based on Semantic Web technologies. Building on them, we developed methods to extract knowledge from the datasets as well as existing research articles that had analyzed these datasets. We also developed a question-answering module that answered novice questions that had been posted on the web. In the following sections, we briefly describe the Semantic Web first and then describe the system design, major components, knowledge representation and extraction, and dataset learning of the DIR framework.

2.2.1 Semantic Web

The Semantic Web is an extension that adds semantics and logic to the well-known World Wide Web (WWW). In the traditional web pages, entities, such as concepts, are dispersed in the text. They are not clearly identified and their relationships are not explicitly represented. In contrast to traditional web pages, the Semantic Web enhances the regular web by coding and linking important concepts. Therefore, it makes semantics behind data understandable not only to human beings but also to machines. The Semantic Web is based on the Resource Description Framework (RDF) [10]. To link entities, RDF provides a straightforward syntax for describing resources, which is called "triple". An RDF triple contains three components—the subject, predicate, and the object, where the predicate represents the relationship between the subject and the object. To query the linked entities, a query language, SPARQL Protocol and RDF Query Language (SPARQL) [11], is designed, which is the key to reasoning. With the support of these techniques, a number of RDF-based resource frameworks have already been developed that show the power of the Semantic Web, such as DBpedia [12] and the Neuroscience Information Framework (NIF) [13].

2.2.2 DIR Framework Overview

The proposed architecture of the DIR system is shown in Figure 2.1. It consists of three major components: 1) knowledge representation (requires the ability to represent metadata in a flexible, extendable, and reusable way to meet and surpass the FAIR Data Principles [14]), 2) question answering (delivers exact knowledge to novices), and 3) metadata extraction (extracts metadata tailored to novices from a large number of diverse dataset resources). With these components, the system can integrate and represent knowledge from scattered datasets, allow flexible research questions, and provide precise answers at a suitable level of comprehension.

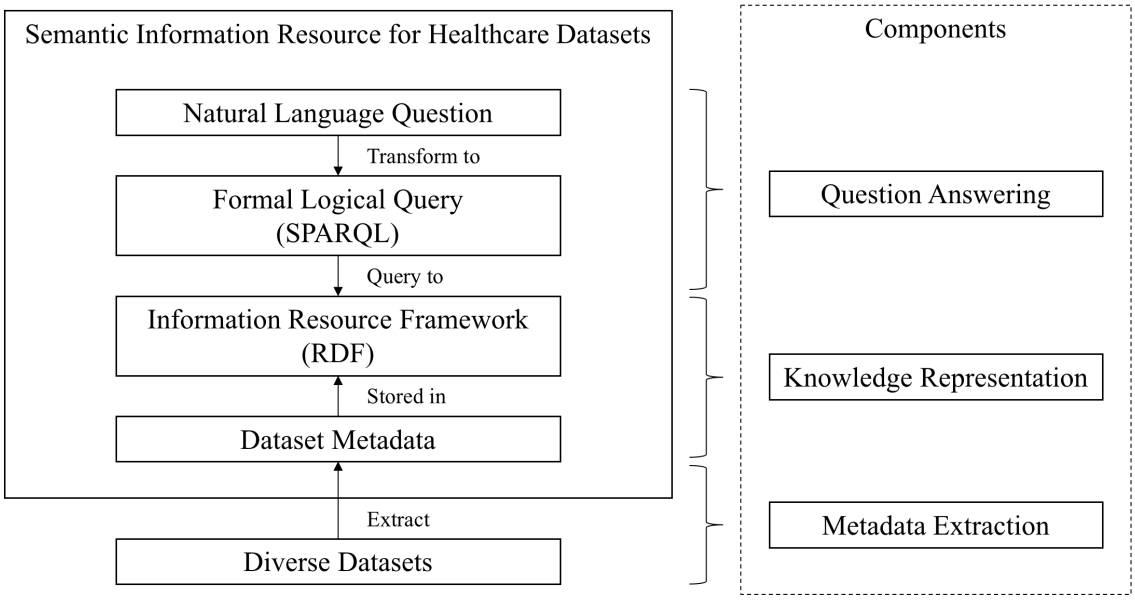

Figure 2.1: Proposed architecture of DIR system.

The DIR prototype is built on top of the open-sourced Semantic MediaWiki (SMW) platform [15] for knowledge representation and question answering. SMW, which tightly couples traditional web pages with an RDF representation to capture essential knowledge, is an extension of MediaWiki (MW) [16] (see Figure 2.2). Additionally, MW is well-known as the foundation of Wikipedia, whose English site contains 5,605,853 articles. Therefore, advantages of MW—such as stability facing massive content and heavy traffic—and advantages of SMW—including the embedded functionality to represent RDF triples by using properties, classes, and semantic forms—can be fully leveraged. Once the knowledge of diverse datasets is extracted, SMW provides a platform for representation and a SPARQL-like mechanism for the semantic query.

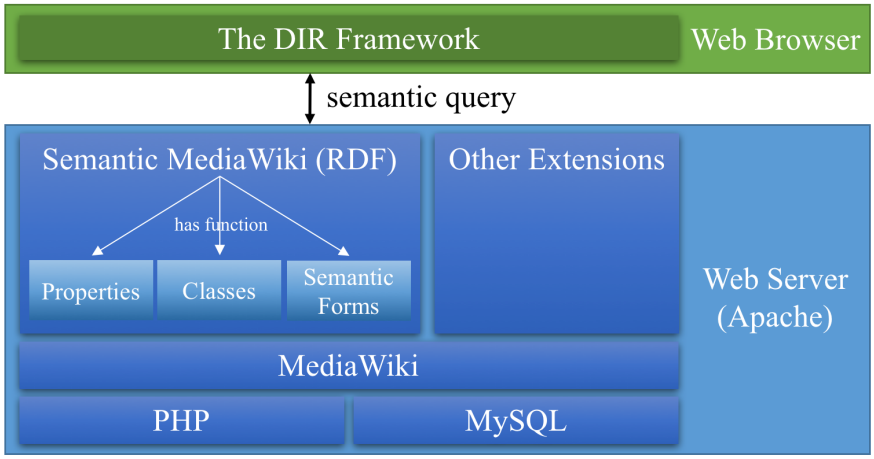

Figure 2.2: Infrastructure of DIR prototype.

2.2.3 Knowledge Representation in DIR

To represent dataset metadata in a standard manner that is findable, accessible, interoperable, and reusable, we adopt the W3C Dataset Description Profile [9] as the basis of a metadata description model. This profile categorizes dataset metadata in three levels: summary, version, and distribution (see Figure 2.3). The summary level is the highest-level description of datasets for the most common information that is independent of specific versions, such as titles, publishers, and homepage links. The version level, as an intermediary of summary and distribution levels, captures version-specific metadata, such as version identifiers and issue dates. The distribution level describes specific forms of a specific version. It includes the most detailed information and guidance, such as data items and links to achieve data. In the DIR prototype, each level of a dataset is a page. Since a dataset can have multiple versions and each version may have various forms, each dataset is described by at least three pages—a summary level (the entrance), at least one version level, and at least one distribution level. For each level, the W3C profile defines a set of suggested data elements, properties, and ranges. The properties that describe datasets are all selected from existing ontologies, such as the Provenance Authoring and Versioning ontology (pav) [17], Data CATalog

vocabulary (dcat) [18], and the CItation Typing Ontology (cito) [19]. Since levels depend on each other, several specific properties are defined to link different level pages of a dataset, such as pav:hasCurrentVersion (links the summary level to the version level) and dcat:distribution (links the version level to the distribution level).

The DIR framework further extends the W3C Dataset Description Profile standard to incorporate properties that represent specific knowledge needed to address the learning needs of health informatics novices. Figure 2.3 presents the extended properties in bold font, and Table 2.1 illustrates the detailed extension. As shown in Table 2.1, four major types of knowledge are currently extended: descriptive information, publication-related metadata, detailed data elements, and blogs. Among publication-related metadata, the Publication-based Popularity Index (PPI) is a special property used to compare and rank datasets (see Chapter 3). Blogs of each dataset are unique and important metadata in the DIR and elaborate on concrete instructions, sample codes, and results that guide an easy start for practice. These blogs targeting novices are written by experienced dataset users, so direct support is strongly provided.

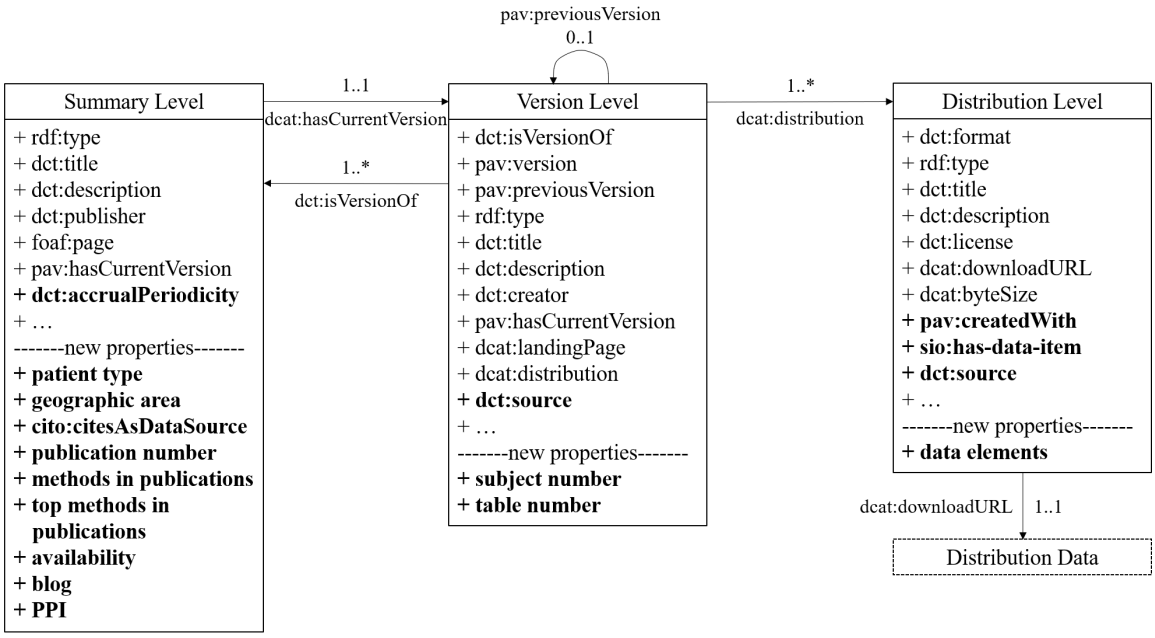

Figure 2.3: Schema of extended W3C dataset description profile.

2.2.4 Datasets in Current DIR Prototype

The current implementation of the DIR includes 12 representative datasets in healthcare, of which 3 datasets—Healthcare Cost and Utilization Project (HCUP) [20], Truven Health MarketScan (MarketScan) [21], and Medical Information Mart for Intensive Care (MIMIC) [22]—are retained from the previous DIR version [23]; nine others are selected from working group notes discussed by domain experts at the UNC Charlotte Health Informatics and Outcomes Research Academy [24]. The nine extended datasets are National Health and Nutrition Examination Survey (NHANES) [25], SEER-Medicare Linked Database (SEER-Medicare) [26], National Longitudinal Study of Adolescent to Adult Health (Add Health) [27], Minimum Data Set (MDS) [28], Clinical Practice Research Datalink (CPRD) [29], The Health Improvement Network (THIN) [30], Premier Healthcare Database (Premier) [31], Clinformatics Data Mart (Clinformatics) [32], and Humedica NorthStar (Humedica) [33].

There are several reasons to choose these datasets. To verify the universality of DIR knowledge representation, they cover most types of healthcare datasets, includ-

Table 2.1: Extended dataset metadata based on W3C dataset description profile.

Property	**Original Value in W3C profile**	**Extended Value in DIR**	**Level**	**Description**
Descriptive Information				
dct:accrualPeriodicity	IRI	IRI or xsd:string	Summary Level	Dataset update frequency
patient type	N/A	xsd:string	Summary Level	Patient type in a dataset (e.g., ICU patients)
geographic area	N/A	xsd:string	Summary Level	Geographic area of a dataset (e.g., city, region, and state)
availability	N/A	xsd:string	Summary Level	availability of a dataset (e.g., public or proprietary)
dct:source	IRI	IRI or xsd:string	Version Level Distribution Level	Data source provenance
subject number	N/A	xsd:integer	Version Level	Number of subjects (e.g., number of patients)
table number	N/A	xsd:integer	Version Level	Number of tables
pav:createdWith	IRI	IRI or xsd:string	Distribution Level	Tools used to create a dataset
Publication-Related Metadata				
cito:citesAsDataSource	N/A	IRI	Summary Level	Link to publications or a collection of publications using a dataset
publication number	N/A	xsd:integer	Summary Level	Number of publications that analyze a dataset
methods in publications	N/A	xsd:string	Summary Level	Methods used in publications to analyze a dataset
top methods in publications	N/A	xsd:string	Summary Level	Top (usually top 10) methods used in publications to analyze a dataset
PPI	N/A	xsd:float	Summary Level	A publication-based popularity index for dataset ranking
Detailed Data Elements				
sio:has-data-item	IRI	IRI or xsd:string	Distribution Level	Item listing (e.g., tables and entities)
data elements	N/A	xsd:string	Distribution Level	Data elements (e.g., attributes)
Blogs				
blog	N/A	IRI	Summary Level	Links to blogs of a dataset

ing claims data (SEER-Medicare, CPRD, MarketScan, Premier, and Clinformatics), electronic medical records (MDS and THIN), hospital data (SEER-Medicare, HCUP, MIMIC, and Humedica), laboratory data (Clinformatics), surveys (NHANES and Add Health), and contextual data (Add Health).

Additionally, these datasets are all large and complex datasets in healthcare, of which four (SEER-Medicare, Add Health, and Clinformatics) even include multiple types listed in the prior paragraph. Most of them have a large number of subjects. For example, HCUP includes the largest collection of longitudinal hospital care data in the United States, and MarketScan consists of nearly 240,000,000 patients' fully integrated, de-identified, individual-level healthcare claims data. In addition to the large amount of data, the diversity of data and the complexity of the structure make novices more difficult to understand and begin to analyze the datasets. For example, MIMIC contains not only numeric and textual data stored in tabular forms, such as lab results and electronic documentation but also graphical data that are stored separately, such as bedside monitor trends and waveforms. Adopting these graphical signals requires not only a deep understanding of data themselves but also sufficient computer skills to convert them into analyzable data and adequate knowledge to decide analytical methods.

Moreover, these datasets are all widely used in healthcare data analytics. A large number of research articles have been published based on these datasets. By searching in PubMed Central (PMC) [34]—an authoritative electronic archive of free full-text biomedical and life sciences journal articles supported by U.S. National Institutes of Health's National Library of Medicine (NIH/NLM)—the most studied dataset, NHANES, was mentioned in 37,485 articles, while the least discussed dataset among them, Humedica, was mentioned in up to 22 articles. On average, each dataset contributes to more than 4,000 publications in PMC.

Finally, these datasets are representative of both public and proprietary datasets.

Among the 12 datasets, two of them (NHANES and MIMIC) are public for research purposes, nine of them (SEER-Medicare, HCUP, MDS, CPRD, MarketScan, THIN, Premier, Clinformatics, and Humedica) are proprietary, and one dataset (Add Health) provides both public- and contractual-use data. In novices' perspectives, complex proprietary datasets are even more challenging than public datasets because they have difficulty retrieving information elsewhere to help them build up a good understanding accurately and quickly.

In the current implementation, we manually extracted most of the metadata from dataset documentations and semi-automatically extracted metadata about analytical methods from publications. The extracted metadata was first stored in the RDF triple format in Excel spreadsheets and imported into MW, using a Python script that converts spreadsheets to MW importable XML files. To ensure the accuracy of manually extracted metadata, a team of health informatics research assistants was formed to review and correct these metadata iteratively.

2.2.5 Extraction of Analytical Methods from Publications

Data analytical methods that have been successfully applied to datasets are important knowledge for data science learners. To deliver this knowledge, we developed a semi-automatic method to extract various analytical methods that had been used in published articles that analyzed the specific datasets in the DIR. For this task, we first developed an ontology of data analytical methods, Method Ontology (MethodOntology.owl [35]), which extended an existing ontology. Based on the Method Ontology, we developed a rule-based Named Entity Recognition (NER) pipeline to extract instances of analytical methods reported in selected publications.

We used PMC as the data resource and downloaded full-text articles that mentioned the 12 datasets, using the keyword identification method in Section 3.3.2. In total, 48,282 PDF-format publications were obtained. The publication number of each dataset is shown in Table 2.2. To preprocess these publications, we developed

a pre-processor, written in the Bash command and Python programming language, which included three major steps: (1) converted PDF files to plain text; (2) excluded proceedings and articles that only cited a dataset without analyzing it; and (3) selected relevant content by removing reference sections. After preprocessing, 25,201 publications remained.

Table 2.2: Publication numbers of 12 datasets.

Dataset	**# of PDF-format articles in PMC**	**# for method extraction after preprocessing**	**# that analyzing datasets**
NHANES	37,485	16,213	10,674
SEER-Medicare	2,569	2,276	1,627
Add Health	1,881	1,477	1,028
HCUP	1,785	1,398	993
MDS	1,337	1,053	584
CPRD	1,014	735	477
MarketScan	985	920	614
THIN	733	678	434
MIMIC	237	206	152
Premier	165	158	95
Clinformatics	69	65	49
Humedica	22	22	9
Total	**48,282**	**25,201**	**16,736**

The Method Ontology describes data analytical methods, which include all major machine learning, data mining, and statistical methods. This ontology extends the Data Mining Knowledge Base (DMKB.owl) of the Data Mining OPtimization Ontology (DMOP version 5.4), which was originally designed to support informed decision-making in the data mining (DM) process [36]. The DMKB.owl describes instances of DMOP concepts, including individual algorithms in popular data mining software, such as RapidMiner and Weka. For the method extraction purpose, the Method Ontology extended it by adding and linking new methods, which were extracted in a training set of dataset publications, and synonyms of all method instances. Figure 2.4 shows the structure of major method classes and a few examples of extended instances in the Method Ontology.

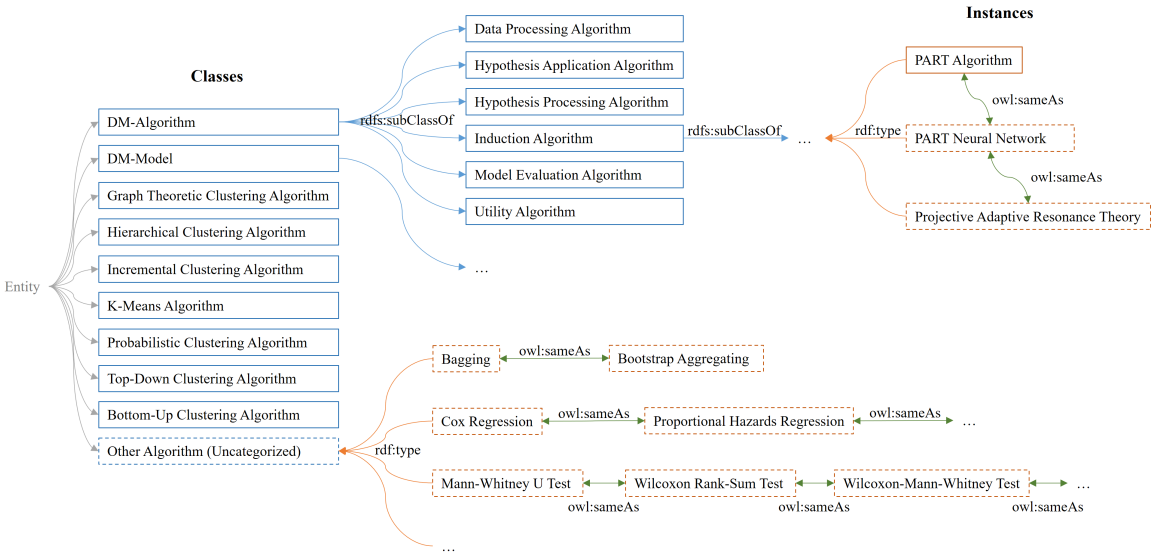

Figure 2.4: Structure of major method classes and some examples of extended instances in Method Ontology. Extended elements are shown in dashed boxes.

A rule-based NER was carried out in the Clinical Language Annotation, Modeling, and Processing Toolkit (CLAMP) [37]—a Natural Language Processing (NLP) software—and was handled by a pipeline that included a sentence detector, a tokenizer, and a dictionary lookup component. The input to this pipeline included the preprocessed publications as well as a method dictionary with semantic labels generated from the Method Ontology. After all potential method entities in the publications were extracted, a post-processor was developed to refine these entities and to combine synonyms for further metadata representation. As a result, method entities were extracted and were represented on summary level pages of the datasets. Assume that a publication that analyzed a dataset mentioned at least one analytical method in the full text. In that case, more than half (16,736 out of 25,201) of the preprocessed publications would have analyzed these datasets. According to the publications, the most frequently used methods for the 12 datasets, as well as proportions of publications that utilized the corresponding method, are shown in Table 2.3. Among these methods, logistic regression, mentioned in 4,229 publications, was the most frequently

used (see Figure 2.5).

We evaluated the pre-processor and the method extraction steps separately. The results showed that the 95% confidence interval of the pre-processor's accuracy was [92.26%, 99.39%], and the precision and recall of the analytical method extraction were 93.82% and 90.53%, respectively.

2.2.6 Dataset Learning and Question Answering

Once the dataset knowledge is extracted and represented, the direct way to query the knowledge is to write SPARQL-like queries in the semantic search mechanism provided by SMW. While this direct method is powerful, it requires an understanding of the Semantic Web and SPARQL, which is clearly burdensome to novices. Our current approach to addressing this issue is to offer a simplistic question-answering functionality by identifying the most popular questions that novices ask and providing ready-to-use queries. We created a parameterized question page for each representative question, where users can simply input words and click the Run Query button to obtain precise answers. The list of current parameterized question pages is shown in Table 2.4.

For example, if users are curious about which datasets can successfully utilize the Support Vector Machine, they can simply visit the "Which datasets can I apply the method to" question page, choose or type in "Support Vector Machine," and click the Run Query button to obtain "Answer: NHANES, CPRD, THIN, HCUP, MDS, MIMIC." The dataset result is in order based on the PPI recommendation. In this example, the query below has already been embedded in the question page template:

```
Answer:
{{#ask:
[[Category:Summary Level]]
[[Methods in publications::{{{method|}}}]]
|sort=PPI
|order=desc
}}.
```

As another example, if users need to investigate large datasets that have more than 1,000,000 subjects, they can refer to the parameterized question page—"Which datasets have more than a specific number of subjects"—that includes the following query:

```
{{#ask:
[[Category:Summary Level]]
[[-Dct:isVersionOf::<q>
[[Category:Version Level]]
[[Subject number::>={{{subject_number|}}}]]
</q>]]
|sort=PPI
|order=desc
}}.
```

To determine the most popular questions that novices ask, we analyzed a variety of resources, including a publication that guides novices to conduct high-value dataset analysis [38], questions labeled as "dataset" on question-and-answer sites (e.g., Quora [39] and Stack Exchange [40]), and opinions from health informatics novices through interviews.

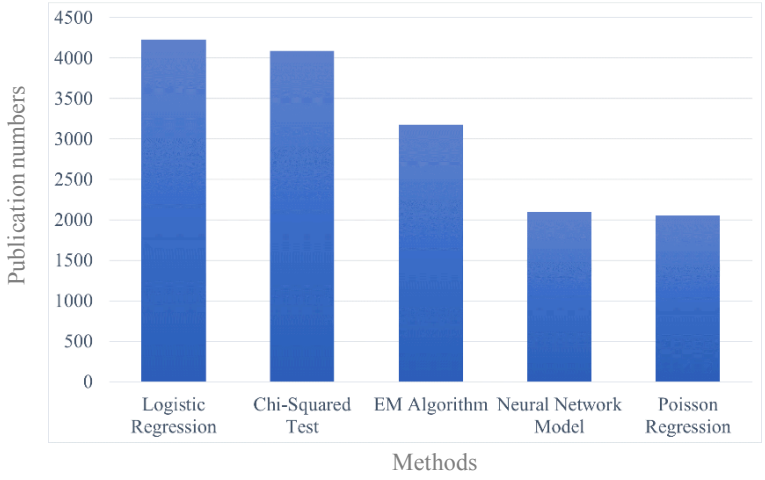

Figure 2.5: The most frequently used methods in publications of 12 datasets.

Table 2.3: Ten most frequently used methods to analyze each dataset.

