

**FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO**

# **Hybrid coding taxonomy for clinical search harmonization in Safe Havens**

**Michael André Pinto Domingues**



Doctoral Program in Informatics Engineering

Supervisor: Rui Carlos Camacho de Sousa Ferreira da Silva

Second Supervisor: Pedro Pereira Rodrigues

September 17, 2020



# **Hybrid coding taxonomy for clinical search harmonization in Safe Havens**

**Michael André Pinto Domingues**

Doctoral Program in Informatics Engineering

September 17, 2020



# Abstract

Over the last decades clinical practice has been driven by informatics changes nourished by distinct sources and affected by several constraints. In addition to the issues inherent to confidential data access and physical distribution of individual records, the terminological and classification systems used for coding and data harmonization (essential for the normalization of terminologies) represent two significant problems in clinical research. With that in mind, the Data Safe Haven (DSH) paradigm emerged to promote a newborn architecture, better reasoning, safe and easy access to distinct Clinical Data Repository (CDR) storing data from the Electronic Health Records (EHR).

EHR have proven to be extremely useful and valuable but they are still unable to support universally, a wide spectrum of coding systems and by that reason lack extensive clinical knowledge with more trustworthy data. With these capabilities and a secure access control layer, researchers could link records from different domains, locations and institutions producing improved studies in an easier and safer way. The limitations of EHR, existing Health Information Systems (HIS), the human error, disparate ontologies for clinical assistance and cross-terminology mapping challenges, are degrading data quality and interoperability.

To tackle these problems, research in federated systems, ontologies mapping and virtualization techniques have been applied to minimize the issues and provide suitable solutions in healthcare investigation. Researchers working in this domain have some routine tasks, that can be automatically accomplished through the application of Information Retrieval (IR) techniques. We have identified one of those tasks: within a DSH context, executing a query against all the available repositories based on medical reasoning (e.g. diagnosis and treatment) and return a set of related records. To properly implement this task we have to solve two main problems: i) to formulate a clinical query scheme capable of being executed in all type of repositories; ii) to execute these queries from the previous stage and retrieve all results preserving the relationships.

In this thesis we propose a novel method of IR to address the problem of retrieving relevant clinical records from a cluster of repositories, when codification may be unknown or diverse. We have developed the Comprehensive Medical Information Identifier (CMIID), a search harmonization tool involving a scheme and a framework for retrieving dynamically the information from the repositories based on their characteristics and the nature of the query. The methodology works as follows. A clinical query using the developed scheme and containing the intended clinical knowledge is submitted to the framework. This scheme uses Unified Medical Language System (UMLS) as a foundation, supporting the largest amount of up-to-date ontologies. The framework, using UMLS services to clinically validate the codes and the contexts relationships defined by the user, parses the query and builds a knowledge graph. This indexing enables properties management and search to be fast and optimized. This graph is then used by a search module to build the repository queries based on the registered characteristics - type (e.g. Application Programming Interface (API) or database), existent clinical contexts, etc. Boolean algebra laws are applied to translate the defined properties to the conditional search operators supported by repositories. Results from the distinct sub-queries are then aggregated based on matching characteristics (e.g. clinical context),

returning the maximum of variables possible to allow an enhanced comprehension of the results.

A prototype was developed and tested in a real world scenario, using the US Food, Drugs and Administration (FDA) web service and the Medical Information Mart for Intensive Care (MIMIC-III) database. These sources were carefully chosen because they are publicly accessible, possess real up-to-date data and are used by research centers and healthcare institutions for scientific and professional purposes. Furthermore, a collaboration to integrate the developed software into a Brazilian National Tuberculosis Network was initiated with the intent of improve the way researcher can access the data.

Two different evaluations with experts were conducted and are also presented in this thesis: a Focus Group (with 15 participants) was conducted with researchers and physicians (from different domains) from a Portuguese health technology and research center; a System Usability Scale targeting another distinct and diverse group (10 participants) in order to evaluate the usability of the hybrid coding scheme in clinical practice. This group of people contained: researchers from the beforementioned site, physicians from various hospitals of Portugal and national experts in clinical practice focused on R&D. Overall, participants valued the scope and features of our hybrid coding scheme, considering it is really important for research and clinical investigation. Moreover, they identified it as simple and important tool for their professional routine, valuing the true meaning of data and easing research across multiple sites with an universal and common language, agnostic to repositories. In terms of limitations they mentioned that using UMLS as core medical thesaurus does not necessarily guarantee similar terms are logically equivalent and does not provide 100% representation/linkage between terminology domains. Having this said, contexts may not be fine-grained as needed. Moreover, the lack of a protocol for repositories registration supporting rich metadata to complement the UMLS limitations, was also an identified gap. Score from the System Usability Scale questionnaire positions the scheme concept within a range that classifies in-between "OK" and "GOOD", with a median value of 72.5.

# Resumo

Durante as últimas décadas a prática clínica tem sido sujeita a diversas alterações no domínio informático, que têm vindo a impôr novos condicionantes. Os problemas relativos ao acesso a dados confidenciais e os sistemas terminológicos e de classificação usados para codificar e harmonizar dados (que são essenciais para normalização de terminologias), representam dois problemas significativos na investigação clínica. Com isto em mente, o paradigma "Data Safe Havens" foi desenvolvido, promovendo um novo conceito de arquitetura, com mais aceitação e uniformização de domínio permitindo um acesso seguro e mais fácil a diversos repositórios de dados que armazenam Registos de Saúde Eletrónicos (RSE).

Os RSE têm-se revelado como sendo extremamente úteis e valiosos não estando ainda preparados para suportar um espectro alargado de sistemas de codificação e por esse motivo, carecem de suporte para uma maior extensão de informação clinicamente relevante e de confiança. Com estas capacidades e assegurando uma camada de controlo de acesso seguro, os investigadores poderiam associar registos de diferentes domínios, localizações e instituições produzindo estudos mais avançados, de uma forma fácil e segura. As limitações dos RSE, dos sistemas de informação, o erro humano, a disparidade entre ontologias médicas para suporte clínico e os desafios de mapeamento inter-terminologias têm vindo a deteriorar a qualidade dos dados e a interoperabilidade.

A fim de mitigar estes problemas, a investigação em sistemas federados, o mapeamento de ontologias e as técnicas de virtualização, têm sido aplicadas para minimizar e resolver as barreiras e para providenciar soluções adequadas em investigação na área da saúde. Muitas das tarefas neste domínio, são rotineiras e podem ser automatizadas através da aplicação de técnicas de Pesquisa e Extração de Informação (PEI). No trabalho da dissertação foi identificada uma dessas tarefas: num prisma de "Data Safe Havens", executar uma pesquisa clínica em vários repositórios considerando um processo médico (p.e. diagnóstico e tratamento) e obter os resultados associados. Para implementar efectivamente esta tarefa é preciso resolver dois