Dataset	Methods				
NHANES	EM Algorithm 29.55%	Neural Network Model 19.63%	Wilcoxon Signed-Rank Test 16.69%	Poisson Regression 15.02%	Chi-Squared Test 14.85%
	Kruskal-Wallis Test 14.32%	Logistic Regression 12.56%	Log-Rank Test 12.17%	Linear Regression 10.04%	T-Test 8.51%
SEER-Medicare	Chi-Squared Test 54.52%	Logistic Regression 50.83%	Cox Regression 39.64%	Log-Rank Test 17.46%	Survival Analysis 14.87%
	T-Test 11.12%	Regression Model 10.45%	Kaplan-Meier Survival Estimates 9.34%	Linear Regression 8.85%	Propensity Score Matching 7.01%
Add Health	Logistic Regression 50.00%	Chi-Squared Test 33.17%	Linear Regression 13.13%	Regression Model 9.82%	Principal Component Analysis 8.07%
	ANOVA 7.49%	Poisson Regression 5.74%	T-Test 5.06%	Propensity Score Matching 3.40%	Cox Regression 3.40%
HCUP	Logistic Regression 57.91%	Chi-Squared Test 48.44%	Linear Regression 20.24%	T-Test 18.03%	Regression Model 15.61%
	ANOVA 9.87%	Poisson Regression 9.06%	Cox Regression 7.45%	Mann-Whitney U Test 7.35%	Bootstrap 4.23%
MDS	Logistic Regression 42.12%	Chi-Squared Test 39.73%	Linear Regression 17.29%	Regression Model 14.90%	T-Test 13.53%
	ANOVA 13.18%	Cox Regression 9.93%	Mann-Whitney U Test 7.19%	Bootstrap 4.11%	Survival Analysis 3.77%
CPRD	Logistic Regression 42.35%	Cox Regression 31.03%	Chi-Squared Test 18.87%	Poisson Regression 12.37%	Propensity Score Matching 10.48%
	Linear Regression 9.85%	Regression Model 8.60%	Survival Analysis 6.08%	T-Test 5.66%	Kaplan-Meier Survival Estimates 4.61%
MarketScan	Chi-Squared Test 47.88%	Logistic Regression 43.32%	Cox Regression 19.22%	T-Test 12.87%	Poisson Regression 12.21%
	Propensity Score Matching 10.91%	Linear Regression 9.93%	Regression Model 9.77%	ANOVA 6.68%	Fisher's Exact Test 5.86%
THIN	Logistic Regression 37.33%	Cox Regression 26.04%	Chi-Squared Test 23.27%	Poisson Regression 12.44%	Regression Model 9.91%
	Inverse Probability Weighting 8.99%	Linear Regression 8.53%	T-Test 8.06%	Survival Analysis 6.91%	Propensity Score Matching 6.68%
MIMIC	Logistic Regression 45.39%	Chi-Squared Test 20.39%	T-Test 17.76%	Mann-Whitney U Test 15.79%	Regression Model 14.47%
	Support Vector Machine 14.47%	Linear Regression 11.84%	Cox Regression 11.18%	Kolmogorov-Smirnov Test 9.87%	K-Nearest Neighbors 9.21%
Premier	Chi-Squared Test 41.05%	K-Means 38.95%	Decision Tree Model 27.37%	Logistic Regression 21.05%	Propensity Score Matching 14.74%
	Kruskal-Wallis Test 13.68%	Linear Discriminant Analysis 11.58%	Regression Model 11.58%	Linear Regression 8.42%	T-Test 8.42%
Clinformatics	Linear Regression 44.90%	Bootstrap 28.57%	Regression Model 20.41%	Kruskal-Wallis Test 14.29%	Chi-Squared Test 12.24%
	F-Test 12.24%	Cox Regression 10.20%	Logistic Regression 10.20%	ANOVA 8.16%	Survival Analysis 6.12%
Humedica	Chi-Squared Test 33.33%	Logistic Regression 22.22%	Bootstrap 22.22%	Fisher's Exact Test 22.22%	Cox Regression 11.11%
	T-Test 11.11%	Linear Regression 11.11%	Propensity Score Matching 11.11%	Survival Analysis 11.11%	Ensemble Learning 11.11%

Table 2.4: Eighteen parameterized question pages in current DIR.

Data-Driven Questions	Which datasets include some specific information/data elements?
	Which datasets have more than a specific number of subjects?
Method-Driven Questions	Which datasets can I apply a specific method to?
Introduction Questions	What does a dataset talk about?
	How to get a specific dataset?
	What are the methods that publications used with a specific dataset?
	What are the publications using a specific dataset?
	Is a specific dataset open to the public?
	How many subjects are there in a specific dataset?
	How many tables are there in a specific dataset?
	What are the different tables/files in a database?
	What are the data elements in a specific dataset?
	What are the patient types that a specific dataset handles?
	How frequently are data updated in a dataset?
	How many times is a dataset cited?
	Who reports the data in a specific dataset?
	What is the geographic area of a dataset?
	What is the full name of a dataset?

2.3 Results and Discussion

A prototype of the DIR has been developed and released. It is accessible via https://cci-hit.uncc.edu/dir/. The current DIR homepage is shown in Figure 2.6. Built on the foundation of the Semantic Web and the extended W3C dataset description profile, we have provided knowledge about 12 representative datasets in healthcare—NHANES, SEER-Medicare, Add Health, HCUP, MDS, CPRD, MarketScan, THIN, MIMIC, Premier, Clinformatics, and Humedica—and five blogs. To facilitate novices' question answering, 18 ready-to-use questions (Table 2.4) have been provided. In addition, the more powerful semantic search function is available for users who are familiar with SPARQL. To ease usability, a tutorial and a support page with an issue tracker and a feedback form are also provided.

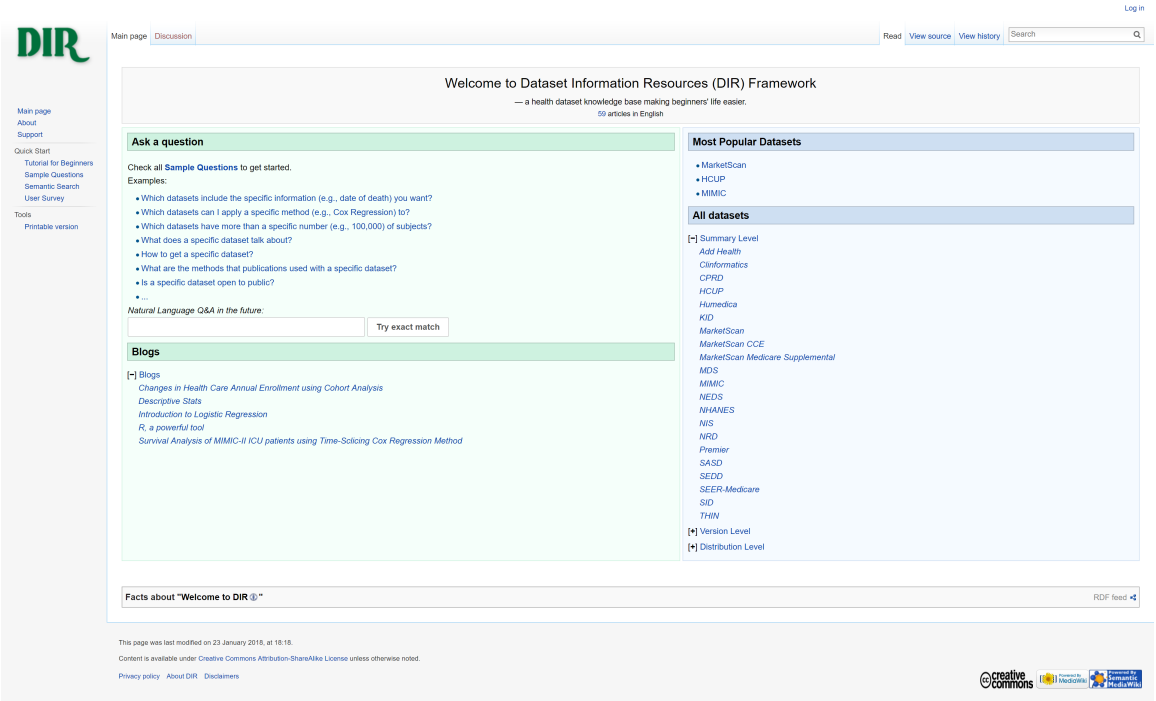

Figure 2.6: Current DIR homepage.

At the time of this paper's submission, the DIR prototype contained 264 pages. The average page loading time was 1.44 seconds. The current approach to add a new dataset includes both manual metadata extractions (from documentations with

team review-based quality check) and semi-automatic knowledge extractions (from publications using NLP technologies). To add a new dataset, the current approach takes approximately one day, in general, for manual extractions and a few minutes for semi-automatically extracting analytical methods, excluding the time to collect publications.

We have conducted a survey and collected feedback from 15 target users who were novices in healthcare data research. Of the target users, 40% had a background in health informatics, and 86.7% had a background in data analytics. We asked the subjects to compare Google, DIR, and other resources in seven use cases and also asked for general comments. The survey results indicated that 73.3% of users, on average, preferred the DIR in these use cases. Significantly, 100% of them preferred the DIR in the case of finding datasets that included a particular data element; 93.3% preferred the DIR when they wanted to adopt a specific analytical method; and 86.7% preferred the DIR in the case of discovering large-enough datasets, such as a dataset that had more than 1,000,000 subjects. In terms of more general knowledge, users tended to rely on broader resources. For example, only 60% of users chose the DIR when they were looking for basic descriptions of a dataset or tutorials about gaining access, while others felt more comfortable on searching in Google, browsing the official website, or using both DIR and other resources simultaneously. Overall, the DIR obtained a score of 86.7% in helpfulness, 83.8% in ease of discovering datasets, 82.9% in ease of question answering, and 82.9% in the scale of meeting users' expectations about healthcare dataset information resources.

According to comments in survey responses, users highlighted the advantages of the DIR as targeted and novice-friendly. As some users commented: "It filters out the irrelevant information and is more structural"; "Beginner-friendly. Information is exhibit[ed] clearly to the user"; and "Sample questions and semantic search are very useful for researchers to find the right dataset or information, or we can say it looks

more intelligent than other search engine[s] like [G]oogle."

However, the DIR clearly has several limitations in this initial phase. (1) The current DIR prototype still relies on manual extractions in part, which is time-consuming and labor-intensive for DIR developers during dataset extending. This limitation has two possible ways to be improved. One refers to the entity linking and typing topic that is intensely discussed in Semantic Web conferences, such as the Open Knowledge Extraction Challenge (OKE) [41] at the European Semantic Web Conference (ESWC). The other way, mentioned by the CEDAR project, involves promoting an authoring-friendly ecosystem in the healthcare dataset community and encouraging researchers to contribute open metadata. (2) Currently, we do not differentiate subclasses of analytical methods, that is, the statistical methods, such as Chi-Square Test, are listed together with machine learning methods, such as Ensemble Learning. Further classification of methods based on the Method Ontology will be needed to address more detailed user questions. (3) As one user commented in the survey: "For now, finding a question is not that hard. However, if the question set becomes larger, then I think it can cause a problem. Somehow you need to facilitate this part," which reveals that preparing query-embedded question-answering pages can only be a temporary solution. When the system is expanded, a real natural language question-answering functionality should be implemented. Actually, question answering is a stand-alone topic in the Semantic Web community and has been discussed over decades in conferences (e.g., the open challenge on Question Answering over Linked Data (QALD) [42] at ESWC) and publications (e.g., [43][44][45][46][47]). (4) As another user mentioned in the survey: "I'm not sure if researchers will trust the information on DIR." Rely on simple quality check approaches is one of the limitations. To ensure quality and to gain user trust, a systematic quality assurance method needs to be developed and reported.

2.4 Conclusions and Future Work

We conclude that it is feasible to develop a DIR that provides value for entry-level health informatics students and researchers. Knowledge about datasets is effectively represented in Semantic Web technologies. At this stage, the DIR has already been able to provide comprehensive and relevant knowledge of 12 important healthcare datasets, which is expected to improve health informatics novices' ability to learn data analysis using suitable datasets.

In contrast to bioinformatics datasets, of which most data elements have already been represented in RDF at the knowledge level, the DIR will continue focusing on the healthcare datasets that are usually at a lower level granularity.

Further development is underway to improve efficiency, accuracy, and scalability. Suitable directions for expansion include two levels: content and function. The content level adds more healthcare datasets, identifies more types of knowledge for target users, and involves a systematic quality assurance method to ensure the quality of metadata. The function level includes developing a natural language-based question-answering component, more automated methods to extract knowledge, intelligent functionalities to compare similar datasets, and collaborative features, such as discussion forums that allow users to help each other and suggest new content.

Chapter 2 Reference List

[1] "Healthdata.gov." https://www.healthdata.gov/.

[2] "Data | Centers for Disease Control and Prevention." https://data.cdc.gov/.

[3] "Dataset Compendium Overview | sgim.org." https://www.sgim.org/communities/research/dataset-compendium.

[4] M. A. Musen, C. A. Bean, K.-H. Cheung, M. Dumontier, K. A. Durante, O. Gevaert, A. Gonzalez-Beltran, P. Khatri, S. H. Kleinstein, M. J. O'connor, *et al.*, "The center for expanded data annotation and retrieval," *Journal of the American Medical Informatics Association*, vol. 22, no. 6, pp. 1148–1152, 2015.

[5] R. S. Goncalves, M. J. O'Connor, M. Martinez-Romero, A. L. Egyedi, D. Willrett, J. Graybeal, and M. A. Musen, "The cedar workbench: An ontology-assisted environment for authoring metadata that describe scientific experiments," in *International Semantic Web Conference*, pp. 103–110, Springer, 2017.

[6] "bioCADDIE | biomedical and healthCAre Data Discovery and Indexing Ecosystem." https://biocaddie.org/.

[7] L. Ohno-Machado, S.-A. Sansone, G. Alter, I. Fore, J. Grethe, H. Xu, A. Gonzalez-Beltran, P. Rocca-Serra, A. E. Gururaj, E. Bell, E. Soysal, N. Zong, and H.-E. Kim, "Finding useful data across multiple biomedical data repositories using DataMed," *Nature Genetics*, vol. 49, pp. 816–819, May 2017.

[8] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 28–37, 2001.

[9] M. Dumontier, A. J. Gray, M. S. Marshall, V. Alexiev, P. Ansell, G. Bader, J. Baran, J. T. Bolleman, A. Callahan, J. Cruz-Toledo, *et al.*, "The health care and life sciences community profile for dataset descriptions," *PeerJ*, vol. 4, p. e2331, 2016.

[10] E. Miller, "An introduction to the resource description framework," *Bulletin of the American Society for Information Science and Technology*, vol. 25, no. 1, pp. 15–19, 1998.

[11] E. Prud'Hommeaux and A. Seaborne, "SPARQL query language for RDF," *W3C recommendation*, vol. 15, 2008.

[12] P. N. Mendes, M. Jakob, and C. Bizer, "DBpedia: A Multilingual Cross-domain Knowledge Base.," in *LREC*, pp. 1813–1817, 2012.

[13] S. D. Larson and M. E. Martone, "NeuroLex. org: an online framework for neuroscience knowledge," *Frontiers in neuroinformatics*, vol. 7, no. 18, 2013.

[14] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, and P. E. Bourne, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific data*, vol. 3, 2016.

[15] M. Krotzsch, D. Vrandecic, and M. Volkel, "Semantic mediawiki," in *The Semantic Web-ISWC 2006*, pp. 935–942, 2006.

[16] "MediaWiki." `https://www.mediawiki.org/wiki/MediaWiki`.

[17] P. Ciccarese, S. Soiland-Reyes, K. Belhajjame, A. J. Gray, C. Goble, and T. Clark, "PAV ontology: provenance, authoring and versioning," *Journal of Biomedical Semantics*, vol. 4, no. 1, p. 37, 2013.

[18] "Data Catalog Vocabulary (DCAT)." `https://www.w3.org/TR/vocab-dcat/`.

[19] "CiTO, the Citation Typing Ontology." `http://www.sparontologies.net/ontologies/cito/source.html`.

[20] C. Steiner, A. Elixhauser, and J. Schnaier, "The healthcare cost and utilization project: an overview.," *Effective clinical practice : ECP*, vol. 5, pp. 143–151, Dec. 2001.

[21] "MarketScan Research Data." `https://marketscan.truvenhealth.com/marketscanportal/`.

[22] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, May 2016.

[23] J. Shi, M. Zheng, L. Yao, and Y. Ge, "DIR - A semantic information resource for healthcare datasets," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 805–810, Nov. 2017.

[24] "Health Informatics and Outcomes Research Academy | UNC Charlotte." `https://hinora.uncc.edu/`.

[25] "NHANES - National Health and Nutrition Examination Survey Homepage." `https://www.cdc.gov/nchs/nhanes/index.htm`.

[26] "SEER-Medicare Linked Database." `https://healthcaredelivery.cancer.gov/seermedicare/`.

[27] "Add Health." `http://www.cpc.unc.edu/projects/addhealth`.

[28] "Minimum Data Set 3.0 Public Reports Overview." https://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/Minimum-Data-Set-3-0-Public-Reports/index.html.

[29] "Clinical Practice Research Datalink - CPRD." https://www.cprd.com/home/.

[30] "THIN Database." https://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database.

[31] "Premier Healthcare Database White Paper." https://learn.premierinc.com/pharmacy-and-research/premier-healthcare-database-whitepaper.

[32] "Clinformatics Data Mart." http://www.optum.ca/life-sciences/differentiate-products/marketing-analytics/clinformatics-data-mart.html.

[33] "Humedica NorthStar." https://www.optum.com/solutions/life-sciences/explore-data/advanced-analytics/humedica-northstar.html.

[34] "Home - PMC - NCBI." https://www.ncbi.nlm.nih.gov/pmc/.

[35] J. Shi, "Method Ontology." https://cci-hit.uncc.edu/dir/ontologies/MethodOntology.owl.

[36] C. M. Keet, A. Lawrynowicz, C. d'Amato, A. Kalousis, P. Nguyen, R. Palma, R. Stevens, and M. Hilario, "The Data Mining OPtimization Ontology," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 32, pp. 43–53, May 2015.

[37] E. Soysal, J. Wang, M. Jiang, Y. Wu, S. Pakhomov, H. Liu, and H. Xu, "Clamp–a toolkit for efficiently building customized clinical natural language processing pipelines," *Journal of the American Medical Informatics Association*, vol. 25, no. 3, pp. 331–336, 2017.

[38] A. K. Smith, J. Z. Ayanian, K. E. Covinsky, B. E. Landon, E. P. McCarthy, C. C. Wee, and M. A. Steinman, "Conducting High-Value Secondary Dataset Analysis: An Introductory Guide and Resources," *Journal of General Internal Medicine*, vol. 26, pp. 920–929, Aug. 2011.

[39] "Quora - The best answer to any question." https://www.quora.com/.

[40] "Hot Questions - Stack Exchange." http://stackexchange.com/.

[41] R. Speck, M. Roder, S. Oramas, L. Espinosa-Anke, and A.-C. N. Ngomo, "Open Knowledge Extraction Challenge 2017," in *Semantic Web Challenges*, Communications in Computer and Information Science, pp. 35–48, May 2017.

[42] R. Usbeck, A.-C. N. Ngomo, B. Haarmann, A. Krithara, M. Roder, and G. Napolitano, "7th Open Challenge on Question Answering over Linked Data (QALD-7)," in *Semantic Web Challenges*, Communications in Computer and Information Science, pp. 59–69, May 2017.

[43] V. Lopez, V. Uren, M. Sabou, and E. Motta, "Is question answering fit for the semantic web?: a survey," *Semantic Web*, vol. 2, no. 2, pp. 125–155, 2011.

[44] J. Jeon, W. B. Croft, and J. H. Lee, "Finding similar questions in large question and answer archives," in *Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 84–90, 2005.

[45] I. Androutsopoulos, G. D. Ritchie, and P. Thanisch, "Natural language interfaces to databases–an introduction," *Natural language engineering*, vol. 1, no. 1, pp. 29–81, 1995.

[46] L. Hirschman and R. Gaizauskas, "Natural language question answering: the view from here," *natural language engineering*, vol. 7, no. 4, pp. 275–300, 2001.

[47] S. Shekarpour, D. Lukovnikov, A. J. Kumar, K. Endris, K. Singh, H. Thakkar, and C. Lange, "Question Answering on Linked Data: Challenges and Future Directions," *arXiv:1601.03541 [cs]*, Jan. 2016. arXiv: 1601.03541.

CHAPTER 3: A PUBLICATION-BASED POPULARITY INDEX (PPI) FOR HEALTHCARE DATASET RANKING

3.1 Introduction

In the big data era, healthcare as one of the largest and fastest-growing industries has experienced a data explosion. An increasing number of datasets are emerging rapidly due to the adoption of electronic platforms in hospitals, insurance companies, and surrounding facilities. Taking advantage of these datasets, more researchers are leveraging health informatics and data analytics to promote medical science development and improve patient care. Healthcare datasets are usually complex, diverse, and hard to understand. To facilitate researchers in dataset discovery, a number of websites and data portals, such as HealthData.gov managed by U.S. Department of Health and Human Services and DataMed.org supported by the National Institutes of Health (NIH), have integrated basic metadata (e.g., descriptions, formats, and landing pages) of millions of datasets. In addition, we have developed a semantic Dataset Information Resource (DIR) to address the special needs for health informatics novices in learning and selecting datasets [1]. The DIR represents tailored metadata (e.g., analytical methods that can be utilized on a dataset) and provides parameterized question answering functionality. Although these efforts can help dataset discovery to a certain extent, choosing a quality dataset that is suitable for specific research inquiries are still challenging for researchers.

Obviously, data quality is one of the most important features of a dataset. It is the foundation of valid and reliable research findings. However, no known websites or data portals have taken data quality into account, which is largely due to the difficulty in measuring quality. To attempt to address issues in data quality for secondary use,

researchers have studied this specific area for more than a decade. To the best of our knowledge, almost all existing data quality research focuses on the Electronic Health Record (EHR) data (e.g., [2][3][4][5]) but rarely on other types, such as survey data (e.g., the National Health and Nutrition Examination Survey (NHANES)) and administrative claims data (e.g., the MarketScan dataset) that account for a large portion of research. For a better EHR data quality assessment, studies have tried to harmonize terms, methods, and practices in a conceptual way, which provides good guidance to empirically evaluate the quality. However, there is still a lack of consistent definition and categorization of quality dimensions. For example, while [2] organized data quality in three categories—conformance, completeness, and plausibility—[5] classified data quality as data accuracy, completeness, consistency, credibility, and timeliness. Aside from the inconsistency, the measurements of these dimensions are qualitative, which are not suitable to precisely compare a large number of candidate datasets. Additionally, intrinsic quality dimensions are more often discussed while other extrinsic dimensions of datasets, such as data accessibility, are equally important for researchers but lack consideration [2].

Therefore, we believe that a quantified method to comprehensively measure both intrinsic quality and extrinsic value properties for all kinds of healthcare datasets will be an important tool for researchers during the dataset selection. Moreover, because the acquisition of healthcare datasets are costly and time-consuming, the ability to identify valuable datasets and analyze salient design features of these datasets will also be significant for data providers to improve future healthcare datasets.

In this article, we propose one way of measuring the value of healthcare datasets from the perspective of popularity. We define a dataset as popular if it has been successfully analyzed by numerous researchers and an increasing number of researchers continue to analyze it. One of the best ways to identify successful analyses is to refer to the publications that have analyzed the dataset. Unlike traditional quality

dimensions, popularity of a dataset naturally reflects its intrinsic quality dimensions (especially plausibility and timeliness) and extrinsic value dimensions (including data accessibility and system availability) because to successfully obtain publishable results from a dataset depends on almost all dimensions indispensably. Therefore, we believe that a dataset with higher popularity is likely to be more valuable in general.

In order to identify datasets with higher popularity, there are two intuitive approaches. One is to simply count the number of related publications in a specific time period. However, this approach ignores the trend, which particularly indicates data quality and research opportunity over time. The other approach is to observe the histograms showing the relationship between years and numbers of publications. For example, Fig. 3.1 is a histogram of the National Longitudinal Study of Adolescent to Adult Health (Add Health) dataset, showing the publication numbers indexed annually by PubMed, which is the most authoritative medical literature search engine in the U.S. The publication of this dataset was first indexed by PubMed in 1998. Fig. 3.2 shows the histogram for the SEER-Medicare Linked Database (SEER-Medicare) since 1998. However, there are two obvious drawbacks to this approach: 1) Researchers can only subjectively observe the trends from figures without a measurement. 2) It is hard to compare both trends and numbers simultaneously across datasets (e.g., comparing Add Health to SEER-Medicare). Both of the drawbacks are even more noticeable when researchers have a large number of candidate datasets.

For these reasons, we have developed a Publication-based Popularity Index (PPI) that takes both the number of publications that have analyzed a dataset and the trend of analyzing the dataset into account in order to quantitatively evaluate and compare the goodness of healthcare datasets. As a result, researchers can eventually establish a ranking of their candidate datasets to support the dataset selection process, and data providers can identify the most valuable datasets for future dataset designs.

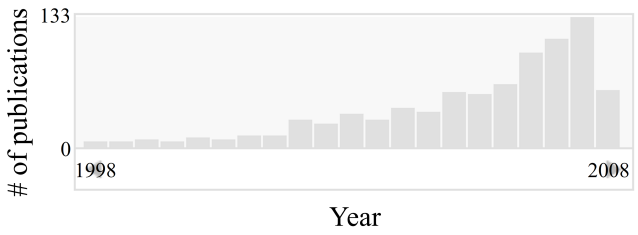

Figure 3.1: Add Health publications by year histogram drawn by PubMed.gov.

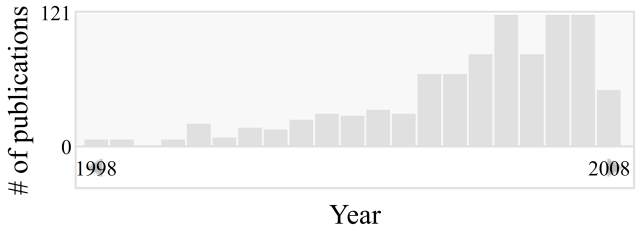

Figure 3.2: SEER-Medicare publications by year histogram drawn by PubMed.gov.

In this article, we define the PPI and discuss its properties in Section 3.2. In Section 3.3, we elaborate on our data source that consists of 14 representative healthcare datasets and present a method to identify publications that have analyzed these datasets. In Section 3.4, we rank the 14 representative datasets using the PPI and discuss the results. In Section 3.5, we summarize our contributions.

3.2 PPI for Healthcare Datasets

As popularity is a time-dependent measurement, let N equal the number of years that users want to consider. We define the PPI of a dataset for the past N years as:

$$PPI = \begin{cases} \bar{P} \cdot \exp\left(\ln|\ln|\beta|| \frac{\beta}{|\beta|}\right) & \text{for } |\beta| > e \\ \bar{P} & \text{for } |\beta| \leq e \end{cases}, \tag{3.1}$$

where

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P_i,$$

average number of publications that have analyzed the dataset in the past N years, where N is a user selection feature,

P_i =number of publications that have analyzed the dataset in i-th-to-last year,

β =slope of the simple linear regression model $y = \alpha + \beta x$,

$$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n \approx 2.71828.$$

In the simple linear regression model, $\{y\} = \{P_i$ sorted by year$\}$, and $\{x\} = \{$calendar years$\}$. For example, when $N = 5$, $\{y\} = \{$number of publications in 2013, ..., number of publications in 2017$\}$, and $\{x\} = \{2013, 2014, 2015, 2016, 2017\}$.

The PPI has two properties.

First, the PPI is monotone increasing in $\bar{P}$ and in β. The monotonicity in $\bar{P}$ is obvious. It is monotone in β because the PPI can be simplified as:

$$PPI = \begin{cases} \bar{P}\ln\beta & \text{for } \beta > e \\ \bar{P} & \text{for } -e \le \beta \le e \\ \bar{P}/\ln(-\beta) & \text{for } \beta < -e \end{cases};$$

therefore,

$$\partial PPI/\partial\beta = \begin{cases} \bar{P}/\beta > 0 & \text{for } \beta > e \\ 0 & \text{for } -e \le \beta \le e \\ -\bar{P}/\beta(\ln(-\beta))^2 > 0 & \text{for } \beta < -e \end{cases}.$$

Second, the range of the PPI is $[0, \infty)$. It is always positive as long as $\bar{P} \neq 0$. If $\bar{P} = 0$, which means that no publication has analyzed the dataset during the selected period, then $PPI = 0$. This makes perfect sense because the absence of publications means that the dataset is unpopular.

The design of the PPI takes both the average number of publications ($\bar{P}$) and the trend (β) into account. $\bar{P}$ can be viewed as a base, while the trend can make a heavy impact on the index. To restrict the impact of the trend and regard $\bar{P}$ as a more important factor, β has been balanced by logarithms. If an increasing trend is significant, i.e., $\beta \gg e$, then $\exp\left(\ln|\ln|\beta|| \frac{\beta}{|\beta|}\right) \gg 1$. That is, the trend gives a boost to $\bar{P}$ (e.g., $PPI = \bar{P} \cdot 3$ when $\beta = 20.1$). The higher the beta, the stronger the boosting. If a decreasing trend is significant, i.e., $\beta \ll -e$, then $0 < \exp\left(\ln|\ln|\beta|| \frac{\beta}{|\beta|}\right) \ll 1$. That is, the trend reduces the value of the PPI from the base, $\bar{P}$ (e.g., $PPI = \bar{P} \cdot 0.3$ when $\beta = -28.0$). If the trend is insignificant, i.e., $|\beta| \leq e$, then $\bar{P}$ itself will be the index. Note that with $|\beta| \leq e$, the difference, on average, between the numbers of publications in any two consecutive years during the selected period is less than three ($e < 3$).

Under this design, suppose we have two datasets, A and B, with corresponding indices PPI_A and PPI_B. If $PPI_A > PPI_B$, then either one of the following two statements is true:

- There are more publications that have analyzed dataset A during the selected period than that of dataset B, and the trend of analyzing dataset B does not show a significant advantage over that of dataset A. In other words, $\bar{P}_A > \bar{P}_B$ and $\beta_B \not\gg \beta_A$.

- The number of publications that have analyzed dataset A is no more than that of dataset B, but the trend of analyzing dataset A shows a significant advantage over that of dataset B. In other words, $\bar{P}_A \leq \bar{P}_B$ and $\beta_A \gg \beta_B$.

To calculate the PPI, we need to fit the simple linear regression model $y = \alpha + \beta x$ to estimate β (i.e., to obtain $\hat{\beta}$). Generally, we can use the least squares (LS) method to fit the model. Using this method, we assume that each of the calendar years has the same level of impact on popularity. However, some users may prefer to assign different weights to different years. It can be achieved by using the weighted least squares (WLS) method, which allows another user selection feature, the weight. For example, users can set the weights for recent five years as 1/15, 2/15, 3/15, 4/15, and 5/15. The latest year has the highest weight. As an advantage, weights allow the PPI calculated by $\hat{\beta}_{WLS}$ to pay more attention to more recent years so that it increases the accuracy of the measurement when users especially care about the timeliness of popularity. Without preferred weights, the WLS method is simply reduced to the LS method. The results and discussion in Section 3.4 further illustrate the advantages of using the WLS.

To summarize, the PPI is defined in (3.1). In addition, two user selection features are allowed: 1) N, the number of years that the PPI takes into account, and 2) weight, the level of the impact that each year makes on the PPI.

3.3 Data Source and Method to Identify Publicaitons

Our data source consists of 14 representative healthcare datasets. To implement the PPI on these datasets, we present a method to identify publications that have analyzed a specific dataset in this section.

3.3.1 Data Source

We selected healthcare datasets from two sources: 1) Three representatives of both public and proprietary datasets in the DIR: Healthcare Cost Utilization Project (HCUP), MarketScan, and Medical Information Mart for Intensive Care (MIMIC). 2) Thirteen representative datasets selected from working group notes discussed by domain experts at the UNC Charlotte Health Informatics and Outcomes Research

Academy [6]: Humedica NorthStar (Humedica), Cliniformatics Data Mart (Cliniformatics), MedMining, Premier Healthcare Database (Premier), Clinical Practice Research Datalink (CPRD), MarketScan, The Health Improvement Network (THIN), RealHealthData, Long-Term Care Minimum Data Set (MDS), HCUP, SEER-Medicare, Add Health, and NHANES. As MarketScan and HCUP were involved in both the DIR and the group notes, 14 representative datasets were finally involved (listed in Table 3.1).

3.3.2 Method to Identify Publicaitons

To identify publications that have analyzed a dataset, we used the PubMed search engine, which consisted of more than 27 million citations for biomedical literature from MEDLINE, life science journals, and books. We assumed that a publication utilized a dataset as the data source if and only if it mentioned at least one of the dataset keywords (e.g., full names and abbreviations) in either the title or the abstract. That is, we could make *"dataset keyword"[tiab]* (*[tiab]* was the same as *[Title/Abstract]*) queries to obtain all possible publications that have analyzed a dataset in PubMed.

However, identifying dataset keywords for searching was actually the most challenging step in the implementation because the search results were likely to encounter over- and under-matching. Over-matching meant that publications mentioning the keywords but not analyzing the corresponding dataset were returned in the search results. For example, when we queried *"CPRD"[tiab]* for the Clinical Practice Research Datalink, the results falsely included publications discussing the Chronic Parenchymal Renal Disease whose abbreviation was also CPRD. Under-matching happened when not all corresponding publications were returned, and it was usually because keywords were missing. For example, all possible keywords of MIMIC included not only the current full name (i.e., Medical Information Mart for Intensive Care) but also its previous full name (i.e., Multiparameter Intelligent Monitoring in Intensive Care).

Thus, we have created a generic method to identify keywords for each dataset to prevent over-matching and under-matching. Fig. 3.3 shows the workflow. Six remarks help readers understand the workflow.

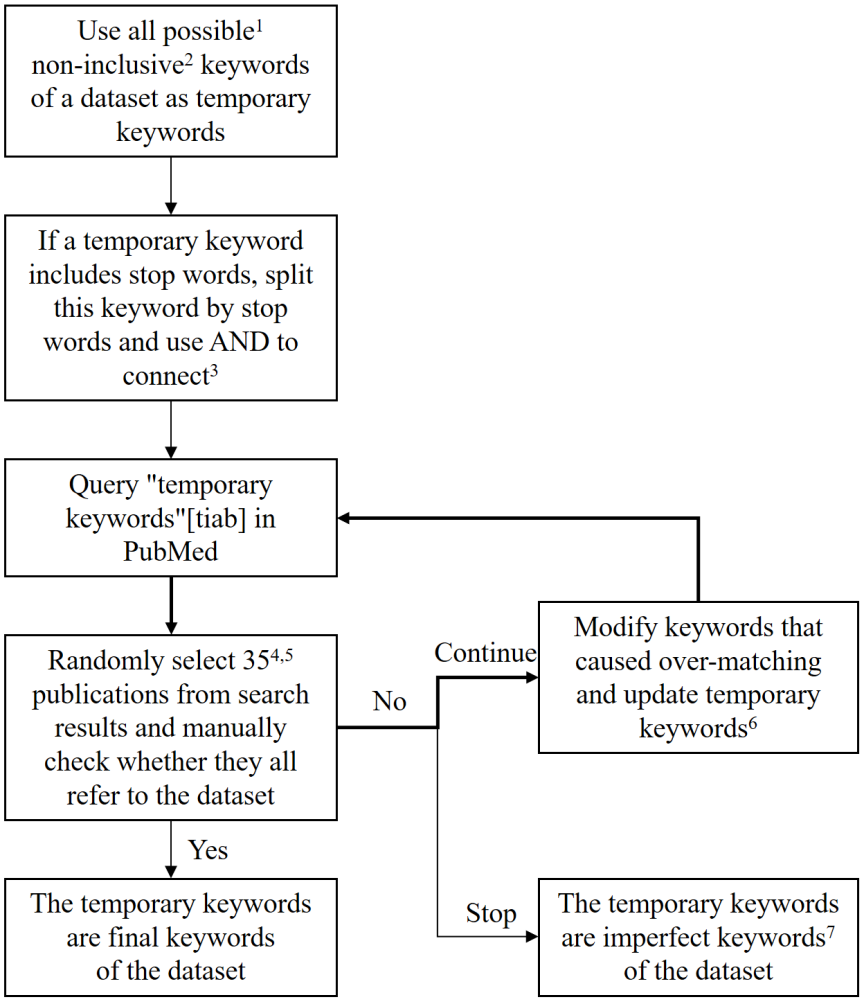

Figure 3.3: Method to identify dataset keywords. Superscripts refer to remarks.

- Remark 1: Using all "possible" keywords prevents under-matching.
- Remark 2: To clarify what "inclusive" means, let's take HCUP as an example. "HCUP" and "the HCUP database" are inclusive keywords because "HCUP" is contained in "the HCUP database." Only the shorter one (i.e., "HCUP") between them needs to be included when searching because of the absorption law (i.e., $A \wedge (A \vee B) \equiv A$).

- Remark 3: PubMed does not allow stop words when field descriptors (e.g., *[tiab]*) are used. A list of stop words for PubMed is in [7]. An example of splitting keywords is: *"Medical Information Mart for Intensive Care"[tiab]* becomes *("Medical Information Mart"[tiab] AND "Intensive Care"[tiab])*.

- Remark 4: By selecting 35, if all of the 35 publications refer to the dataset, the 95% confidence interval of the proportion of publications that refer to the dataset in the corresponding search results is [90%, 100%]. In other words, we are 95% confident that at least 90% of all the corresponding search results refer to the dataset.

  This number can be adjusted to obtain a different accuracy level. For example, if the number becomes 20 and all of the 20 publications refer to the dataset, then one will be 95% confident that at least 83.16% of the search results refer to the dataset. If the number is 50 and all of the 50 publications refer to the dataset, then one will be 95% confident that at least 92.89% of the search results refer to the dataset.

- Remark 5: If the total number of search results is less than 35, then one can manually check all the search results to obtain a precise accuracy level. For example, if there are only 24 search results under the temporary keywords, and all of them refer to the dataset, then the temporary keywords are the final keywords with 100% accuracy.

- Remark 6: An example of modifying keywords is: *"CPRD"[tiab]* becomes *"CPRD Database"[tiab]*.

- Remark 7: If one indeed cannot further improve the temporary keywords to pass the random selection test, the workflow can be stopped by producing imperfect keywords of the dataset. In such a situation, the accuracy level needs to be calculated accordingly by using exact binomial confidence intervals. For

example, if one randomly select 35 publications from the search results and find that 34 out of the 35 refer to the dataset, then one will be 95% confident that the proportion of publications that refer to the dataset in the search results is between 85.08% and 99.93%.

3.4 Results and Discussion

In this section, we demonstrate the utility of the PPI by ranking the 14 representative healthcare datasets and discuss the results.

Using the method in Fig. 3.3, final keywords of the 14 datasets are listed in Table 3.1.

Let $N = 5$ in the PPI. We obtained the numbers of publications from the years 2013 to 2017, respectively, by adding *AND ("201x/01/01"[EPDAT] : "201x/12/31"[EPDAT])* to the PubMed queries. For example, to obtain the number of publications that analyzed HCUP in 2013, we made the following query: *"HCUP"[tiab] OR "Healthcare Cost Utilization Project"[tiab] AND ("2013/01/01"[EPDAT] : "2013/12/31"[EPDAT])*. Here, *[EPDAT*] refers to electronic dates of publications, which are more timely than print dates (i.e., *[PPDAT]*).

There were two reasons that we set $N = 5$ instead of a larger number: 1) The PPI is the index of popularity. A dataset that was widely used years ago but has not been used often recently should not be defined as popular. That said, we should focus on only the information of recent years in terms of popularity. 2) A popular dataset can be new. It does not make sense to find numbers of publications over a long period because the majority of the numbers will be zero.

By querying PubMed, we obtained the data in Table 3.2.

Table 3.1: Final keywords of 14 representative datasets in PubMed queries.

Dataset	**Part of the PubMed Query**
HCUP	"HCUP"[tiab] OR "Healthcare Cost Utilization Project"[tiab]
MarketScan	"MarketScan"[tiab]
MIMIC	("Medical Information Mart"[tiab] AND "Intensive Care"[tiab]) OR ("Multiparameter Intelligent Monitoring"[tiab] AND "Intensive Care"[tiab]) OR "MIMIC-II"[tiab] OR "MIMIC-III"[tiab]
Humedica	"Humedica"[tiab]
Clinformatics	"Clinformatics"[tiab]
MedMining	"MedMining"[tiab]
Premier	"Premier Database"[tiab] OR "Premier Healthcare"[tiab]
CPRD	"CPRD Database"[tiab] OR "Clinical Practice Research Datalink"[tiab] OR "Clinical Practice Research Database"[tiab]
THIN	"THIN Database"[tiab] OR "Health Improvement Network"[tiab]
RealHealthData	"RealHealthData"[tiab]
MDS	"Long-Term Care Minimum Data Set"[tiab] OR "Minimum Data Set 2.0"[tiab] OR "Minimum Data Set 3.0"[tiab] OR "MDS 2.0"[tiab] OR "MDS 3.0"[tiab] OR ("Minimum Data Set"[tiab] AND ("Nursing Home"[tiab] OR "Nursing Homes"[tiab]))
SEER-Medicare	"SEER-Medicare"[tiab] OR ("Surveillance Epidemiology"[tiab] AND "End Results-Medicare"[tiab])
Add Health	("National Longitudinal Study"[tiab] AND "Adolescent"[tiab] AND "Adult Health"[tiab]) OR "Add Health"[tiab]
NHANES	("National Health"[tiab] AND "Nutrition Examination Survey"[tiab]) OR "NHANES"[tiab] NOT ("KOREA"[tiab] OR "KOREAN"[tiab])

Table 3.2: Numbers of publications of 14 representative datasets from 2013 to 2017[a].

Dataset	2013	2014	2015	2016	2017
HCUP	24	29	27	42	42
MarketScan	71	105	121	155	200
MIMIC	8	15	24	21	29
Humedica	0	2	8	2	8
Clinformatics	3	1	3	7	19
MedMining	0	1	1	0	0
Premier	5	8	5	14	25
CPRD	51	110	134	166	167
THIN	30	48	45	58	55
RealHealthData	0	0	0	0	0
MDS	41	32	44	38	49
SEER-Medicare	62	102	70	108	72
Add Health	45	40	77	107	93
NHANES	592	551	601	672	640

[a]Numbers were collected on 01/31/2018.

```
data=read.csv(file.choose(),header=T);
#Table II

x=c(2013, 2014, 2015, 2016, 2017);
#years in the selected period (user selection feature)

w=c(1, 2, 3, 4, 5);
#weights for WLS (user selection feature)

results=matrix(nrow=length(data[,1]),ncol=5);
rownames(results)=data[,1];
colnames(results)=c("Mean","BetaHatLS","BetaHatWLS","IndexLS",
"IndexWLS");
#define results matrix

for(i in 1:length(data[,1])){
    fitLS=lm(as.numeric(as.character(data[i,2:length(data[i,])]))~x);
    #run regression by least squares (LS) method

    betahatLS=summary(fitLS)$coefficients[2,1];
    #pick the LS estimate of beta from the summary

    exponentLS=log(abs(log(abs(betahatLS))))*sign(betahatLS);

    fitWLS=lm(as.numeric(as.character(data[i,2:length(data[i,])]))~x,
    weights=w);
    #run regression by weighted least squares (WLS) method

    betahatWLS=summary(fitWLS)$coefficients[2,1];
    #pick the WLS estimate of beta from the summary

    exponentWLS=log(abs(log(abs(betahatWLS))))*sign(betahatWLS);

    if(abs(betahatLS)<=exp(1)){betahatLS=0;exponentLS=0;}
    #judge if the slope estimated by LS is insignificant

    if(abs(betahatWLS)<=exp(1)){betahatWLS=0;exponentWLS=0;}
    #judge if the slope estimated by WLS is insignificant

    results[i,1]=mean(as.numeric(as.character(data[i,2:length(data[i,
    ])])));
    results[i,2]=betahatLS;
    results[i,3]=betahatWLS;
    results[i,4]=results[i,1]*exp(exponentLS);
    results[i,5]=results[i,1]*exp(exponentWLS);
}
results;
```

Figure 3.4: R code used to calculate PPIs and related attributes for 14 representative datasets.

PPIs for the 14 representative datasets have been calculated in R (code see Fig. 3.4) and summarized in Table 3.3. We have also developed an R package for easier usage and have released it on GitHub (https://github.com/JingyiShi/PPI). In Table 3.3, the datasets are ranked by PPI_{LS}. $\hat{\beta}_{\text{LS}}$s are estimated slopes calculated by the LS method, and $\hat{\beta}_{\text{WLS}}$s are estimated slopes calculated by the WLS method with the weights 1/15, 2/15, 3/15, 4/15, and 5/15 for the years 2013 to 2017. PPI_{LS}s are PPIs calculated using $\hat{\beta}_{\text{LS}}$s and PPI_{WLS}s are PPIs calculated using $\hat{\beta}_{\text{WLS}}$s. All numbers in this table have been rounded to one decimal digit.

Table 3.3: Dataset ranking by PPI_{LS} and related attributes.

Rank	**Dataset**	$\bar{\boldsymbol{P}}$	$\hat{\boldsymbol{\beta}}_{\text{LS}}$	$\hat{\boldsymbol{\beta}}_{\text{WLS}}$	$\textbf{PPI}_{\textbf{LS}}$	$\textbf{PPI}_{\text{WLS}}$
1	NHANES	611.2	21.7	23.4	**1880.9**	1926.2
2	MarketScan	130.4	30.8	32.5	**446.9**	454.0
3	CPRD	125.6	28.8	24.2	**422.1**	400.1
4	Add Health	72.4	16.3	15.2	**202.1**	197.2
5	THIN	47.2	6.0	4.9	**84.6**	74.9
6	SEER-Medicare	82.8	0.0	0.0	**82.8**	82.8
7	HCUP	32.8	4.9	5.2	**52.1**	54.1
8	MDS	40.8	0.0	3.1	**40.8**	46.7
9	MIMIC	19.4	4.8	4.4	**30.4**	28.6
10	Premier	11.4	4.6	5.8	**17.4**	20.0
11	Clinformatics	6.6	3.8	5.1	**8.8**	10.7
12	Humedica	4.0	0.0	0.0	**4.0**	4.0
13	MedMining	0.4	0.0	0.0	**0.4**	0.4
14	RealHealthData	0.0	0.0	0.0	**0.0**	0.0

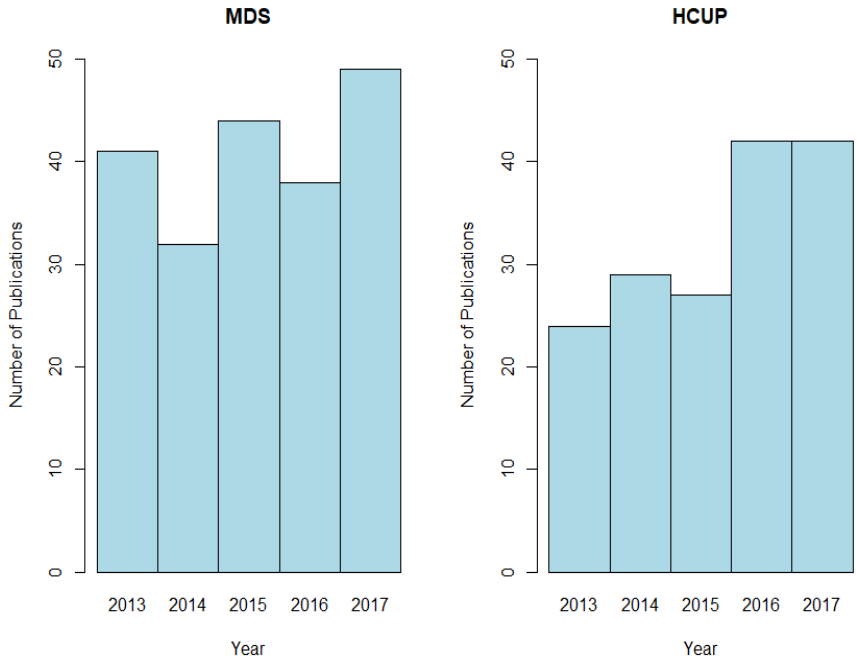

Figure 3.5: Numbers of publications that have analyzed MDS and HCUP from 2013 to 2017.

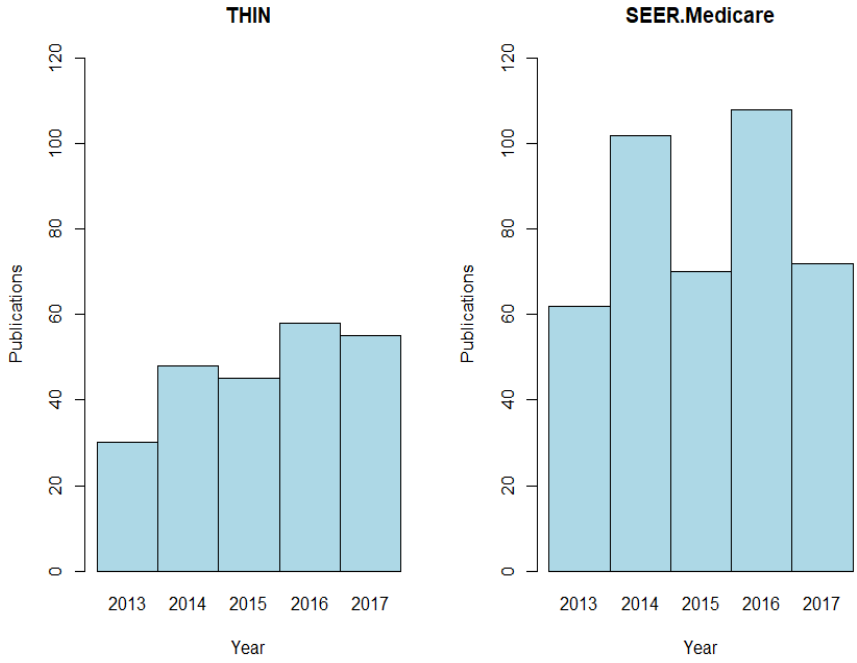

Figure 3.6: Numbers of publications that have analyzed THIN and SEER-Medicare from 2013 to 2017.

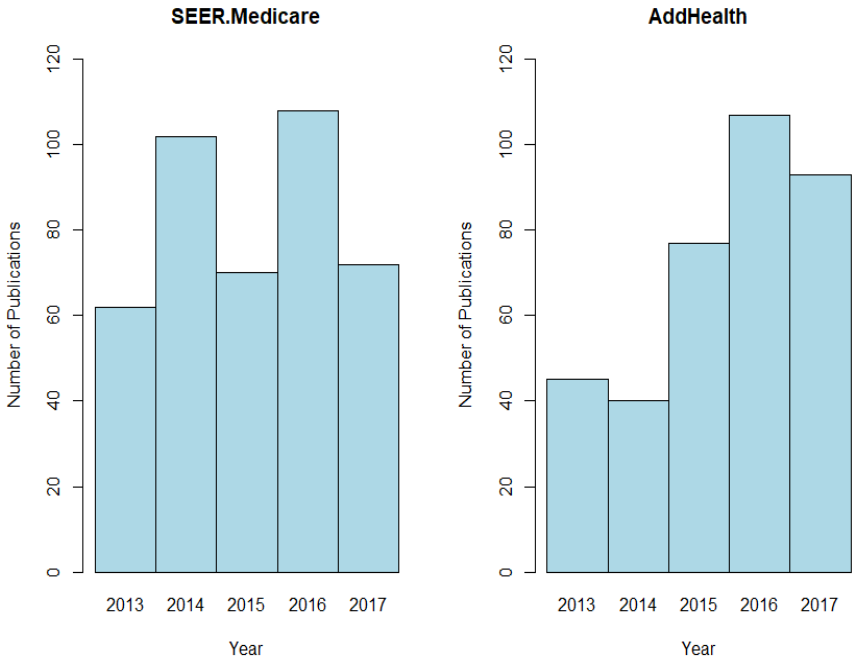

Figure 3.7: Numbers of publications that have analyzed SEER-Medicare and Add Health from 2013 to 2017.

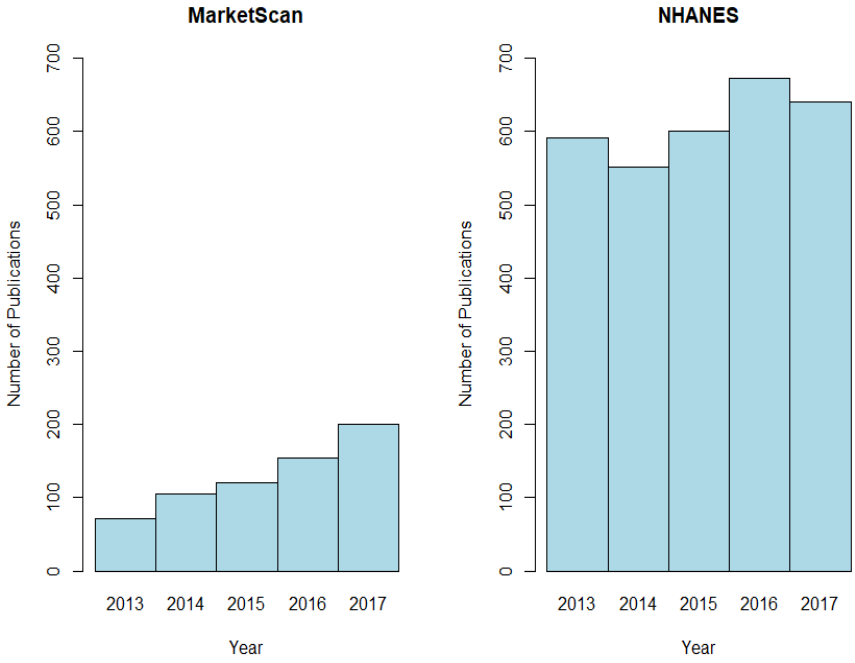

Figure 3.8: Numbers of publications that have analyzed MarketScan and NHANES from 2013 to 2017.

From Table 3.3, it is interesting to note that $\bar{P}$ of MDS is higher than that of HCUP; however, the PPI (using both methods) of MDS is lower than that of HCUP. This reveals the advantage of the PPI, which takes not only the number of publications but also the trend into account. Fig. 3.5 shows the number of publications that analyzed MDS and HCUP in each year from 2013 to 2017. As we can see in both Table 3.3 and Fig. 3.5, there is a much stronger increasing trend for HCUP than for MDS, while their average numbers of publications are close. Therefore, we believe that it is correct to conclude that HCUP has been more popular than MDS in recent years. In contrast, if one focuses on only the number of publications, it will lead to the opposite conclusion. The opposing results can also be found when comparing THIN to SEER-Medicare (see Fig. 3.6) and SEER-Medicare to Add Health (see Fig. 3.7). Further, it is interesting to compare MarketScan to NHANES. As we can see in Fig. 3.8, the increasing trend for MarketScan is much stronger than for NHANES. Nevertheless, the PPI results still indicate that NHANES is more popular. This is obviously true because the number of publications of NHANES dominates that of MarketScan. This reveals another advantage of the PPI—it takes the trend into account, but it balances so well that the number of publications still plays an important role.

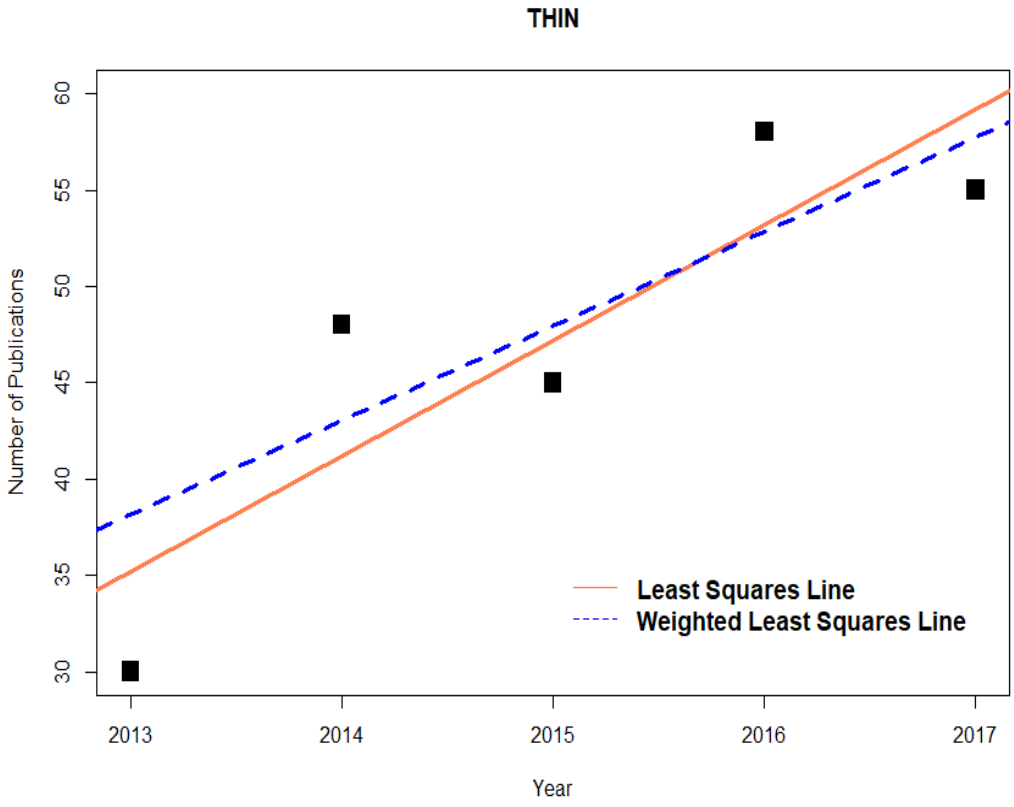

Figure 3.9: Comparison of LS and WLS methods on THIN.

Using the WLS versus the LS method can influence the rank to a certain extent. For example, it changes the ranks of THIN and SEER-Medicare in Table 3.3. That is, PPI_{LS} of THIN is higher than that of SEER-Medicare, but PPI_{WLS} of THIN is lower than that of SEER-Medicare. As $\hat{\beta}_{\text{LS}}$ of THIN is 6.0, and $\hat{\beta}_{\text{WLS}}$ of it is 4.9, the trend boosting for THIN is weaker when using the WLS method, which makes a difference in the ranking. Furthermore, Fig. 3.9 shows the advantage of using the WLS method in a clearer way. In Fig. 3.9, the slope of the WLS line decreases, compared to the LS line, because of the weak increment of the publication numbers of THIN in the most recent years. As we can see in Fig. 3.6, SEER-Medicare seems to be more popular because the increasing trend for both of them are not significant in the most recent years, but SEER-Medicare has a significantly higher number of publications.

Note that in Table 3.3, RealHealthData receives zero PPI values. According to

the design of the PPI, zero means the least popularity. Although there were some researchers analyzing RealHealthData, no related publications have been indexed by PubMed. Indeed, according to the RealHealthData website, all the publications it posted were abstracts and workshop discussions (as of 05/07/2018). The absence of conference and journal papers can indicate that no significant research has been conducted by analyzing RealHealthData. Hence, there was no evidence to verify the value of RealHealthData. In this regard, we believe that the PPI, as a measurement to indicate goodness, should be zero.

Table 3.4: Numbers of publications of 14 representative datasets in 2018[a].

Rank	**Dataset**	**$\mathbf{PPI_{LS}}$**	**Number of Publications in 2018[a]**
1	NHANES	1880.9	247
2	MarketScan	446.9	97
3	CPRD	422.1	67
4	Add Health	202.1	26
5	THIN	84.6	22
6	SEER-Medicare	82.8	32
7	HCUP	52.1	13
8	MDS	40.8	16
9	MIMIC	30.4	8
10	Premier	17.4	11
11	Clinformatics	8.8	11
12	Humedica	4.0	0
13	MedMining	0.4	0
14	RealHealthData	0.0	0

[a]As of 05/07/2018.

Although ranking methods are difficult to be systematically evaluated (e.g., PageRank [8] and H-index [9]), we still attempt to show the overall performance of the PPI to a certain extent. We have retrieved the current (as of 05/07/2018) number of publications that have analyzed each of the 14 representative datasets in 2018. These data are summarized in Table 3.4, where the 14 representative datasets are ranked by their corresponding PPIs from Table 3.3. The results in Table 3.4 indicate that the

PPI provides a legitimate ranking in terms of popularity because the ranking by the number of publications in 2018 is almost identical to that by the PPI. As can be seen, there is a surge in the SEER-Medicare. From its publication numbers in Table 3.2 and Table 3.4, it is interesting to notice that there is a plausible periodic trend that the number of publications surges every other year (in 2014, 2016, and 2018 (to-date)). However, we did not use time series analysis because it may cause an overfitting issue. Additionally, the rankings for MDS and MIMIC are slightly different. Nevertheless, this is possible because the PPI provides a legitimate ranking of popularity instead of a definite ranking of the future number of publications.

We note that the current method for identifying publications that have analyzed a dataset can be further improved. We intend to leverage text mining and machine learning methods to reduce manual work in future work. In addition, other reasons rather than goodness might affect popularity of a dataset. For example, a dataset that covers a more prevalent disease is probably used more often. Therefore, an analysis of dataset coverage is encouraged in the future to help researchers understand features of popularity.

3.5 Conclusions and Future Work

We have developed the Publication-based Popularity Index (PPI), which provides an overall quantitative ranking of all types of healthcare datasets. The PPI incorporates the quantity of successful analyses and the trend simultaneously. According to the design of the PPI, users can easily identify the high-quality datasets with outstanding research value and make an objective comparison among similar datasets. We evaluated and ranked 14 representative healthcare datasets based on the PPI, and the final results were promising. We believe that the PPI provides one important measurement of the value of healthcare datasets that is currently lacking. In the future, the PPI can be used to provide an overall ranking of all healthcare datasets for all types of dataset integration systems to sort search results. This ranking can also be

used as a starting point for identifying and examining the design of the most valuable healthcare datasets so that features of these datasets can inform future designs.

Chapter 3 Reference List

[1] J. Shi, M. Zheng, L. Yao, and Y. Ge, "DIR - A semantic information resource for healthcare datasets," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 805–810, Nov. 2017.

[2] M. G. Kahn, T. J. Callahan, J. Barnard, A. E. Bauck, J. Brown, B. N. Davidson, H. Estiri, C. Goerg, E. Holve, S. G. Johnson, S.-T. Liaw, M. Hamilton-Lopez, D. Meeker, T. C. Ong, P. Ryan, N. Shang, N. G. Weiskopf, C. Weng, M. N. Zozus, and L. Schilling, "A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data," *eGEMs*, vol. 4, Sept. 2016.

[3] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research," *Journal of the American Medical Informatics Association*, vol. 20, pp. 144–151, Jan. 2013.

[4] N. G. Weiskopf, G. Hripcsak, S. Swaminathan, and C. Weng, "Defining and measuring completeness of electronic health records for secondary use," *Journal of Biomedical Informatics*, vol. 46, pp. 830–836, Oct. 2013.

[5] Shelli L. Feder, "Data Quality in Electronic Health Records Research: Quality Domains and Assessment Methods," *Western Journal of Nursing Research*, Jan. 2017.

[6] "Health Informatics and Outcomes Research Academy | UNC Charlotte." `https://hinora.uncc.edu/`.

[7] "PubMed Help [Table, Stopwords]." `https://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.stopwords/`.

[8] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web.," tech. rep., Stanford InfoLab, 1999.

[9] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences*, vol. 102, pp. 16569–16572, Nov. 2005.

CHAPTER 4: AN ENTROPIC FEATURE SELECTION METHOD IN PERSPECTIVE OF TURING'S FORMULA

4.1 Introduction

Inspired by the recent advancement in Big Data, health informaticians are attempting to assist health care providers and patients from a data perspective, with the hope of improving quality of care, detecting diseases earlier, enhancing decision making, and reducing healthcare costs [1]. In the process, health informaticians have been confronted with the issue of generalization [2]. Analyzing real health data involves many practical problems that could contribute to the issue of generalization; for example, the unknown amount of information (signal) versus error (noise), the curse of dimensionality, and the generalizability of models. All these trivial problems boil down to the essential problem issued by a limited sample. With the limitation of the sample size, the information from the sample cannot represent the information of the population to a desirable extent. For this reason, a simple way to address these trivial problems is to collect a sufficiently large sample, which is unfortunately often impractical in healthcare because of multiple reasons. For example:

1. The term sufficiently large is relative to the dimensionality of data and the complexity of feature spaces. Health data are generally large in dimensionality, particularly when dummy variables (one-hot-encoding) are adopted to represent enormous categories of complex qualitative features (such as extracted words from clinical notes). As a result, a dataset with a sample size of 1,000,000 may not be sufficient, depending on its feature spaces.

2. There may not be sufficient patient cases for a rare disease. Even if there are

ample potential cases, it may be cost-prohibitive for clinical trials to achieve a sufficient sample.

Without a sufficiently large sample, dimension reduction becomes a major research direction in health data analytic as reducing the dimensionality can partly relieve the issues from a limited sample. These dimension reduction techniques mainly focus on feature selection and feature projection, where feature selection can be further applied to the features created by feature projection. In this article, we focus on feature selection. It has become an important research area, dating back at least to 1997 [3][4]. Since then, many feature selection methods have been proposed and well discussed in multiple recent review papers, such as [5], [6], and [7]. To apply these feature selection methods to health data, domain-specific challenges must be considered.

Health data can be numerical and categorical. For example, many machine readings (*e.g.*, heart rate, blood pressure, and blood oxygen level) are numerical, while gene expression data are categorical. A healthcare dataset could contain numerical data only, categorical data only, or a combination of both data types. The fundamental distinction between numerical data and categorical data is whether the data space is ordinal or non-ordinal. As a result, data consisting of only numbers are not necessarily numerical data; for example, gene expression data can be coded to numbers using dummy variables, but it should be still considered as categorical. When the data space is ordinal (numerical data only), classical methods—which detect the association using ordinal information—are more powerful in capturing the associations in data. When the data space is non-ordinal (categorical data only), ordinal information does not naturally exist; hence, continuing to use classical methods onto coded data loses their original advantages and has additional estimation issues. Namely, involving dummy variables increases the dimensionality of data and further exacerbates the estimation problem using a limited sample. This particularly happens when

an involved categorical feature has a complex feature space that requires a tremendous number of dummy variables to represent all the different categories. To deal with the categorical data, only information-theoretic quantities (*e.g.*, entropy and mutual information [8]) serve the purpose. When a dataset is a combination of both data types, it is inconclusive about whether to use classical or information-theoretic methods. In general, if one believes that the numerical data in the dataset carry more information than the categorical data, then classical methods can be used. If one believes the categorical data carry more information, then information-theoretic methods should be used, and the numerical data should be binned to categorical data. One should be advised that coding categorical data for classical methods increases dimensionality and issues more difficulties in estimation, while binning numerical data for information-theoretic methods inevitably loses ordinal information. It should also be noted that, although ordinal information could provide extra information about associations among the data, the ordinal information could also mislead a person's judgment when associations actually exist, but there is no visual pattern among the data. The way that classical methods work is very similar to our visualization; if there is a pattern that can be visually observed, then it can also be detected by some classical methods. However, not all associations among numerical data are visually observable, in which case, classical methods would fail to detect the associations. On the other hand, if there is a visual pattern among data, binning the data (losing the ordinal information) would not necessarily lead to a loss of associations among data; it depends on the binning methods and performance of the information-theoretic methods.

Classical feature selection methods include, but are not limited to, Fisher Score [9], ReliefF [10], Trace Ratio [11], Laplacian Score [12], SPEC [13], l_p-regularized [14], $l_{p,q}$-regularized [14], Efficient and Robust Feature Selection (REFS) [15], Multi-Cluster Feature Selection (MCFS) [16], Unsupervised Feature Selection Algorithm

(UDFS) [17], Nonnegative Discriminative Feature Selection (NDFS) [18], T-score [19], and LASSO [20]. All these classical feature selection methods require information from ordinal spaces, such as moments (*e.g.*, mean and variance) and spacial information (*e.g.*, nearest location and norms). Information-theoretic feature selection methods include, but are not limit to, Mutual Information Maximisation (MIM) [21], Mutual Information Feature Selection (MIFS) [22], Joint Mutual Information (JMI) [23], minimal Conditional Mutual Information Maximisation (CMIM) [24][25], Minimum Redundancy Maximal Relevancy (MRMR) [26], Conditional Infomax Feature Extraction (CIFE) [27], Informative Fragments (IF) [24], Double Input Symmetrical Relevance (DISR) [28], minimal Normalised Joint Mutual Information Maximisation (NJMIM) [29], Chi-square Score [30], Gini Index [31], and CFS [32]. All these information-theoretic methods use ordered probabilities, which always exist in non-ordinal spaces. For example, frequencies, category probabilities (proportions), Shannon's entropy, mutual information, and symmetric uncertainty are all functions of ordered probabilities.

In many cases, all (or most) of the data in a healthcare dataset could be categorical. To analyze the categorical data in such a dataset, information-theoretic feature selection methods are preferred because they could capture the associations among features without using dummy variables, where classical methods require dummy variables that would increase the dimensionality. Most existing information-theoretic methods use entropy or mutual information (a function of entropy) to measure associations among data. Information-theoretic methods that do not use entropy include Gini Index and Chi-square Score. Gini Index focuses on whether a feature is separative, but does not indicate probabilistic associations. Chi-square Score relies on the performance of asymptotic normality on each component, and when there are categories with low frequencies (*e.g.*, less than 5), the Chi-square Score is very unstable. However, under a limited sample, we should expect at least a few, if not many, cat-

egories would have relatively low frequencies. For the existing information-theoretic methods that use entropy (we call these *entropic methods*), all of them estimate entropy with the classical maximum likelihood estimator (the plug-in estimator). The plug-in entropy estimator performs very poorly when the sample size is not sufficiently large [33][34], and we have discussed that the sample size is usually relatively limited in healthcare datasets. As a result, to use entropic methods in healthcare data analytics, the estimation of entropy under small samples must be improved.

In addition to estimation based on small samples, the unhelpful association is another issue with these samples. While the issue of estimation can be addressed by using a better estimator, the problem of unhelpful association is trickier. The unhelpful association is partially a result of sample randomness, and it could be severe when the sample size is small. Suppose there is a healthcare dataset with multiple features and one outcome, and there is a feature in the dataset that could distinguish the values of the outcome based on the sample information, then there are three possible situations:

Situation 1. The feature has abundant real information toward the outcome, and the real information is well preserved by the sample data.

Situation 2. The feature has abundant real information toward the outcome, but the real information is not well preserved by the sample data.

Situation 3. The feature has little real information but seems relevant to the outcome because of randomness in the sample.

The term *real information* of a feature means the feature-carried information that could indicate the values of the outcome at the population level. All three situations are conceptual classifications. At the population level, situation 1 and 2 features are relevant features, and situation 3 features are irrelevant features. It is clear that situation 1 features should be selected while situation 3 features should be dropped. For situation 2, caution should be exercised. Intuitively, situation 2 features should

be kept as they are relevant features at the population level. However, as a result of a limited sample, the information carried by these situation 2 features are very subtle. There are at least two constitutional problems about the information from situation 2 features. First, although the feature could distinguish the values of the outcome based on the sample information, the sample-preserved information possibly provides only a meager coverage of all the possible values of the feature. As a result, when there is a new observation (*e.g.*, a new patient), it is very likely that the new observation's corresponding label has not been observed by the preserved information, in which case no outcome information is available to assist prediction based on the information of such a situation 2 feature. Second, because of the limited sample, the predictability of the situation 2 features revealed by the sample may not be complete; hence, it could contribute as an (a) error (noise). For example, based on the sample information, different values of a situation 2 feature could possibly uniquely determine a corresponding value of the outcome (particularly when a feature space is complex while the sample size is small), but this deterministic relationship revealed by a limited sample is unlikely to be true at the population level. As a result, using this information in further modelling and prediction would be wrong and could further contribute to the issue of generalization. Therefore, we suggest omitting situation 2 features. In addition, one should note that a relevant feature being categorized as situation 2 is a consequence of a limited sample. All situation 2 features would eventually become situation 1 when the sample size grows (because more real information would be revealed). As a summary, under a limited sample, situation 1 features should be kept, and situation 2 and 3 features should be dropped.

Focusing on the domain-specific challenges from health data, we develop the proposed entropic feature selection method based on the concept of Coverage Adjusted Standardized Mutual Information (CASMI). The proposed method aims at improving the performance of estimation and addressing the issue of unhelpful association

under relatively small samples. The rest of the article is organized as follows. The concept, intuition, and estimation of CASMI are discussed in Section 4.2. The proposed method is described in detail in Section 4.3 and evaluated by a simulation study in Section 4.4. A brief discussion is in Section 4.5.

4.2 CASMI and its Estimation

In this section, we introduce the concept, intuition, and estimation of CASMI. Before we proceed, let us state the notations first.

Let $\mathscr{X} = \{x_i; i = 1, \cdots, K_1\}$ and $\mathscr{Y} = \{y_j; j = 1, \cdots, K_2\}$ be two finite alphabets with cardinalities $K_1 < \infty$ and $K_2 < \infty$, respectively. Consider the Cartesian product $\mathscr{X} \times \mathscr{Y}$ with a joint probability distribution $\mathbf{p} = \{p_{i,j}\}$. Let the two marginal distributions be respectively denoted by $\mathbf{p}_x = \{p_{i,\cdot}\}$ and $\mathbf{p}_y = \{p_{\cdot,j}\}$, where $p_{i,\cdot} = \sum_j p_{i,j}$ and $p_{\cdot,j} = \sum_i p_{i,j}$. Assume that $p_{i,\cdot} > 0$ and $p_{\cdot,j} > 0$ for all $1 \le i \le K_1$ and $1 \le j \le K_2$ and that there are $K = \sum_{i,j} 1[p_{i,j} > 0]$ non-zero entries in $\{p_{i,j}\}$. We re-enumerate these K positive probabilities in one sequence and denote it as $\{p_k; k = 1, \cdots, K\}$. Let X and Y be random variables following distributions $\mathbf{p}_x$ and $\mathbf{p}_y$, respectively. For every pair of i and j, let $f_{i,j}$ be the observed frequency of the random pair (X, Y) taking value (x_i, y_j), where $i = 1, \cdots, K_1$ and $j = 1, \cdots, K_2$, in an *iid* sample of size n from $\mathscr{X} \times \mathscr{Y}$ under $\mathbf{p}$, and let $\hat{p}_{i,j} = f_{i,j}/n$ be the corresponding relative frequency. Consequently, we write $\hat{\mathbf{p}} = \{\hat{p}_k\}$ (*i.e.*, $\{\hat{p}_{i,j}\}$), $\hat{\mathbf{p}}_x = \{\hat{p}_{i,\cdot}\}$, and $\hat{\mathbf{p}}_y = \{\hat{p}_{\cdot,j}\}$ as the sets of observed joint and marginal relative frequencies. Shannon's mutual information between X and Y is defined as

$$MI(X, Y) = H(X) + H(Y) - H(X, Y), \tag{4.1}$$

where

$$H(X) = -\sum_i p_{i,\cdot} \ln p_{i,\cdot},$$
$$H(Y) = -\sum_j p_{\cdot,j} \ln p_{\cdot,j},$$
$$H(X,Y) = -\sum_i \sum_j p_{i,j} \ln p_{i,j} = -\sum_{k=1}^{K} p_k \ln p_k.$$

We define the CASMI as follows:

Definition (CASMI)**.** *κ^*, the Coverage Adjusted Standardized Mutual Information (CASMI) of a feature X to an outcome Y, is defined as*

$$\kappa^*(X,Y) = \kappa(X,Y) \cdot (1 - \pi_0(X)), \tag{4.2}$$

where

$$\kappa(X,Y) = \frac{MI(X,Y)}{H(Y)}, \tag{4.3}$$

and $(1 - \pi_0)$ is the sample coverage that was first introduced by Good [35] as "the proportion of the population represented by (the species occurring in) the sample".

4.2.1 Intuition of CASMI

Many entropic concepts can measure the associations among non-ordinal data; for example, mutual information (MI), Kullback-Leibler divergence ([36]), conditional mutual information ([37]), and weighted variants ([38]). Among them, MI is the fundamental concept as all the other entropic association measurements are developed based on or equivalent to MI. For this reason, we develop the CASMI starting with MI. It is well known that $MI \geq 0$, and $MI(X,Y) = 0$ if and only if X and Y are independent. However, MI is not bounded from above; hence, using the values of MI to compare the degrees of dependence among different pairs of random variables is

inconvenient. Therefore, it is necessary to standardize the mutual information, which yields to the so-called standardized mutual information (SMI) or normalized variants. [39] provides several forms of SMI, such as $MI/H(X)$ (also known as information gain ratio if X is a feature and Y is the outcome), $MI/H(Y)$, and $MI/H(X,Y)$. All these forms of SMI can be proven to be bounded by $[0,1]$, where 0 stands for independence between X and Y, and 1 stands differently for different SMIs. For $MI/H(X)$ (information gain ratio), 1 means that, given the value of Y (outcome), the value of X (feature) is determinate. For $MI/H(Y)$, 1 means that, given the value of X, the value of Y is determinate. For $MI/H(X,Y)$, 1 means a one-to-one correspondence between X and Y.

The goal of feature selection is to separate the predictive features from non-predictive features. In this regard, $MI/H(Y) = 1$ is most desirable because $MI/H(X) = 1$ does not indicate the predictability of X and $MI/H(X,Y) = 1$ is too strong and unnecessary. Therefore, we select κ in (4.3) as the SMI in CASMI.

As we have discussed, detecting unhelpful associations under small samples is important in health data analytics as involving unhelpful associations would bring too much noise or unnecessary dimensions to model-building or prediction. In other words, we would like to detect situation 2 and 3 features in a limited sample. The common characteristics among situation 2 and 3 features is the information revealed by the limited sample covers little of the total information in the population. For this reason, we can use sample coverage $(1 - \pi_0)$, the concept introduced by Good, to detect these features. A feature with high predictability but low sample coverage must belong to either situation 2 or 3. In CASMI, we multiply the SMI by the sample coverage. Under this setting, although features from situations 2 and 3 have high SMI values, their CASMI scores would be low because of their low sample coverages; hence, these features would not be selected in a greedy selection. On the other hand, the CASMI score for a situation 1 feature would be high because both SMI and

the sample coverage are high. As a result, by selecting features greedily, situation 1 features would be selected, while situation 2 and 3 features would be dropped.

The purpose of CASMI is to capture the association between a feature and the outcome, with a penalized term from the sample coverage, so that features under situations 2 and 3 would be eliminated. By selecting features under only situation 1, the issue of generalization under small samples is expected to be reduced. (See Section 4.3 for a discussion on feature redundancy (or feature interaction).)

It may be interesting to note that the CASMI is an information-theoretic quantity that is related to both the population and the sample. It is neither a parameter nor a statistic, and it is only observable when both the population and the sample are known. Next, we introduce its estimation.

4.2.2 Estimation

To estimate κ^* (CASMI), we need to estimate π_0 and κ. $\pi_0(X)$ can be estimated by Turing's formula [35]

$$T_1(X) = N_1(X)/n, \tag{4.4}$$

where $N_1(X)$ is the number of singletons in the sample. For example, if a sample of English letters consists of $\{a, a, a, b, c, c, d, e, e, f\}$, then the corresponding $N_1 = 3$ (b, d, and f are the three singletons). Discussions on the performance of estimating π_0 by T_1 can be found in [39] and [40]. In experimental categorical data, singletons could possibly indicate the sample size is small. As the sample size grows, the chance of obtaining a singleton in the sample approaches zero. It may be interesting to note that using (5.4) to estimate the sample coverage would automatically separates ID-like features. This is because an ID-like feature is naturally all (or almost all) singletons and would result in a zero (or very small) estimated sample coverage that further leads to a zero (or very low) CASMI score; hence, such an ID-like feature would not be selected.

Estimating $\kappa(X,Y)$ is equivalent to estimating $MI(X,Y)$ and $H(Y)$. As we have discussed, thus far, all the existing entropic information-theoretic methods use the plug-in estimator of entropy ($\hat{H}$). However, the plug-in entropy estimator has a huge bias, particularly when sample size is small. [33] showed that the bias of $\hat{H}$ is

$$\mathrm{E}(\hat{H}) - H = -\frac{K-1}{2n} + \frac{1}{12n^2}\left(1 - \sum_{k=1}^{K}\frac{1}{p_k}\right) + \mathcal{O}\left(n^{-3}\right),^{1}$$

where n is the sample size and K is the cardinality of the space on which the probability distribution $\{p_k\}$ lives. Based on the expressions of the bias, it is easy to see that the plug-in estimator underestimates the real entropy, and the bias approaches 0 as n (sample size) approaches infinity, with a rate of n^{-1} (power decay). Because of the power decaying rate, the bias is not small when sample size (n) is relatively low.

To improve the estimation under a small sample, we adopt the following $\hat{H}_z$ [41] as the estimator of H:

$$\hat{H}_z = \sum_{v=1}^{n-1}\frac{1}{v}\frac{n^{1+v}[n-(1+v)]!}{n!}\sum_{k}\left[\hat{p}_k\prod_{j=0}^{v-1}\left(1-\hat{p}_k-\frac{j}{n}\right)\right]. \tag{4.5}$$

Compared to the power decaying bias of $\hat{H}$, $\hat{H}_z$ has an exponentially decaying bias

$$\mathrm{E}(\hat{H}_z) - H = \mathcal{O}\left(\frac{(1-p_\wedge)^n}{n}\right),$$

where $p_\wedge = \min\{p_k > 0\}$.

To help understand the differences between the power decaying bias and exponentially decaying bias, we conduct a simulation. In the simulation, the real underlying distribution is $p_k = k/2001000$, where $k = 1, 2, \ldots, 2000$ (*i.e.*, a triangle distribution). Under this setting, the true entropy $H = 7.408005$. To compare the two estimators, we independently generate 10,000 samples following the triangle distribution for each

[1] We write $f = \mathcal{O}(g(n))$ to denote $\limsup_{n\to\infty}|f(n)/g(n)| < \infty$.

of the six sample size settings (*i.e.*, we generate 60,000 random samples in total). The average values of $\hat{H}$ and $\hat{H}_z$ under different sample sizes are summarized in Table 4.1.

Table 4.1: Estimation comparison between $\hat{H}$ and $\hat{H}_z$.

n	100	300	500	1000	1500	2000
avg. of $\hat{H}$	4.56	5.57	6.00	6.51	6.75	6.89
avg. of $\hat{H}_z$	5.11	6.09	6.49	6.92	7.11	7.21

The calculation shows that $\hat{H}$ would consistently underestimate H more than $\hat{H}_z$. The underestimation is more severe when the sample size is smaller. Therefore, from a theoretical perspective, we expect adopting $\hat{H}_z$ in estimating the entropies in CASMI would provide a better estimation, particularly under small samples. Furthermore, we expect CASMI would capture the associations among features and the outcome more accurately under small samples because of the improvement in estimation. Interested readers can find additional discussions on comparison among more entropy estimators in [41], and comparison about mutual information estimators using $\hat{H}$ and $\hat{H}_z$ in [42].

Consequently, we let

$$\widehat{MI}_z(X,Y) = \hat{H}_z(X) + \hat{H}_z(Y) - \hat{H}_z(X,Y), \tag{4.6}$$

and we estimate κ as

$$\hat{\kappa}_z(X,Y) = \frac{\widehat{MI}_z(X,Y)}{\hat{H}_z(Y)}. \tag{4.7}$$

As a summary, we estimate κ^* by the following estimator, which is the scoring function of the selection stage in the proposed method.

$$\hat{\kappa}^*(X,Y) = \hat{\kappa}_z(X,Y) \cdot (1 - T_1(X)), \tag{4.8}$$

where $\hat{\kappa}_z$ is defined in (4.7) and T_1 is defined in (5.4). $\hat{\kappa}^*$ adopts an entropy estimator

with an exponentially decaying bias to improve the performance in estimating κ^* and capturing the associations when the sample size is not sufficiently large. Furthermore, we expect involving the sample coverage would separate and drop situation 2 and 3 features under small samples.

4.3 CASMI Based Feature Selection Method

In this section, we introduce the proposed feature selection method in detail. The proposed method contains two stages. Before we present the two stages, let us first discuss data preprocessing.

4.3.1 Data preprocessing

To use the proposed method, all features and the outcome data must be preprocessed to categorical data. Continuous numerical data must be discretized, and there are numerous discretization methods [43]. While binning continuous features, the estimated sample coverage (5.4) should be checked to avoid over-discretization, which increases the risk of wrongly shifting a feature from situation 1 to situation 2.

If the data are already categorical, one may need to combine some of the categories to improve the sample coverage, when necessary. When most observations of a feature are singletons, then the coverage is close to 0, in which case it is difficult to draw any reliable and generalizable statistical inference. Therefore, for features that may carry real information but have low sample coverages (below 50%), it is suggested to regroup them to create repeats and improve coverages. Note that not all features are worth regrouping; for example, if a feature is the IDs of patients, regrouping should be avoided as there is no reason to believe an ID can contribute to the outcome. The proposed method does not select features with low sample coverages; hence, ID-like features are eliminated automatically.

When a feature contains missing (or invalid) data that cannot be recovered by the data collector, without deleting the feature, there are several possible remedies,

such as deleting the observation, making an educated guess, predicting the missing values, and listing all missing values as NA. While it is the user's preference on how to handle the missing data, one should be advised that manipulating (guessing or predicting) the missing data could create (or enhance) false associations; therefore, one should be cautious. Assigning all the missing values as NA generally would not create false associations, but it may reduce the predictive information of the feature. The performance of each remedy method could vary from situation to situation. Additional discussions on handling missing data can be found in [44], [45], and [46]. We suggest dealing with the missing data at the beginning of the data preprocessing.

The processed data should contain only categorical features and outcome(s). A feature with only integer values could be considered as categorical as long as the sample coverage is satisfactory.

4.3.2 Stage 1: Eliminate Independent Features

In this stage, we eliminate the features that are believed to be independent of the outcome based on a statistical test. This step filters out the features that are very unlikely to be useful; hence, the computation time for feature selection is reduced.

Suppose there are p features, $X_1, X_2, \ldots, X_p$, and one outcome, Y, in a dataset. Note that there could be multiple outcome attributes in a dataset. Because each outcome attribute has its own related features, when making a feature selection, we consider one outcome attribute at a time.

In finding independent features, we adopt a chi-squared test of independence using $\widehat{MI}_z$ as the statistic.

Theorem 1. *[47] Provided that $MI = 0$,*

$$\chi^2 = 2n\widehat{MI}_z + (K_1 - 1)(K_2 - 1) \xrightarrow{L} \chi^2\left((K_1 - 1)(K_2 - 1)\right), \tag{4.9}$$

where $\widehat{MI}_z$ is defined in (5.3). K_1 and K_2 are the effective cardinalities of the selected

feature X and the outcome Y, respectively.[2]

Compared to Pearson's chi-squared test of independence, testing independence using Theorem 1 has more statistical power, particularly when the sample size is small [47]. We test hypothesis $H_0 : MI(X, Y) = 0$ against $H_a : MI(X, Y) > 0$ between the outcome and each of the features. At a user-chosen level of significance (α), any feature whose test decision fails to reject H_0 is eliminated at this stage. It is suggested to let $\alpha = 0.10$. A smaller α increases the chance of Type-II error (eliminating useful features); a larger α reduces the ability of the elimination, which results in a longer selection computation time in the next stage.

Let $X_1, X_2, \ldots, X_s$ denote the s features (out of the p features) that have passed the test of independence. The other $(p - s)$ features are eliminated at this stage. Note that the $X_1, \ldots, X_s$ are temporary notations for features. Namely, the X_1 in $\{X_1, \ldots, X_p\} := \{X\}_p$ and the X_1 in $\{X_1, \ldots, X_s\} := \{X\}_s$ are different if the X_1 in $\{X\}_p$ is eliminated in this stage. Note that we do not consider feature redundancy at Stage 1. Redundant features could all pass the test of independence as long as they appear to be relevant to the outcome based on sample data. Feature redundancy would be considered at Stage 2.

4.3.3 Stage 2: Selection

In this stage, we make a greedy selection among the s remaining features from Stage 1.

The selection algorithm is:

1. $X_{(1)} = \underset{X_i \in \{X\}_s}{\arg\max} [\hat{\kappa}^*(X_i, Y)]$;
2. $X_{(2)} = \underset{X_i \in \{X\}_s \setminus \{X_{(1)}\}}{\arg\max} \left[\hat{\kappa}^*(X_{(1)} \times X_i, Y)\right]$;
3. $X_{(3)} = \underset{X_i \in \{X\}_s \setminus \{X_{(1)}, X_{(2)}\}}{\arg\max} \left[\hat{\kappa}^*(X_{(1)} \times X_{(2)} \times X_i, Y)\right]$;

   ...

[2] We write $\xrightarrow{L}$ to denote convergence in distribution.

The algorithm stops at time c when $\hat{\kappa}^*(X_{(1)} \times \cdots \times X_{(c+1)}, Y) < \hat{\kappa}^*(X_{(1)} \times \cdots \times X_{(c)}, Y)$.

To clarify the notations, $\hat{\kappa}^*(X_{(1)} \times X_i, Y)$ stands for the estimated CASMI of the joint feature $X_{(1)} \times X_i$ to the outcome Y, and $\{X\}_s$ is the collection of the s remaining features.

The proposed method handles feature redundancy by considering joint-distributions among features. Taking $X_{(1)}$ and $X_{(2)}$ as examples, the first step yields the feature $X_{(1)}$, which is the most relevant feature (measured by the estimated CASMI) to the outcome. In the second step, we joint the selected $X_{(1)}$ with each of the remaining $(s-1)$ features, and we evaluate the estimated CASMIs between each of the joint-features and the outcome. The joint-feature with the highest estimated CASMI is selected, which becomes $X_{(2)}$. It should be noted that $X_{(1)}$ and $X_{(2)}$ are neither necessarily independent nor necessarily the least dependent. Selecting $X_{(2)}$ only indicates that based on the information provided from $X_{(1)}$, $X_{(2)}$ provides the most additional information about the outcome among the remaining $(s-1)$ features. In addition, CASMI is an information-theoretic quantity that does not use ordinal information of features; therefore, both linear and nonlinear redundancy are captured, evaluated, and considered.

The proposed algorithm stops when the term $\max[\hat{\kappa}^*(\cdot, Y)]$ starts to decrease. The features selected by the proposed method are $X_{(1)}, \ldots, X_{(c)}$.

In some situations, a researcher may want to select a desired number of features (d) that is different from c. For example, let $c = 10$, $d_1 = 6$, and $d_2 = 15$. When $c = 10$ and $d_1 = 6$, because $6 \leq 10$, we can stop the algorithm at the time 6. When $c = 10$ and $d_2 = 15$, because $15 > 10$, the user needs to select 5 additional features. We propose two choices on how to select the additional features.

Choice 1. Keep running the proposed algorithm until time 15.

Choice 2. Use any other user-preferred feature selection methods to select the 5 additional features.

Choice 2 could be complicated. If the user-preferred feature selection method has a ranking on the selected features, such as filter methods, then one can find the additional features by looking for the top 5 features other than the already-selected 10 features. If the user-preferred feature selection method does not have a ranking among the selected features, one can start by selecting 15 features using the preferred method, and then check if there are exactly 5 new features in the group compared to the 10 features selected by the proposed method. If the number of new features in the group is more than 5, then one needs to reduce the number of selected features, using the preferred method, until a point that there are exactly 5 new features in the group, so that the 5 additional features can be determined.

After the two stages, the proposed method is completed. The performance of the proposed method is evaluated in the following section.

4.4 Simulations

In this section, we provide a simulation study to evaluate the performance of the proposed feature selection method. We first discuss the evaluation metric and then introduce the simulation setup and results.

4.4.1 Evaluation Metric

The proposed feature selection method selects only relevant features but does not provide an associated model or classifier. In evaluating such a feature selection method, there are two possible approaches [48]. The first approach is to embed a classifier and compare the accuracy of the classification process based on a real dataset. The results obtained with this approach are difficult to generalize as they depend on the specific classifier used in the comparison. The second approach is based on a scenario defined by an initial set of features and a relation between these

features and the outcome. Under this situation, a feature selection method could be evaluated by the truth. Focusing on the evaluation of the selected features, we adopt the second approach to evaluate the proposed feature selection method based on the truth. Under this approach, there are several strategies. One can calculate the percentage (success rate) of all relevant features that are selected. For example, let us consider an outcome T that is relevant to three features F_1, F_2, and F_3, where F_1 contributes the most information (variability) of T, F_2 contributes the second most, and F_3 contributes the least. Also, there is an irrelevant feature F_4 in the dataset. Suppose there are four different selection results: $S_1 = \{F_1\}$, $S_2 = \{F_1, F_4\}$, $S_3 = \{F_2, F_4\}$, and $S_4 = \{F_3, F_4\}$. Evaluating their performances using the success rate would achieve the same result (33.3% or 1/3) for all of them as they all identify one correct feature out of the three. The success rate is simple to calculate because the ground truth is known, and it works well when we focus on the number of correctly selected features or if we assume all the relevant features contribute evenly to the outcome. However, under the restriction of a limited sample, it may be more important to select the group of features that could jointly and efficiently provide the most information instead of selecting all relevant features regardless of the degrees of relevance and redundancy. Although ignoring low relevant or vastly redundant features may lose information, dropping them would further reduce the dimensionality and benefit the estimation. This can be considered as a trade off between estimation (dimensionality) and information: the more information, the more difficult the estimation. When the estimation is overly difficult, the results could be biased and hardly generalizable.

Because the success rate does not take the degrees of relevance and redundancy into consideration, we introduce the following evaluation metric to measure the ratio of the relevant information from the joint of selected features to the total relevant information from the joint of all the relevant features using mutual information.

Definition 1 (Information Recovery Ratio (IRR))**.**

$$IRR = \frac{MI(\mathcal{X}_{selected}, Y)}{MI(\mathcal{X}_{relevant}, Y)}, \tag{4.10}$$

where $\mathcal{X}_{selected}$ is the random variable that follows the joint-distribution of the selected features, and $\mathcal{X}_{relevant}$ is the random variable that follows the joint-distribution of all the features on which Y depends.

The IRR is not calculable in real datasets because 1) there is no knowledge on which features are relevant to the outcome, and 2) the true underlying distributions and associations (including redundancy) of the features and outcomes in real data are unknown. Given the setup of a simulation, we have all the knowledge; hence, the IRR for any group of selected features is calculable.

The IRR represents the percentage of relevant information in the joint of selected features. It considers feature redundancy by evaluating the mutual information between the joint-feature and the outcome. The range of the IRR is $[0, 1]$. If no relevant features are selected, the IRR is 0. If all the features in the dataset are selected regardless of relevance, the IRR is 1 for certain; therefore, when comparing the performance using the IRR, the number of selected features must be controlled. When the number of selected features from different methods are the same, a larger IRR means the joint of the selected features contains more relevant information; hence, the method is more efficient in dimension reduction. The efficiency of a feature selection method is desirable, particularly under small samples.

To make a comparison between the IRR and the success rate, both evaluate the performance of feature selection methods only when the ground truth is known. The success rate focuses on the ratio of the number of relevant features selected to the total number of relevant features, while the IRR focuses on the ratio of the relevant information in the joint of the selected features to the total relevant information.

4.4.2 Simulation Setup

A good evaluation scenario must include a representative set of features, containing relevant, redundant, and irrelevant ones [48]. In the simulation, we generate ten X variables ($X_1, \ldots, X_{10}$) and one outcome (Y). Among these variables, X_1, X_2, X_3, X_4 (or X_6), and X_5 are relevant features; X_6 (or X_4) is a redundant feature; X_7, X_8, X_9, and X_{10} are irrelevant features. The detailed settings are as follows.

$$Y = X_1 + X_2 + X_3^3 - 0.5 \cdot X_4^2 + |X_5| + X_6 + \varepsilon, \tag{4.11}$$

where

$$
\begin{aligned}
X_1 =& -3.5 \cdot \mathbb{1}[Z_1 < -3] - 1.4 \cdot \mathbb{1}[-3 \le Z_1 \le -0.5] \\
&+ \mathbb{1}[0.5 \le Z_1 \le 3] + 2.2 \cdot \mathbb{1}[Z_1 > 3], \\
X_2 =& -5 \cdot \mathbb{1}[Pois_1 = 0] - 3 \cdot \mathbb{1}[Pois_1 = 1] \\
&+ 2.4 \cdot \mathbb{1}[Pois_1 = 3 \text{ or } 4] + 5.4 \cdot \mathbb{1}[Pois_1 \ge 5], \\
X_3 =& -2 \cdot \mathbb{1}[U_1 \le -0.6] - \mathbb{1}[-0.6 < U_1 < -0.2] \\
&+ \mathbb{1}[0.2 < U_1 < 0.6] + 2 \cdot \mathbb{1}[U_1 \ge 0.6], \\
X_4 =& (B_1 - 2) \cdot \mathbb{1}[B_1 \ne 4] + 5 \cdot \mathbb{1}[B_1 = 4], \\
X_5 =& -2.5 \cdot \mathbb{1}[Z_2 < -0.5] - 2 \cdot \mathbb{1}[-0.5 \le Z_2 \le -0.2] \\
&+ 1.7 \cdot \mathbb{1}[-0.2 \le Z_2 \le 0.2] + 2 \cdot \mathbb{1}[0.2 \le Z_2 \le 0.6] \\
&+ 4 \cdot \mathbb{1}[Z_2 > 0.6], \\
X_6 =& X_4, \\
X_7 =& (Pois_2 - 2) \cdot \mathbb{1}[Pois_2 < 2] + 2 \cdot \mathbb{1}[Pois_2 \ge 2], \\
X_8 =& -2 \cdot \mathbb{1}[U_2 \le -0.6] - \mathbb{1}[-0.6 < U_2 < -0.2] \\
&+ \mathbb{1}[0.2 < U_2 < 0.6] + 2 \cdot \mathbb{1}[U_2 \ge 0.6], \\
X_9 =& B_2 - 1.2, \\
X_{10} =& -2 \cdot \mathbb{1}[Z_3 < -1.5] - 1.5 \cdot \mathbb{1}[-1.5 \le Z_3 \le -0.7] \\
&+ 1.5 \cdot \mathbb{1}[0.7 \le Z_3 \le 1.5] + 2 \cdot \mathbb{1}[Z_3 > 1.5], \\
\varepsilon =& -\mathbb{1}[U_3 \le \frac{1}{3}] + \mathbb{1}[U_3 \ge \frac{2}{3}],
\end{aligned}
$$

and

$$
\begin{aligned}
Z_1, Z_2, Z_3 \sim & N(0,1), \\
Pois_1, Pois_2 \sim & Poisson(2), \\
B_1 \sim & Binomial(4, 0.1), \\
B_2 \sim & Binomial(6, 0.2), \\
U_1, U_2 \sim & Uniform(-1, 1), \\
U_3 \sim & Uniform(0, 1).
\end{aligned}
$$

Usually, a simulation setup should include varieties to justify the challenges in real world data. Namely, it is often desirable to have complex feature spaces and complicated relationships among the features and the outcome. However, the above simulation setup is not complicated for the following reasons.

1. The purpose of this simulation is to evaluate the performance of the proposed method, particularly when the sample size is relatively small. The complexity of the feature spaces and the relationships among the features and the outcome would determine the threshold of what constitutes a sufficiently large sample. As they are not complex, we sample with smaller sizes to evaluate the performances in simulation. This is fair to all feature selection methods in comparison as they select features based on the same sample data with the same sample size.

2. The proposed feature selection method is one of the entropic methods. In the simulation, we would compare the performance of the proposed method to only other entropic methods because of the domain-specific challenges discussed in Section 4.1. During the simulation, we assign numerical values to the X variables so that we can generate the value of the outcome Y based on a model.

But entropic methods do not use the ordinal information from the numerical data as the inputs of the entropic methods are the frequencies of different numbers. Therefore, involving a complicated model (linear or nonlinear) does not affect the entropic methods because they regard the numbers as labels without ordinal information. However, complicating the model could make the outcome variable Y more complex and result in a higher threshold of a sufficiently large sample, which does not affect the comparison and evaluation among different methods, as discussed previously.

3. In calculating IRR, we need the two joint-distributions, $\mathcal{X}_{selected} \times Y$ and $\mathcal{X}_{relevant} \times Y$. To obtain the true joint-distributions, we have to enumerate the combinations among all possible values of the selected relevant features and of all the relevant features with their probabilities, respectively. Complicating the relevant X variables would make the calculation of the joint-distributions unnecessarily complex.

Note that the major benefit of a simple simulation setup is the ease in calculating the true joint-distributions, which are components of the IRR. In real world data, we do not need such calculations as the true joint-distributions and the IRR are not calculable. Hence, when applying the proposed method on real world high dimensional and complex data, the main calculation is just the estimated CASMI, which is not a problem.

With the simulation setup, one can consider that we create a dataset for evaluation. In this case, we know the ground truth that the features X_1, X_2, X_3, X_4 (or X_6), and X_5 should be selected. We would evaluate the performances by calculating the IRRs for features selected by different methods.

4.4.3 Simulation Results

In the simulation, we compare the IRR of the proposed feature selection method to the IRRs of six widely cited entropic feature selection methods: MIM, JMI, CMIM, MRMR, DISR, and NJMIM.

These six entropic methods all require users to set the number of features to be selected, while the proposed method can automatically decide the most appropriate number of features based on data. As we must control the number of selected features to validate the comparison of IRRs, we use the number of selected features from the proposed method as the number of features to be selected in the six entropic methods in each iteration. It should be noted that we are not claiming the number of features determined by the proposed method is correct. We set them to be the same only for the purpose of validating the comparison. As a matter of fact, the relevant features would not be entirely selected until the sample size is sufficiently large, and the threshold of a sufficiently large sample varies from method to method.

For each sample size N in $\{50, 100, 150, \ldots, 2750, 2800\}$, we re-generate the entire dataset 10,000 times and calculate the average IRRs of each method. The average IRR results are plotted in Figure 4.1.

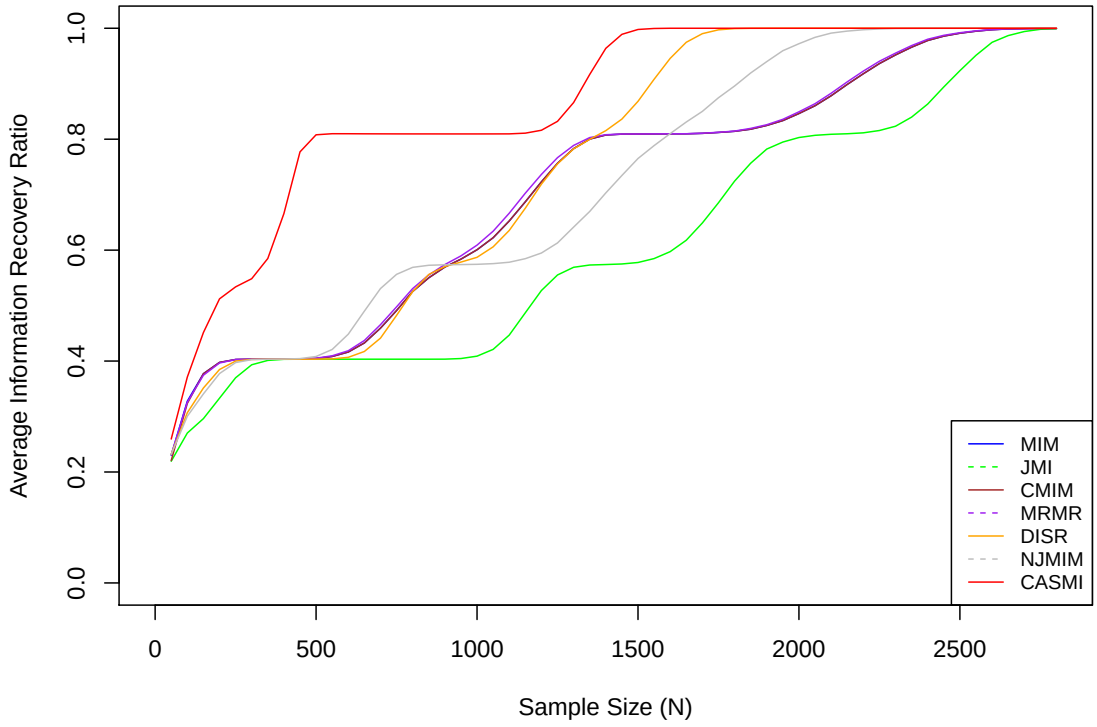

Figure 4.1: The average IRRs for seven methods, where CASMI refers to the proposed method. The proposed method is the most efficient method when sample size is limited. In the simulation, the threshold of a sufficiently large sample for the proposed method is approximately $N = 1500$, which is the smallest among all methods. The vertical index is the IRR, not the success rate. An IRR of 0.8 means 80% of the total mutual information has been accounted for by the selected features. It does not mean 80% of relevant features are selected. The proposed method does not select all relevant features when the sample size is small because some relevant features are in situation 2 under a limited sample; hence, they are not selected. As the sample size grows, all situation 2 features eventually become situation 1 features.

Based on the results, we can see that the average IRR of the proposed method is consistently higher than or equivalent to all the other methods. This is because under the restriction of a limited sample, the proposed method has a much smaller estimation bias so that it captures the associations among features and the outcome more accurately than the existing methods that estimate with the plug-in estimators. Table 4.2 presents the 95% confidence intervals for IRRs based on features selected by

different methods under different sample sizes. Based on the table, we can roughly rank the proposed methods and the six methods as follows: CASMI > DISR > NJMIM > MRMR > MIM $\sim$ CMIM > JMI.

Meanwhile, we recorded the average computation time of the proposed method when implementing feature selection in R. The plot of results is shown in Figure 4.2. The computation time when $N = 50$ was 0.03 seconds; the time when $N = 2800$ was 1.97 seconds; the longest time during the simulation was 3.37 seconds.

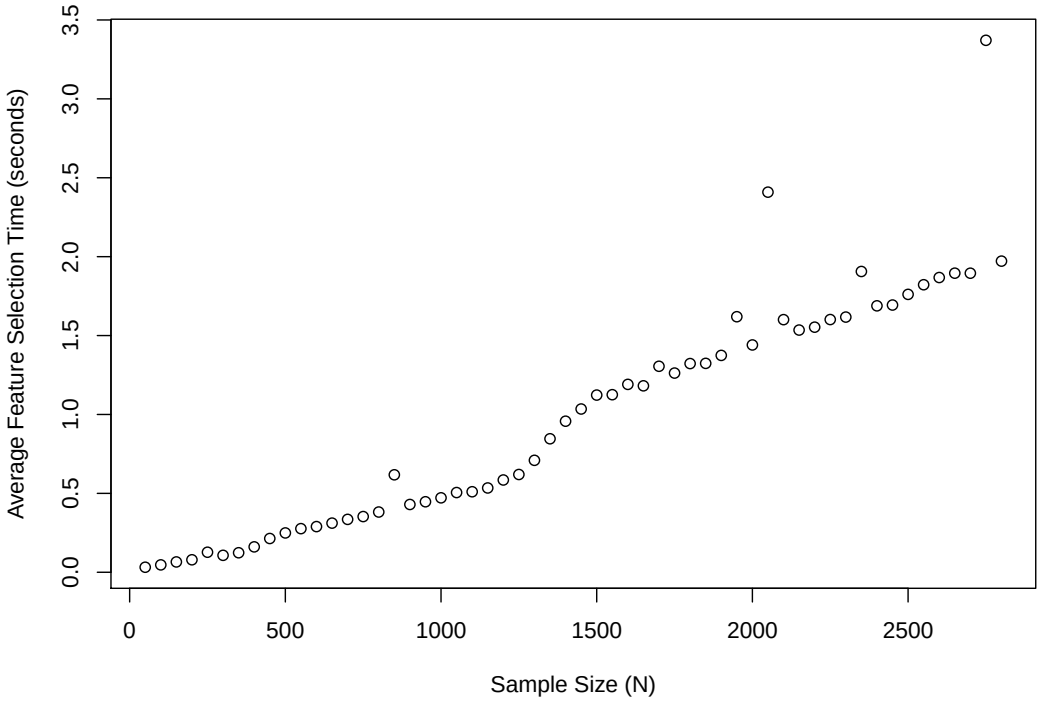

Figure 4.2: The average computation time of the proposed method when implementing feature selection in R.

Based on the simulation result in Figure 4.1, different methods achieve 1 (in average IRRs) at different sample sizes. One should realize that the threshold of a sufficiently large sample greatly depends on the probability spaces of the underlying associated features and the outcome. The probability spaces of real datasets are generally significantly more complicated than that of the simulation. Consequently,

in reality, particularly in health data, the majority of samples should be considered small; hence, the efficiency of a feature selection method is very important.

The simulation codes are available at [49]. The proposed feature selection method using CASMI are implemented in the R package at [50].

4.5 Discussion

In this article, we have proposed a new entropic feature selection method based on CASMI. Compared to existing methods, the proposed method has two unique advantages: 1) it is very efficient as the joint of selected features provides the most relevant information compared to features selected by other methods, particularly when the sample size is relatively small, and 2) it automatically learns the number of features to be selected from data. The proposed method handles feature redundancy from the perspective of joint-distributions. Although we initially developed the proposed method for the domain-specific challenges in healthcare, the proposed method can be used in many other areas where there is an issue of limited sample.

The proposed method is an entropic information-theoretic method. It aims at assisting data analytics on non-ordinal spaces. However, the proposed method can also be used on numerical data with an appropriate binning technique. Furthermore, using the proposed method on binned numerical data could discover different information as the entropic method looks at the data from a non-ordinal perspective.

In detecting unhelpful associations (situation 2 and 3 features), we implement an adjustment from the sample coverage. The level of this adjustment can be modified by users. For example, users can replace the scoring function of the proposed method by CASMI* with a tuning parameter (u) as follows:

$$\kappa^*(X,Y) = \kappa(X,Y) \cdot (1 - \pi_0(X))^u,$$

and estimate it by

$$\hat{\kappa}^*(X,Y) = \hat{\kappa}_z(X,Y) \cdot (1 - T_1(X))^u,$$

where u is any fixed positive number. The u can be considered as a parameter to determine the requirement for a feature to qualify situation 1. A larger u stands for a heavier penalty from the sample coverage; hence, a feature needs to contain more real information to be categorized to situation 1. A smaller u stands for a less penalty from the sample coverage; hence, a feature with less real information could be categorized to situation 1. However, users should be cautious when using a small u because it may mistakenly classify an irrelevant feature (situation 3) to situation 1, and further exacerbates the issue of generalization. We suggest to begin the proposed feature selection method with $u = 1$. After completing feature selection, if a user desires to select more or less features, the user could re-run the proposed method with a smaller or larger u, respectively, and keep modifying the value of u until satisfactory.

The proposed method only selects features but does not provide a classifier; however, to draw inferences on outcomes, a classifier is needed. To this end, additional techniques are required, such as machine learning (e.g., regressions and random forest). Into the future, it may be interesting to explore 1) methods that can distinguish features under situation 2 and 3 when the sample size is small; and 2) the possibilities of extending the proposed method to tree-based algorithms (e.g., random forest) to help determine which leaves and branches should be omitted. In addition, it may be interesting to investigate the performance of existing entropic methods if we use the $\hat{H}_z$, instead of $\hat{H}$, to estimate the entropies in their score functions.

Table 4.2: The 95% Confidence Intervals for IRRs based on features selected by different methods under different sample sizes.

n	CASMI CI	MIM CI	JMI CI	CMIM CI	MRMR CI	DISR CI	NJMIM CI
50	[0, 0.40]	[0, 0.40]	[0, 0.40]	[0, 0.40]	[0, 0.40]	[0, 0.40]	[0, 0.40]
100	[0.21, 0.57]	[0.21, 0.40]	[0.21, 0.40]	[0.21, 0.40]	[0.21, 0.40]	[0.21, 0.40]	[0.21, 0.40]
150	[0.26, 0.62]	[0.21, 0.40]	[0.21, 0.40]	[0.21, 0.40]	[0.21, 0.40]	[0.21, 0.40]	[0.21, 0.40]
200	[0.32, 0.62]	[0.21, 0.40]	[0.21, 0.40]	[0.21, 0.40]	[0.21, 0.40]	[0.21, 0.40]	[0.21, 0.40]
250	[0.37, 0.62]	[0.40, 0.40]	[0.21, 0.40]	[0.40, 0.40]	[0.40, 0.40]	[0.40, 0.40]	[0.21, 0.40]
300	[0.37, 0.62]	[0.40, 0.40]	[0.21, 0.40]	[0.40, 0.40]	[0.40, 0.40]	[0.40, 0.40]	[0.40, 0.40]
350	[0.37, 0.81]	[0.40, 0.40]	[0.40, 0.40]	[0.40, 0.40]	[0.40, 0.40]	[0.40, 0.40]	[0.40, 0.40]
400	[0.37, 0.81]	[0.40, 0.40]	[0.40, 0.40]	[0.40, 0.40]	[0.40, 0.40]	[0.40, 0.40]	[0.40, 0.40]
450	[0.62, 0.81]	[0.40, 0.40]	[0.40, 0.40]	[0.40, 0.40]	[0.40, 0.40]	[0.40, 0.40]	[0.40, 0.40]
500	[0.81, 0.81]	[0.40, 0.40]	[0.40, 0.40]	[0.40, 0.40]	[0.40, 0.40]	[0.40, 0.40]	[0.40, 0.57]
550	[0.81, 0.81]	[0.40, 0.57]	[0.40, 0.40]	[0.40, 0.57]	[0.40, 0.57]	[0.40, 0.40]	[0.40, 0.57]
600	[0.81, 0.81]	[0.40, 0.57]	[0.40, 0.40]	[0.40, 0.57]	[0.40, 0.57]	[0.40, 0.40]	[0.40, 0.57]
650	[0.81, 0.81]	[0.40, 0.57]	[0.40, 0.40]	[0.40, 0.57]	[0.40, 0.57]	[0.40, 0.57]	[0.40, 0.57]
700	[0.81, 0.81]	[0.40, 0.57]	[0.40, 0.40]	[0.40, 0.57]	[0.40, 0.57]	[0.40, 0.57]	[0.40, 0.57]
750	[0.81, 0.81]	[0.40, 0.57]	[0.40, 0.40]	[0.40, 0.57]	[0.40, 0.57]	[0.40, 0.57]	[0.40, 0.57]
800	[0.81, 0.81]	[0.40, 0.57]	[0.40, 0.40]	[0.40, 0.57]	[0.40, 0.57]	[0.40, 0.57]	[0.40, 0.57]
850	[0.81, 0.81]	[0.40, 0.57]	[0.40, 0.40]	[0.40, 0.57]	[0.40, 0.57]	[0.40, 0.57]	[0.57, 0.57]
900	[0.81, 0.81]	[0.40, 0.62]	[0.40, 0.40]	[0.40, 0.62]	[0.40, 0.81]	[0.40, 0.57]	[0.57, 0.57]
950	[0.81, 0.81]	[0.57, 0.81]	[0.40, 0.40]	[0.57, 0.81]	[0.57, 0.81]	[0.57, 0.57]	[0.57, 0.57]
1000	[0.81, 0.81]	[0.57, 0.81]	[0.40, 0.57]	[0.57, 0.81]	[0.57, 0.81]	[0.57, 0.81]	[0.57, 0.57]
1050	[0.81, 0.81]	[0.57, 0.81]	[0.40, 0.57]	[0.57, 0.81]	[0.57, 0.81]	[0.57, 0.81]	[0.57, 0.57]
1100	[0.81, 0.81]	[0.57, 0.81]	[0.40, 0.57]	[0.57, 0.81]	[0.57, 0.81]	[0.57, 0.81]	[0.57, 0.57]
1150	[0.81, 0.81]	[0.57, 0.81]	[0.40, 0.57]	[0.57, 0.81]	[0.57, 0.81]	[0.57, 0.81]	[0.57, 0.81]
1200	[0.81, 1]	[0.57, 0.81]	[0.40, 0.57]	[0.57, 0.81]	[0.57, 0.81]	[0.57, 0.81]	[0.57, 0.81]
1250	[0.81, 1]	[0.57, 0.81]	[0.40, 0.57]	[0.57, 0.81]	[0.57, 0.81]	[0.57, 0.81]	[0.57, 0.81]
1300	[0.81, 1]	[0.57, 0.81]	[0.40, 0.57]	[0.57, 0.81]	[0.57, 0.81]	[0.57, 0.81]	[0.57, 0.81]
1350	[0.81, 1]	[0.57, 0.81]	[0.57, 0.57]	[0.57, 0.81]	[0.57, 0.81]	[0.57, 0.81]	[0.57, 0.81]
1400	[0.81, 1]	[0.81, 0.81]	[0.57, 0.57]	[0.81, 0.81]	[0.81, 0.81]	[0.81, 1]	[0.57, 0.81]
1450	[0.81, 1]	[0.81, 0.81]	[0.57, 0.57]	[0.81, 0.81]	[0.81, 0.81]	[0.81, 1]	[0.57, 0.81]
1500	[1, 1]	[0.81, 0.81]	[0.57, 0.57]	[0.81, 0.81]	[0.81, 0.81]	[0.81, 1]	[0.57, 0.81]
1550	[1, 1]	[0.81, 0.81]	[0.57, 0.81]	[0.81, 0.81]	[0.81, 0.81]	[0.81, 1]	[0.57, 1]
1600	[1, 1]	[0.81, 0.81]	[0.57, 0.81]	[0.81, 0.81]	[0.81, 0.81]	[0.81, 1]	[0.57, 1]
1650	[1, 1]	[0.81, 0.81]	[0.57, 0.81]	[0.81, 0.81]	[0.81, 0.81]	[0.81, 1]	[0.57, 1]
1700	[1, 1]	[0.81, 0.81]	[0.57, 0.81]	[0.81, 0.81]	[0.81, 0.81]	[0.81, 1]	[0.81, 1]
1750	[1, 1]	[0.81, 0.81]	[0.57, 0.81]	[0.81, 0.81]	[0.81, 0.81]	[1, 1]	[0.81, 1]
1800	[1, 1]	[0.81, 0.81]	[0.57, 0.81]	[0.81, 0.81]	[0.81, 1]	[1, 1]	[0.81, 1]
1850	[1, 1]	[0.81, 1]	[0.57, 0.81]	[0.81, 1]	[0.81, 1]	[1, 1]	[0.81, 1]
1900	[1, 1]	[0.81, 1]	[0.57, 0.81]	[0.81, 1]	[0.81, 1]	[1, 1]	[0.81, 1]
1950	[1, 1]	[0.81, 1]	[0.57, 0.81]	[0.81, 1]	[0.81, 1]	[1, 1]	[0.81, 1]
2000	[1, 1]	[0.81, 1]	[0.57, 0.81]	[0.81, 1]	[0.81, 1]	[1, 1]	[0.81, 1]
2050	[1, 1]	[0.81, 1]	[0.81, 0.81]	[0.81, 1]	[0.81, 1]	[1, 1]	[0.81, 1]
2100	[1, 1]	[0.81, 1]	[0.81, 0.81]	[0.81, 1]	[0.81, 1]	[1, 1]	[0.81, 1]
2150	[1, 1]	[0.81, 1]	[0.81, 0.81]	[0.81, 1]	[0.81, 1]	[1, 1]	[0.81, 1]
2200	[1, 1]	[0.81, 1]	[0.81, 0.81]	[0.81, 1]	[0.81, 1]	[1, 1]	[1, 1]
2250	[1, 1]	[0.81, 1]	[0.81, 1]	[0.81, 1]	[0.81, 1]	[1, 1]	[1, 1]
2300	[1, 1]	[0.81, 1]	[0.81, 1]	[0.81, 1]	[0.81, 1]	[1, 1]	[1, 1]
2350	[1, 1]	[0.81, 1]	[0.81, 1]	[0.81, 1]	[0.81, 1]	[1, 1]	[1, 1]
2400	[1, 1]	[0.81, 1]	[0.81, 1]	[0.81, 1]	[0.81, 1]	[1, 1]	[1, 1]
2450	[1, 1]	[0.81, 1]	[0.81, 1]	[0.81, 1]	[0.81, 1]	[1, 1]	[1, 1]
2500	[1, 1]	[0.81, 1]	[0.81, 1]	[0.81, 1]	[0.81, 1]	[1, 1]	[1, 1]
2550	[1, 1]	[0.81, 1]	[0.81, 1]	[0.81, 1]	[1, 1]	[1, 1]	[1, 1]
2600	[1, 1]	[1, 1]	[0.81, 1]	[1, 1]	[1, 1]	[1, 1]	[1, 1]
2650	[1, 1]	[1, 1]	[0.81, 1]	[1, 1]	[1, 1]	[1, 1]	[1, 1]
2700	[1, 1]	[1, 1]	[0.81, 1]	[1, 1]	[1, 1]	[1, 1]	[1, 1]
2750	[1, 1]	[1, 1]	[1, 1]	[1, 1]	[1, 1]	[1, 1]	[1, 1]
2800	[1, 1]	[1, 1]	[1, 1]	[1, 1]	[1, 1]	[1, 1]	[1, 1]

Chapter 4 Reference List

[1] C. S. Kruse, R. Goswamy, Y. Raval, and S. Marawi, "Challenges and opportunities of big data in health care: a systematic review," *JMIR medical informatics*, vol. 4, no. 4, 2016.

[2] C. H. Lee and H.-J. Yoon, "Medical big data: promise and challenges," *Kidney research and clinical practice*, vol. 36, no. 1, p. 3, 2017.

[3] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[4] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[5] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, p. 94, 2017.

[6] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: introduction and review," *Journal of biomedical informatics*, 2018.

[7] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018.

[8] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[9] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[10] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," *Machine learning*, vol. 53, no. 1-2, pp. 23–69, 2003.

[11] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, "Trace ratio criterion for feature selection.," in *AAAI*, vol. 2, pp. 671–676, 2008.

[12] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Advances in neural information processing systems*, pp. 507–514, 2006.

[13] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proceedings of the 24th international conference on Machine learning*, pp. 1151–1157, ACM, 2007.

[14] J. Liu, S. Ji, J. Ye, *et al.*, "Slep: Sparse learning with efficient projections," *Arizona State University*, vol. 6, no. 491, p. 7, 2009.

[15] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint l2, 1-norms minimization," in *Advances in neural information processing systems*, pp. 1813–1821, 2010.

[16] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 333–342, ACM, 2010.

[17] F. Yang and K. Mao, "Robust feature selection for microarray data based on multicriterion fusion," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 4, pp. 1080–1092, 2011.

[18] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised feature selection using nonnegative spectral analysis," in *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[19] J. C. Davis and R. J. Sampson, *Statistics and data analysis in geology*, vol. 646. Wiley New York et al., 1986.

[20] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[21] D. D. Lewis, "Feature selection and feature extraction for text categorization," in *Proceedings of the workshop on Speech and Natural Language*, pp. 212–217, Association for Computational Linguistics, 1992.

[22] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on neural networks*, vol. 5, no. 4, pp. 537–550, 1994.

[23] H. H. Yang and J. Moody, "Data visualization and feature selection: New algorithms for nongaussian data," in *Advances in Neural Information Processing Systems*, pp. 687–693, 2000.

[24] M. Vidal-Naquet and S. Ullman, "Object recognition with informative features and linear classification.," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 3, p. 281, 2003.

[25] F. Fleuret, "Fast binary feature selection with conditional mutual information," *Journal of Machine Learning Research*, vol. 5, no. Nov, pp. 1531–1555, 2004.

[26] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[27] D. Lin and X. Tang, "Conditional infomax learning: an integrated framework for feature extraction and fusion," in *European Conference on Computer Vision*, pp. 68–82, Springer, 2006.

[28] P. E. Meyer and G. Bontempi, "On the use of variable complementarity for feature selection in cancer classification," in *Workshops on applications of evolutionary computation*, pp. 91–102, Springer, 2006.

[29] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using joint mutual information maximisation," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8520–8532, 2015.

[30] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," in *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*, pp. 388–391, IEEE, 1995.

[31] C. Gini, "Variabilita e mutabilita, studi economico-giuridici della r," *Universita di Cagliari*, vol. 3, no. 2, pp. 3–159, 1912.

[32] M. A. Hall and L. A. Smith, "Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper.," in *FLAIRS conference*, vol. 1999, pp. 235–239, 1999.

[33] B. Harris, "The statistical estimation of entropy in the non-parametric case," tech. rep., Wisconsin Univ-Madison Mathematics Research Center, 1975.

[34] L. Paninski, "Estimation of entropy and mutual information," *Neural computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[35] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 3-4, pp. 237–264, 1953.

[36] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[37] A. D. Wyner, "A definition of conditional mutual information for arbitrary ensembles," *Information and Control*, vol. 38, no. 1, pp. 51–59, 1978.

[38] S. Guiasu, *Information theory with applications*, vol. 202. McGraw-Hill New York, 1977.

[39] Z. Zhang, *Statistical Implications of Turing's Formula*. John Wiley & Sons, 2016.

[40] M. I. Ohannessian and M. A. Dahleh, "Rare probability estimation under regularly varying heavy tails," in *Conference on Learning Theory*, pp. 21–1, 2012.

[41] Z. Zhang, "Entropy estimation in turing's perspective," *Neural computation*, vol. 24, no. 5, pp. 1368–1389, 2012.

[42] Z. Zhang and L. Zheng, "A mutual information estimator with exponentially decaying bias," *Statistical applications in genetics and molecular biology*, vol. 14, no. 3, pp. 243–252, 2015.

[43] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," in *Machine Learning Proceedings 1995*, pp. 194–202, Elsevier, 1995.

[44] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[45] R. J. Little, R. D'agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, *et al.*, "The prevention and treatment of missing data in clinical trials," *New England Journal of Medicine*, vol. 367, no. 14, pp. 1355–1360, 2012.

[46] H. Kang, "The prevention and handling of the missing data," *Korean journal of anesthesiology*, vol. 64, no. 5, pp. 402–406, 2013.

[47] J. Zhang and C. Chen, "On 'a mutual information estimator with exponentially decaying bias' by zhang and zheng," *Statistical applications in genetics and molecular biology*, vol. 17, no. 2, 2018.

[48] C. Pascoal, M. R. Oliveira, A. Pacheco, and R. Valadas, "Theoretical evaluation of feature selection methods based on mutual information," *Neurocomputing*, vol. 226, pp. 168–181, 2017.

[49] J. Shi, "Casmi simulation r codes." `https://github.com/JingyiShi/CASMI/blob/master/SimulationEvaluationUsingGroundTruth.R`, 2019. [Online; Github].

[50] J. Shi, "Casmi in r." `https://github.com/JingyiShi/CASMI`, 2019. [Online; Github].

CHAPTER 5: AN ASSOCIATION-BASED INTRINSIC QUALITY INDEX FOR HEALTHCARE DATASET RANKING

5.1 Introduction

With the increasing ease of collecting, managing, and sharing data, a large number of datasets have been available for analytics. Particularly in healthcare, one of the world's largest and fastest-growing industries, emerging datasets have been rapidly generated from electronic health records (EHRs), practice management systems, insurance organizations, clinical trials, and public surveys. These datasets bring opportunities for researchers to assist healthcare providers and patients in improving health care quality and reducing cost. Meanwhile, the data explosion introduces significant challenges in discovering and identifying a suitable dataset for a specified research target. Healthcare datasets are generally complex and diverse. Searching for a suitable healthcare dataset requires many efforts because it takes time to access, set up, and fully understand datasets. As a result, assistance in healthcare dataset identification is needed.

Aiming at facilitating dataset identification, Google released a Dataset Search [1] engine in 2018, which provides organized information of datasets and a keyword search function. Meanwhile, Google has promoted schema.org, a metadata standard, to encourage people to upload structured metadata of new datasets. A few platforms targeting the same goal for specific domains have been developed as well, including the Neuroscience Information Framework (NIF) for neuroscience data and resources [2], HealthData.gov for government healthcare datasets [3], DataMed.org for biomedical datasets and repositories [4], and the Dataset Information Resource (DIR) framework for healthcare datasets [5]. These platforms introduce standards

of dataset structure and help researchers understand contents in datasets. Moreover, some of these platforms provide functions that allow researchers to search for relevant contents; for example, keyword search functions in almost all platforms, filtering in HealthData.org and DataMed.org, and formal language querying and questions answering-like modules in DIR. The matching results can be sorted by relevance (nearly all platforms), data collection time (e.g., HealthData.gov), and alphabetic order (e.g., HealthData.gov).

With the help of these platforms, researchers can quickly find healthcare datasets related to their research purposes. In most situations, there are multiple relevant healthcare datasets. For example, suppose a group of researchers are interested in the causes of breast cancer, then all healthcare datasets that contain the attribute of breast cancer history can be relevant to the research purpose. Among these datasets, the researchers would need to make a decision about which datasets to use for their specific research. One natural criterion for such decisions is some measure of quality. In other words, they need to choose the dataset with the most potential for producing the most valid analytic results. Numerous studies have attempted to evaluate such quality. For example, EHR data quality has been categorized into completeness, correctness, concordance, plausibility, and currency in 2013 [6], conformance, completeness, and plausibility in 2016 [7], and completeness, consistency, credibility, and timeliness in 2018 [8], while such concept harmonization is limited for other types of healthcare datasets. These qualitative measurements provide information on quality from different perspectives, but researchers have to judge each dataset individually and subjectively. These measures are also not conducive for comparing among multiple datasets.

To help make a comparison, several ranking indexes have been developed to rank healthcare datasets based on publication citations, such as the U-index [9] and the Publication-based Popularity Index (PPI) [10]. These methods provide scores of pop-

ularity and focus on extrinsic quality. However, in some situations, intrinsic quality may be more valuable. For example, a healthcare dataset could have high intrinsic quality but is less-analyzed because the data were collected recently. Such datasets are hard to discover when measured using popularity; nevertheless, these recent quality healthcare datasets are more promising toward new insights and deserve more attention than those well studied popular datasets. Therefore, we believe that an index to measure the intrinsic quality of healthcare datasets is needed.

The intrinsic quality of a healthcare dataset has multiple perspectives; for example, as we mentioned earlier, the correctness, credibility, and completeness. For a health data scientist who is searching for a healthcare dataset with high intrinsic quality, the ultimate desire from the dataset is the potential to obtain valid results (productivity). Noticing that while a healthcare dataset could contain more than one potential outcome attribute for different research purposes, and for each individual study, there is usually only one outcome at a time, we begin with the situation of only one possible outcome variable in the dataset, and we discuss the extension to situations of multiple potential outcomes in Section 5.5. When there is only one outcome variable, and to obtain valid results, there must be relevant features in the dataset. If all features carry little association to the outcome, then almost no valid results can be obtained despite what efforts are spent. Therefore, we desire to find a measurement (score) of the association between features and the outcome in a dataset to indicate the intrinsic quality, such that the higher the score, the stronger the association and the better the intrinsic quality.

Nevertheless, measuring the association among attributes (features and outcomes) in healthcare datasets is not a simple task. Some health data are continuous or discrete (ordinal), and some are categorical (non-ordinal). Usually, correlation coefficients (CC) (*e.g.*, Pearson correlation coefficient) are used to measure the association among ordinal data, and information-theoretic quantities (ITQ) (*e.g.*, mutual

information and information gain ratio) are used to measure the association among non-ordinal data. While neither CCs nor ITQs could evaluate the association among both ordinal and non-ordinal data, healthcare datasets are often mixtures of the two data types. A naive approach could be using CCs for the ordinal features and ITQs for the non-ordinal features, in a single dataset. However, this does not solve the issue because the outcome is either ordinal or non-ordinal; if it is ordinal, ITQs cannot measure the association between non-ordinal features and the ordinal outcome, and if it is non-ordinal, CCs cannot measure the association between ordinal features and the non-ordinal outcome.

For non-ordinal data, the absence of metric blocks the transformation from non-ordinal to ordinal. While for ordinal data, a discretization (or binning) that drops the metric information could transform the data from ordinal to non-ordinal. Although binning ordinal data loses some metric information, this does not mean a loss of association. The dropped metric information from binning could help prediction; however, instead of building a classifier or making a prediction, we desire only to measure the degree of association to indicate the intrinsic quality. If there is a linear correlation from the metric space among the data, binning them transforms the linear relationship to a non-ordinal association, which could be captured by ITQs, as well. Furthermore, binning ordinal health data and evaluating the association using ITQs could be more accurate compared to using CCs because numerous associations are nonlinear while most CCs could capture linear associations only. Therefore, in evaluating the association among both types of data, we suggest binning the ordinal health data and using ITQs to measure the association.

Using a score to evaluate the association between features and the outcome is very similar to the spirit of feature selection, specifically filter methods. Majority of existing filter feature selection methods use ITQs to measure the association, such as Mutual Information Maximization (MIM) [11], Joint Mutual Information (JMI)

[12], and Coverage Adjusted Standardized Mutual Information (CASMI) [13]. In evaluating the intrinsic quality of a healthcare dataset other than selecting features, there are two pitfalls: redundancy and inflation. In a healthcare dataset, a number of features could be redundant. Because stacking redundant features does not help improve the intrinsic quality, this issue must be addressed. The problem of inflation occurs if the score does not decrease when introducing an irrelevant feature. One can consider R-square in regression analysis to understand the issue of inflation (R-square never decreases when adding regression variables). Many healthcare datasets, such as claims data, are often collected without a specific research question in mind; hence, there may be a lot of irrelevant features to the targeting outcome. As a result, if the issue of inflation is not addressed, the score would become misleading as it has been inflated by the irrelevant features.

Majority of existing filter feature selection methods either ignore feature redundancy (*e.g.*, MIM) or have the issue of inflation (*e.g.*, MIM and JMI). To address the two issues in quantifying data quality, we design the proposed Association-based intrinsic Quality Index (AQI) based on the CASMI, which considers both issues. Initially, the AQI is expected to help in comparing and ranking multiple healthcare datasets in perspective of intrinsic quality. Additionally, the AQI could be used for discovering research opportunities in a given dataset. The rest of the article is organized as follows. The estimated CASMI and the design of the proposed AQI are described in Section 5.2. Data sources for an AQI usage demonstration and a user study are introduced in Section 5.3. The results of the user study and AQI limitations are discussed in Section 5.4. Finally, we make a conclusion and clarify AQI usage scenarios in Section 5.5.

5.2 AQI for Healthcare Datasets

In this section, we introduce the CASMI first and then the design of the AQI for comparisons of healthcare datasets.

Introduced by [13], the CASMI is a score that can measure the general association between features and the outcome. We state the definition of the estimated CASMI in Definition 2.

Definition 2 (Estimated CASMI with tuning parameter u)**.** *$\hat{\kappa}^*$, the estimated Coverage Adjusted Standardized Mutual Information (CASMI) of a feature X to an outcome Y, with a tuning parameter u, is defined as*

$$\hat{\kappa}^*(X,Y) = \hat{\kappa}_z(X,Y) \cdot (1 - T_1(X))^u, \tag{5.1}$$

where

$$\hat{\kappa}_z(X,Y) = \frac{\widehat{MI}_z(X,Y)}{\hat{H}_z(Y)}, \tag{5.2}$$

$$\hat{H}_z(\cdot) = \sum_{v=1}^{n-1} \frac{1}{v} \frac{n^{1+v}[n-(1+v)]!}{n!} \sum_k \left[\hat{p}_k \prod_{j=0}^{v-1} \left(1 - \hat{p}_k - \frac{j}{n}\right) \right],$$

$$\widehat{MI}_z(X,Y) = \hat{H}_z(X) + \hat{H}_z(Y) - \hat{H}_z(X,Y); \tag{5.3}$$

$$1 - T_1(X) = 1 - N_1(X)/n, \tag{5.4}$$

$N_1(X)$ is the number of singletons in the data of X, and n is the sample size. In (5.1), u is a tuning parameter that can be customized.

The tuning parameter, u, in the CASMI controls the level of penalty from a low sample coverage. Sample coverage means the proportion of the observed categories in population. For a feature, a low sample coverage suggests less reliable information in the feature, and thus its association with the outcome is questionable and deserves

a penalty. In the CASMI, the larger the u value, the heavier the penalty. Moreover, when n is large, u should be small because the issues of low sample coverage are not severe in a relatively large sample (dataset). In other words, u should be a decreasing function of n. For this reason, we set $u = 1/(\ln \ln n)$ for the AQI.

The feature selection method using the CASMI in [13] includes a screening test to filter out irrelevant features at first. We state the test in Theorem 2 as we also add it to reduce the calculation time of the AQI. The screening test is proven in [14].

Theorem 2. *If the real mutual information is zero, then*

$$\chi^2 = 2n\widehat{MI}_z + (K_1 - 1)(K_2 - 1) \xrightarrow{L} \chi^2\left((K_1 - 1)(K_2 - 1)\right), \tag{5.5}$$

where $\widehat{MI}_z$ is defined in (5.3). K_1 and K_2 are the effective cardinalities of the feature X and the outcome Y, respectively.[1]

The AQI is developed based on the CASMI with the tuning parameter $u = 1/(\ln \ln n)$, where n is the number of observations in a healthcare dataset. In calculating the AQI of a dataset for a specified outcome (Y), the steps are:

1. Preprocess all attributes (features and the outcome) in the dataset to categorical.

2. Use (5.5) to test if mutual information (MI) between each of the features and the outcome (Y) in the dataset is 0 against $MI > 0$ at $\alpha = 0.20$, and let $\{X\}_s$ be the collection of all features that reject the null hypothesis in the tests. In other words, $\{X\}_s$ contains features that are associated with the outcome, according to the test.

3. i. $X_{(1)} = \underset{X_i \in \{X\}_s}{\arg\max} \left[\hat{\kappa}^*(X_i, Y)\right]$;

   ii. $X_{(2)} = \underset{X_i \in \{X\}_s \setminus \{X_{(1)}\}}{\arg\max} \left[\hat{\kappa}^*(X_{(1)} \times X_i, Y)\right]$;

[1] We write $\xrightarrow{L}$ to denote convergence in distribution.

iii. $X_{(3)} = \underset{X_i \in \{X\}_s \setminus \{X_{(1)}, X_{(2)}\}}{\arg\max} \left[\hat{\kappa}^*(X_{(1)} \times X_{(2)} \times X_i, Y)\right]$;

...

This stops at time c when $\hat{\kappa}^*(X_{(1)} \times \cdots \times X_{(c+1)}, Y) < \hat{\kappa}^*(X_{(1)} \times \cdots \times X_{(c)}, Y)$; hence, $\hat{\kappa}^*(X_{(1)} \times \cdots \times X_{(c)}, Y)$ is the final score of the estimated CASMI, and we denote it as $\hat{\kappa}_0^*$. Alternatively, it also stops when there is no additional features in $\{X\}_s$.

4. The AQI for the dataset when Y is the outcome is

$$AQI = \frac{100}{1 - \ln \hat{\kappa}_0^{*1/e}}. \tag{5.6}$$

Before explaining these steps, we clarify the notations. $\hat{\kappa}^*(X_{(1)} \times X_i, Y)$ stands for the $\hat{\kappa}^*$ of the joint feature $X_{(1)} \times X_i$ to the outcome Y.

In step 1, all ordinal features and the outcome in the dataset must be binned to non-ordinal. The reason for the discretization has been discussed in Section 5.1. There are different approaches to bin ordinal attributes, which are discussed in [15]. While binning ordinal data, the estimated sample coverage (5.4) should be checked to avoid over-discretization, which unnecessarily reduces the AQI. If the data are already non-ordinal, one may need to combine some of the categories for some features to improve their sample coverages. Including sample coverage to evaluate associations among attributes in a dataset reduces the risk of a falsely high AQI for two reasons. First, ID-like features in a dataset could provide a significant association to the outcome, while such an association is manipulated and unhelpful. Second, non-ID features, particularly continuous features, could seem associated to the outcome if they are excessively discretized. Under an extreme situation, a feature with all singletons (the finest discretization) is a surjection onto the outcome, regardless of the degree of relevance. While a relevant feature could naturally be all singletons in a dataset, we suggest reasonably binning the feature to avert the suspicion from the surjection. We

suggest users control the estimated sample coverage for each potential relevant feature at 50% and above (if possible) to allow it to contribute to the AQI, and obviously irrelevant features (*e.g.*, IDs) should not be regrouped. User should be aware that the AQI is likely to change with the corresponding binning method. As demonstrated at the end of Section 5.4, the oscillation of AQIs from different binning methods is not severe as long as the binning methods and sample coverages are reasonable. In summary, an appropriate binning method is sufficient, and it is unnecessary to strive for the best one.

When a feature contains missing data, there are several remedy options. For example, one can delete the observation, make an educated guess, predict the missing values, or list all missing values as NA. The appropriateness of each remedy option could vary from situation to situation. While it is the user's preference on how to handle the missing data, one should be cautious because manipulating (guessing or predicting) the missing data may create false associations that wrongly increase the AQI. Assigning all the missing values as NA generally would not create false associations, but it might cause an underestimation of the AQI for the dataset. Additional discussions on handling missing data can be found in [16], [17], and [18]. We suggest processing the missing data before discretization.

In step 2, the screening test is performed to filter out features that are not associated with the targeting outcome. This helps reduce the calculation time because healthcare datasets are generally collected without a specific research goal. For this reason, when an attribute Y is selected to be the outcome in a healthcare dataset, many features in the dataset may not be associated with the chosen outcome attribute. After the screening test, we obtain a collection of features $\{X\}_s$ that are believed to be relevant to the outcome, and we later calculate the AQI for the dataset to the outcome Y based on $\{X\}_s$. For the screening test between each feature and the outcome, K_1 is the number of categories in the preprocessed feature, and K_2 is the number of categories

in the preprocessed outcome. If a feature or outcome is naturally non-ordinal with satisfactory estimated sample coverage that does not need a regrouping, in which case the K_1 or (and) K_2 could be unknown, then the Ks can be estimated with the number of categories in the corresponding sample data.

In step 3, we calculate the best score of the estimated CASMI based on the relevant features $\{X\}_s$. This is the core step where the issues of redundancy and inflation are handled. The issue of feature redundancy is addressed by adopting joint-distributions between features. Namely, we first find the $X_{(1)}$ with highest $\hat{\kappa}^*$ among $\{X\}_s$. We use the notation $X_{(1)}$ because we leave X_1 as the notation for the first feature in $\{X\}_s$. $X_{(1)}$ is considered to be the most relevant feature to the outcome. After $X_{(1)}$ is selected, we joint the $X_{(1)}$ with each feature in $\{X\}_s \backslash \{X_{(1)}\}$ to obtain $(s-1)$ joint-features, and then we calculate $\hat{\kappa}^*$ between each of the joint-feature and the outcome. From the $(s-1)$ $\hat{\kappa}^*$s, we pick the highest one to determine $X_{(2)}$. In this way, given $X_{(1)}$, the $X_{(2)}$ provides the most additional information to the outcome. By induction, $X_{(3)}$ provides the most additional information given $X_{(1)}$ and $X_{(2)}$, ...; all the features selected later are based on the amount of the additional information. Therefore, the issue of redundancy is addressed because redundant information among multiple features are counted only once, and the redundant features would not have duplicate contributions to the AQI.

For the issue of inflation, the calculation of $\hat{\kappa}^*$ contains a penalized term from the number of singletons. Note that, when calculating $\hat{\kappa}^*$ for joint-features, the number of singletons is counted in the joint-space but not in the original feature space. For example, suppose there is a dataset with only five observations and two features X_1 and X_2, and let $X_1 = \{1, 1, 2, 2, 3\}$ and $X_2 = \{1, 1, 1, 3, 3\}$. There is one singleton in X_1 and no singleton in X_2; however, $X_1 \times X_2 = \{(1, 1), (1, 1), (2, 1), (2, 3), (3, 3)\}$ contains three singletons. Adding an additional feature to the joint almost certainly increases the number of singletons. If there are too many singletons in the space of

joint-features, then there is a risk of over-fitting; hence, we include a penalized term based on the number of singletons to enforce a balance between extra information and the risk of over-fitting. A feature could contribute to the AQI only if the extra information brought by the feature is more valuable than the extra risk of over-fitting, and this is the reason it stops at time c.

In step 4, the AQI is a scaled $\hat{\kappa}_0^*$ from step 3. The reason for scaling is because, in most cases, $\hat{\kappa}_0^*$s tend to be small values (*e.g.*, $\hat{\kappa}_0^* < 0.5$), from which differences among datasets are not well represented. After scaling, the range of the AQI is [0,100]. A higher AQI means a higher $\hat{\kappa}_0^*$, which indicates a stronger association between the features and the outcome, hence a quality healthcare dataset. To help understand the calculation of the AQI, we summarize the process in Fig. 5.1.

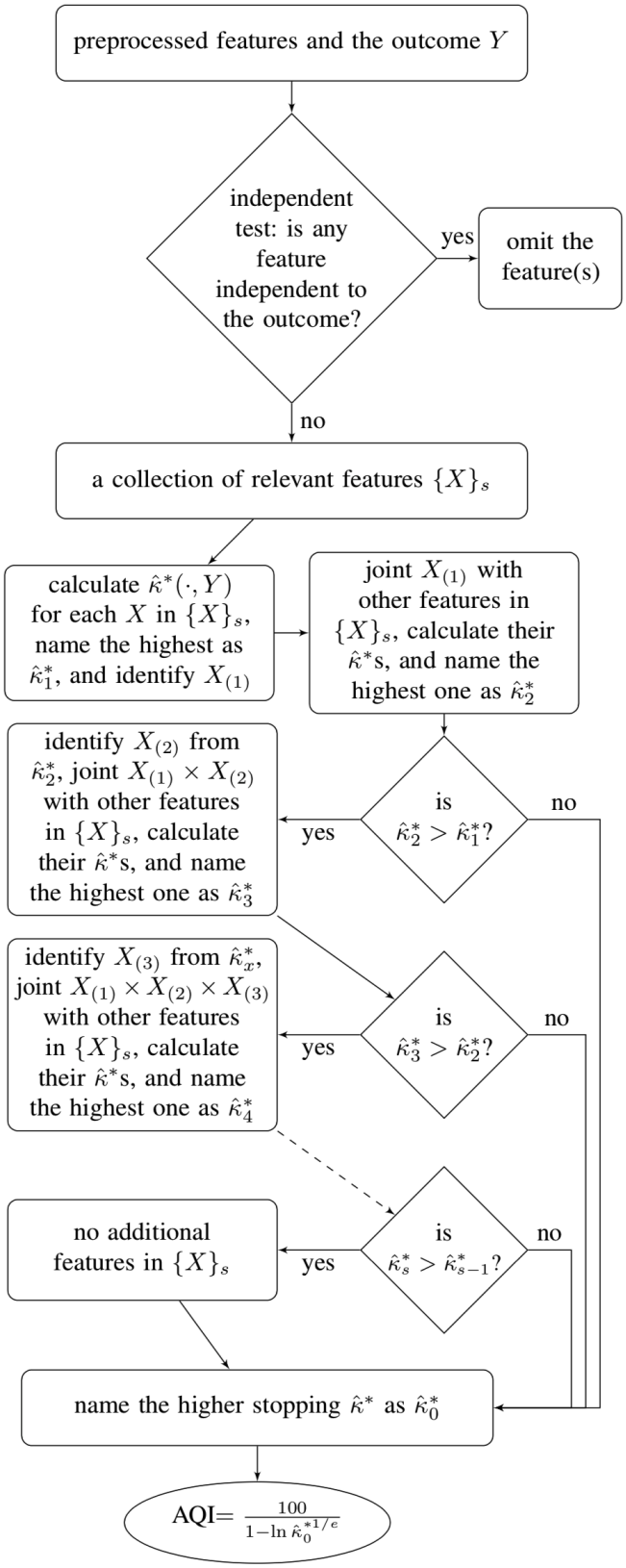

Figure 5.1: Algorithm for AQI calculation.

5.3 Data Source for AQI Usage Demonstration

To propose a preliminary result and help users understand the usage of the AQI, we demonstrate the utility of the AQI by comparing two pairs of datasets from UCI dataset repository [19]. Moreover, we conducted a survey based on the pairs, which we discuss in Section 5.4.

Among the two pairs, the first is about a diagnostic of breast cancer. The possible outcome responses in each of the two datasets are benign or malignant. These datasets were originally obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg [20]. The first dataset was titled Wisconsin Breast Cancer Database (WBCD) [21], and the second one was titled Wisconsin Diagnostic Breast Cancer (WDBC) [22]. For the WBCD, there are 698 patient records with 11 attributes, including the outcome. Calculating the AQI for the WBCD does not require data prepocessing as all attributes are categorical (or discrete). After calculation, the AQI score is 95.5785. For the WDBC, there are 568 patient records with 32 attributes, including the outcome. Calculating the AQI for the WDBC requires data prepocessing on some of the attributes. For each continuous feature in the WDBC, we binned it to five categories with equal width based on the range of continuous values represented in the data. Note that the equal-width binning method is only the method we chose as a demonstration in the example. In practice, one needs to consider both outliers and the distribution of data (if available) when selecting a binning method. As a result, the AQI score of the WDBC is 93.87177.

The second pair of datasets is about a prognostic of breast cancer. The possible outcome responses in each of the two datasets are whether there are recurrence-events of breast cancer (labeled as recurrence-events or no-recurrence-events). The first dataset was titled Breast Cancer Data (BCD) [23]. It was provided by M. Zwitter and M. Soklic and originally obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. The second dataset was titled Wisconsin Prognostic

Breast Cancer (WPBC) [24], which was also originally obtained from the University of Wisconsin. The BCD contains 285 patient records with 10 attributes, including the outcome; the WPBC contains 197 patient records with 35 attributes, including the outcome. Calculating the AQI for the BCD does not require data prepocessing as all attributes are categorical (or discrete), while calculating the AQI for the WPBC requires data prepocessing on some of the attributes. For each continuous feature in the WPBC, we binned it to five categories with equal width, as well. After calculation, the AQI score for the BCD and the WPBC are 70.92659 and 80.84716, respectively. The scores of the two pairs are summarized in Table 5.1.

Table 5.1: AQI scores in demonstration.

Pair #	Dataset Titles	AQI Scores
1	WBCD	95.5785
1	WDBC	93.87177
2	BCD	70.92659
2	WPBC	80.84716

5.4 Results and Discussion

To help dataset comparisons and ranking, we have developed the AQI to quantify intrinsic quality in an association perspective for healthcare datasets. The steps to calculate the AQI are summarized in Fig. 5.1. In addition, we have developed R pakcage CASMI [25], in which function AQI.score provides the calculation. Users could directly input preprocessed data (as discussed in Section 5.2) into the function and obtain a corresponding AQI score.

To study the helpfulness and limitations of the AQI, we conducted a survey based on the two pairs of datasets discussed in Section 5.3. We asked participants to select a better dataset from each pair for analysis and to state reasons. At first, we did not provide any information other than datasets themselves. After participants made a decision, we then provided the calculated AQI scores and asked them to select again. We invited ten participants to response, of whom nine had background in data science

and analytics. Particularly, three out of nine studied in health informatics. None of the participants were familiar with these datasets.

For the first pair, before providing the AQIs, the choices from the ten participants were quite evenly distributed: three selected the WBCD, three selected the WDBC, and four said they would randomly pick one. Their reasons to make such selections were various. For example, someone selected the WBCD because "[its] sample size is larger," some people selected the WDBC because of "[m]ore attributes and [the] slightly balanced class," and people who chose to randomly pick one attempted "[t]o have an unbiased review". The average confident level of their selections was 5.8/10, which was moderate. With the knowledge that the AQI score of WBCD is 95.5785 and of WDBC is 93.87177, six of the participants selected the WBCD, two selected the WDBC, one responded that "I am struggling," and the final user wanted to consult domain experts. The reason of selecting WBCD was mainly because of the higher score, while the major reason of still selecting WDBC was that, "[w]ith comparable score[s], [users] prefer to stick to [their] original choice." The average confidence of the selections improved from 5.8/10 to 7/10.

For the second pair of datasets, before providing the AQIs, the choices from the ten participants were also quite evenly distributed: three selected the BCD, four selected the WPBC, and the remaining three randomly picked one. The average confident level of their selections was 5.6/10. With the knowledge that the AQI score of the BCD is 70.92659, and of the WPBC is 80.84716, only one participant selected the BCD, while the remaining nine all selected the WPBC. The reason for selecting BCD was because the participant "do n[o]t want a dataset with too high dimension but low [number of]instances," while the reason of selecting WPBC was that its score was much higher. The average confidence of the selections improved from 5.6/10 to 6.4/10.

At the end of the survey, the participants rated the overall helpfulness of the AQI

at 6.4/10, which is acceptable but not exceptional. The major reason of low ratings is that participants do not understand the logic and validity of the AQI, and they have doubts when relying on it. Six participants said that the AQI improved their speed of selecting a dataset, while the remaining four were not sure. Promisingly, seven users believed that the AQI improved the confidence of their decisions. From their views, advantages of the AQI included that it is "[q]uantitative, standardized, easy to use," and "a quick measure to exclude low quality data", which "saves time and energy to select data," especially when "the indexes are significantly different." In terms of limitations, the participants commented that "[it is n]ot widely used yet," "it [i]s only one aspect of data quality," and it "maybe lead to mistake when the difference among scores is pretty small."

As a summary of the survey, the responses from participants indicated that the AQI could increase confidence and speed when selecting datasets. However, efforts are needed to help users understand the underlying theoric support and the usage scenarios, so that users could ultimately trust and take all advantages of the AQI. In addition, users should be guided that the AQI measures only one aspect of data quality. It is an index that helps users rapidly compare association-based quality among datasets, while other aspects, such as timeliness, should also be considered, depending on specific research purposes.

It should be noted that, while an AQI of 100 means absolutely the best and a zero AQI means absolutely the worst, the values in between them should be interpreted carefully. The purpose of the AQI is to compare different healthcare datasets. In this regard, the datasets with higher AQIs are absolutely more preferable than the others. Nevertheless, if these datasets contain ordinal features, then the AQIs undoubtedly depend on the appropriateness of the discretization on these ordinal features. Therefore, when comparing datasets using the AQI, one should take the performance of binning methods into consideration. Extra caution should be exercised when one

would like to interpret the AQI of a single dataset without a competitor. The AQI score should not be interpreted linearly or as a percentage. For example, the interpretation of an AQI of 60 is not clear, and the number 60 is useful only when there is another AQI for a comparison dataset. Therefore, there is no standard (threshold) on which values of the AQI could be considered as good or poor.

To briefly investigate the impact from different binning methods to the AQI, we calculate the AQIs for WDBC and WPBC with various binning settings. As we have discussed in Section 5.2, AQI scores change with binning methods. However, based on the results in Table 5.2, the changes in AQIs are relatively small when there is no excessive discretization (less than 15 bins in the example). This suggests the AQI is relatively robust under reasonable binning settings.

Table 5.2: AQI scores under different binning settings.

bins	3	5	10	15	20
WDBC	94.12464	93.87177	94.03538	93.00716	93.15479
WPBC	81.02628	80.84716	80.2148	76.85335	78.77477

5.5 Conclusion and Future Work

We have proposed the AQI to quantify intrinsic quality of healthcare datasets from an association perspective to help dataset identification. An AQI score indicates the degree of association (both linear and nonlinear) between features and the targeting outcome in a dataset. While the AQI does not suggest an underlying model, it helps researchers understand the research potential of healthcare datasets without a time-consuming, rigorous, and detailed analysis. In addition, the AQI is helpful in not only healthcare but also other areas that need dataset comparison assistance. Calculation of AQI is implemented in R package CASMI [25]. The AQI is designed to capture associations among non-ordinal data, and continuous data need to be preprocessed to categorical. It should be noted that an AQI score of a dataset is not unique as it depends on binning methods, when there are ordinal features, and is associated

with the targeting outcome in the dataset. If the outcome changes to a different attribute, the AQI score changes (almost surely). Specifically, because the AQI is a measurement of intrinsic quality for a dataset with a specific targeting outcome, one could calculate AQIs for all possible outcomes and use their average, median, or other measurements to represent the overall intrinsic quality of a dataset.

The brief investigation using real datasets (Table 5.2) suggests that the AQI is relatively robust under reasonable binning settings. Therefore, to achieve a relatively accurate AQI does not require a perfect discretization. While a quality discretization absolutely results in a more accurate AQI, it could require domain specific knowledge. For this reason, we suggest dataset providers calculate and publish AQIs for popular outcomes in their datasets to help researchers quickly understand the intrinsic quality and research potential. Moreover, the binning methods used by the providers on different attributes could also help researchers further understand the background knowledge of data. Finally, data providers could also identify and examine the design of the related datasets with higher AQIs, and features in these datasets could help the providers improve future designs.

Aside from dataset comparisons and ranking, the AQI could also be used to discover research opportunities in a given dataset. Some researchers may have access to a large dataset without a specific research goal and are willing to recognize any research opportunities in the dataset. Without the AQI, these researchers have to keep trying almost randomly until they make discoveries. Moreover, if there are ordinal data in the dataset, usually only linear relationships are examined during the trials because nonlinear relationships are generally difficult to capture. With the help of the AQI, researchers could calculate the AQI for each potential outcome attribute in the dataset to evaluate corresponding possibility toward valid results, a perspective of intrinsic quality. Although they have to discretize ordinal data to calculate the AQI, such discretization needs to be conducted only once for each dataset (unless they would

like to re-perform the discretization with a different binning method to improve the performance). After AQIs of the potential outcome attributes in the dataset are calculated, they could determine a research direction from the possible outcomes with higher AQIs, because an outcome with a higher AQI has a stronger association with features and a brighter potential toward valid results. We summarize the usage of comparing datasets and exploring research opportunities in Fig. 5.2.

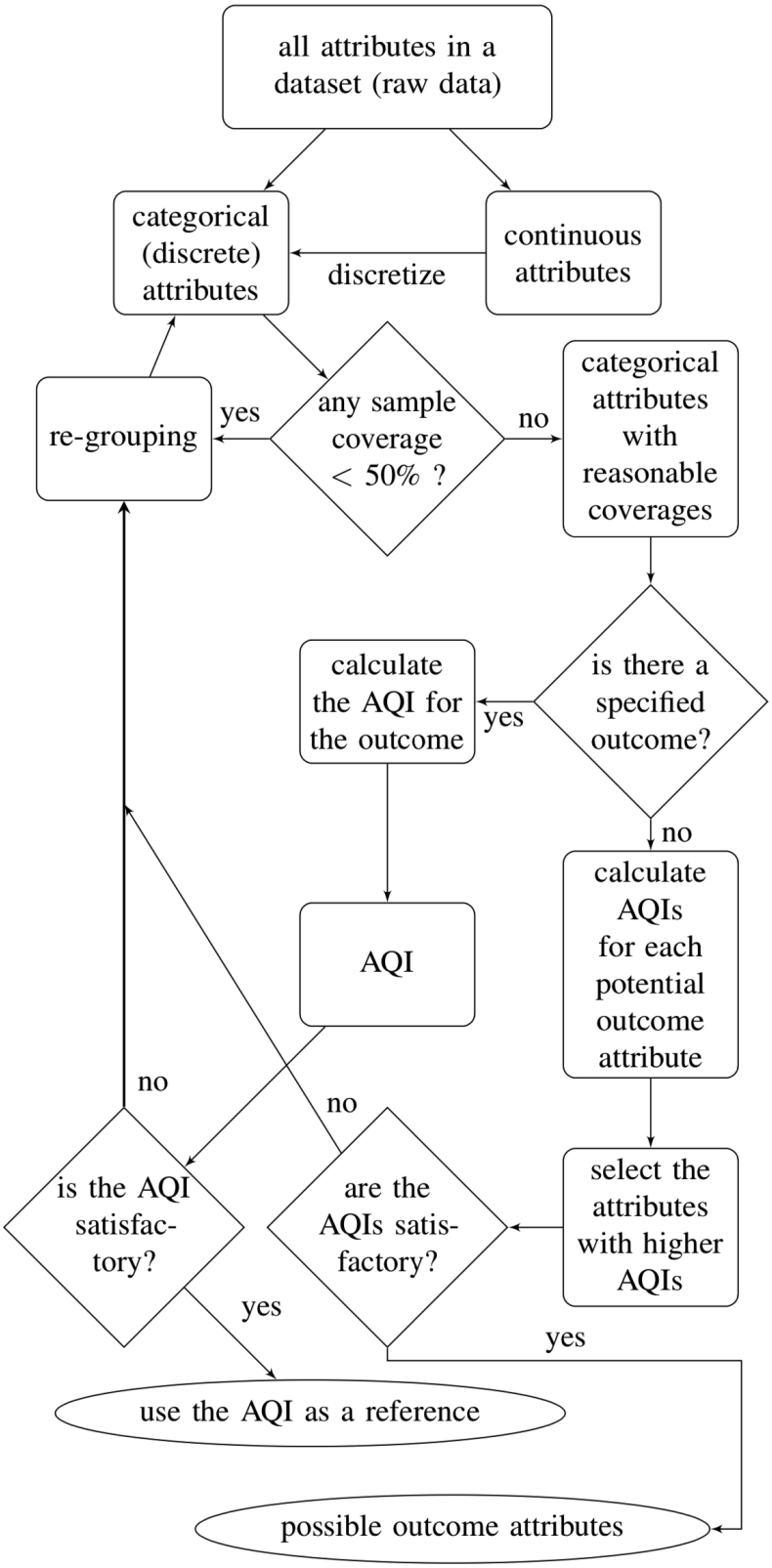

Figure 5.2: Two usage scenarios of AQI.

Future research could explore theoretical interpretations for values of AQI in-between 0 and 100. With an interpretation, thresholds could be determined to help further understand intrinsic quality of a dataset. Also, researchers could investigate performances of different options in evaluating the overall intrinsic quality of a dataset without a specific outcome; for example, the mean and median of the AQIs calculated with respect to each possible outcome attribute. In addition, research on quantifying other perspectives of intrinsic quality is needed to further assist dataset identification.

Chapter 5 Reference List

[1] "Google dataset search." `https://toolbox.google.com/datasetsearch`.

[2] D. Gardner, H. Akil, G. A. Ascoli, D. M. Bowden, W. Bug, D. E. Donohue, D. H. Goldberg, B. Grafstein, J. S. Grethe, A. Gupta, *et al.*, "The neuroscience information framework: a data and knowledge environment for neuroscience," *Neuroinformatics*, vol. 6, no. 3, pp. 149–160, 2008.

[3] "Healthdata.gov." `https://www.healthdata.gov/`.

[4] L. Ohno-Machado, S.-A. Sansone, G. Alter, I. Fore, J. Grethe, H. Xu, A. Gonzalez-Beltran, P. Rocca-Serra, A. E. Gururaj, E. Bell, *et al.*, "Finding useful data across multiple biomedical data repositories using datamed," *Nature genetics*, vol. 49, no. 6, p. 816, 2017.

[5] J. Shi, M. Zheng, L. Yao, and Y. Ge, "Developing a healthcare dataset information resource (dir) based on semantic web," *BMC medical genomics*, vol. 11, no. 5, p. 102, 2018.

[6] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 144–151, 2013.

[7] M. G. Kahn, T. J. Callahan, J. Barnard, A. E. Bauck, J. Brown, B. N. Davidson, H. Estiri, C. Goerg, E. Holve, S. G. Johnson, *et al.*, "A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data," *Egems*, vol. 4, no. 1, 2016.

[8] S. L. Feder, "Data quality in electronic health records research: Quality domains and assessment methods," *Western journal of nursing research*, vol. 40, no. 5, pp. 753–766, 2018.

[9] A. Callahan, R. Winnenburg, and N. H. Shah, "U-index, a dataset and an impact metric for informatics tools and databases," *Scientific data*, vol. 5, p. 180043, 2018.

[10] J. Shi, M. Zheng, L. Yao, and Y. Ge, "A publication-based popularity index (ppi) for healthcare dataset ranking," in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 247–254, IEEE, 2018.

[11] D. D. Lewis, "Feature selection and feature extraction for text categorization," in *Proceedings of the workshop on Speech and Natural Language*, pp. 212–217, Association for Computational Linguistics, 1992.

[12] H. H. Yang and J. Moody, "Data visualization and feature selection: New algorithms for nongaussian data," in *Advances in Neural Information Processing Systems*, pp. 687–693, 2000.

[13] J. Shi, J. Zhang, and Y. Ge, "An entropic feature selection method in perspective of Turing formula," *arXiv e-prints*, p. arXiv:1902.07115, Feb 2019.

[14] J. Zhang and C. Chen, "On 'a mutual information estimator with exponentially decaying bias' by zhang and zheng," *Statistical applications in genetics and molecular biology*, vol. 17, no. 2, 2018.

[15] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," in *Machine Learning Proceedings 1995*, pp. 194–202, Elsevier, 1995.

[16] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[17] R. J. Little, R. D'agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, *et al.*, "The prevention and treatment of missing data in clinical trials," *New England Journal of Medicine*, vol. 367, no. 14, pp. 1355–1360, 2012.

[18] H. Kang, "The prevention and handling of the missing data," *Korean journal of anesthesiology*, vol. 64, no. 5, pp. 402–406, 2013.

[19] M. Lichman, "UCI machine learning repository." `https://archive.ics.uci.edu/ml/datasets.html`, 2013.

[20] O. L. Mangasarian and W. H. Wolberg, "Cancer diagnosis via linear programming," tech. rep., University of Wisconsin-Madison Department of Computer Sciences, 1990.

[21] "breast-cancer-wisconsin.data, breast-cancer-wisconsin.names." `https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/`.

[22] "wdbc.data, wdbc.names." `https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/`.

[23] "breast-cancer.data, breast-cancer.names." `https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer/`.

[24] "wpbc.data, wpbc.names." `https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/`.

[25] J. Shi, "Casmi r package." `https://github.com/JingyiShi/CASMI`, 2019.

CHAPTER 6: CONCLUSIONS

In the big data era, we believe that dataset identification becomes a severe challenge in health data science, and providing complete knowledge of datasets should be an ultimate goal to address this challenge. As with a catalog of books in a library where people can find the desired book easily, with complete knowledge of datasets, users are expected to identify the right datasets quickly. In this dissertation, we describe several contributions toward the ultimate goal of complete knowledge, including both system and method developments about knowledge in both content and quality levels. Particularly, we systematically examined the needs of dataset knowledge for a target group of users (health data science novices), extracted content level knowledge from documentation and publications both automatically and manually, created methods (PPI and AQI) to quantify quality of datasets in two different perspectives, and developed a DIR framework to efficiently represent knowledge based on Semantic Web technologies and the extended W3C standard.

As a result, the current DIR prototype includes metadata and related knowledge of selected representative datasets in healthcare and provides both keyword and semantic search modules as well as the question answering function for frequently asked questions. The prototype is released on https://cci-hit.uncc.edu/dir. In addition, in terms of the two quality quantification methods—PPI and AQI—we not only elaborate on the detailed algorithms in the corresponding chapters but also developed R packages for users to calculate the scores easily. The R package for PPI is available via https://github.com/JingyiShi/PPI, while the package for AQI and its fundamental feature selection method can be accessed via https://github.com/JingyiShi/CASMI.

While the developed system and methods can already benefit the target user popu-

lation in the dataset identification process to a certain extent, there are several major limitations. First, in this initial phase of the DIR, the current prototype still partly relies on manual knowledge extraction, which is time-consuming and labor-intensive for further dataset extension. Second, the current question answering component is based on a pre-defined question pool, while a question answering functionality based on the real natural language would be more user-friendly. Third, quality assurance for both content- and quality-level knowledge has not yet been systematically studied. Particularly, the current content-level knowledge in DIR was only assured by a team-based review, and the quality measurement methods were assured theoretically. Fourth, both PPI and AQI have limitations to certain types of datasets. For example, PPI could be unfair to new datasets that are recently released, and AQI could be used to compare datasets only when they have clear outcomes. Fifth, the existing pieces of knowledge in content and quality levels are still far from covering the full needs of users. Additional elements and perspectives of knowledge are required toward the ultimate goal of complete knowledge.

Recognizing these limitations, we summarize major future directions as follows. First, dataset knowledge should be extended. This includes discovering more perspectives of knowledge other than content and quality, studying more aspects of content relevancy and quality, examining knowledge for broader user populations, involving more datasets, and providing a wider range of metadata. Second, the functionality of the DIR as an efficient knowledge representation framework should be improved. This involves 1) providing a real natural language question answering component, which requires further studies on the separate topic: question answering over Semantic Web (e.g., conferences of Question Answering over Linked Data (QALD) [11] at the European Semantic Web Conference (ESWC) and publications, such as [12][13][14][15][16]); 2) developing a pipeline that can automatically extract dataset metadata, which refers to other topics on named entity recognition, linking,

and typing (e.g., studies from the Open Knowledge Extraction Challenge (OKE) [17] at the ESWC); and 3) establishing collaborative features, such as a discussion forum that allows users to suggest new content. Finally, we should create approaches to assure the quality of the represented knowledge. To achieve a complete knowledge of healthcare datasets, significant additional effort is required from the data science community.

References

[1] "Google dataset search." https://toolbox.google.com/datasetsearch.

[2] "Healthdata.gov." https://www.healthdata.gov/.

[3] "Data | Centers for Disease Control and Prevention." https://data.cdc.gov/.

[4] L. Ohno-Machado, S.-A. Sansone, G. Alter, I. Fore, J. Grethe, H. Xu, A. Gonzalez-Beltran, P. Rocca-Serra, A. E. Gururaj, E. Bell, *et al.*, "Finding useful data across multiple biomedical data repositories using datamed," *Nature genetics*, vol. 49, no. 6, p. 816, 2017.

[5] "bioCADDIE | biomedical and healthCAre Data Discovery and Indexing Ecosystem." https://biocaddie.org/.

[6] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 144–151, 2013.

[7] M. G. Kahn, T. J. Callahan, J. Barnard, A. E. Bauck, J. Brown, B. N. Davidson, H. Estiri, C. Goerg, E. Holve, S. G. Johnson, *et al.*, "A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data," *Egems*, vol. 4, no. 1, 2016.

[8] S. L. Feder, "Data quality in electronic health records research: Quality domains and assessment methods," *Western journal of nursing research*, vol. 40, no. 5, pp. 753–766, 2018.

[9] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 28–37, 2001.

[10] M. Dumontier, A. J. Gray, M. S. Marshall, V. Alexiev, P. Ansell, G. Bader, J. Baran, J. T. Bolleman, A. Callahan, J. Cruz-Toledo, *et al.*, "The health care and life sciences community profile for dataset descriptions," *PeerJ*, vol. 4, p. e2331, 2016.

[11] R. Usbeck, A.-C. N. Ngomo, B. Haarmann, A. Krithara, M. Roder, and G. Napolitano, "7th Open Challenge on Question Answering over Linked Data (QALD-7)," in *Semantic Web Challenges*, Communications in Computer and Information Science, pp. 59–69, May 2017.

[12] V. Lopez, V. Uren, M. Sabou, and E. Motta, "Is question answering fit for the semantic web?: a survey," *Semantic Web*, vol. 2, no. 2, pp. 125–155, 2011.

[13] J. Jeon, W. B. Croft, and J. H. Lee, "Finding similar questions in large question and answer archives," in *Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 84–90, 2005.

[14] I. Androutsopoulos, G. D. Ritchie, and P. Thanisch, "Natural language interfaces to databases–an introduction," *Natural language engineering*, vol. 1, no. 1, pp. 29–81, 1995.

[15] L. Hirschman and R. Gaizauskas, "Natural language question answering: the view from here," *natural language engineering*, vol. 7, no. 4, pp. 275–300, 2001.

[16] S. Shekarpour, D. Lukovnikov, A. J. Kumar, K. Endris, K. Singh, H. Thakkar, and C. Lange, "Question Answering on Linked Data: Challenges and Future Directions," *arXiv:1601.03541 [cs]*, Jan. 2016. arXiv: 1601.03541.

[17] R. Speck, M. Roder, S. Oramas, L. Espinosa-Anke, and A.-C. N. Ngomo, "Open Knowledge Extraction Challenge 2017," in *Semantic Web Challenges*, Communications in Computer and Information Science, pp. 35–48, May 2017.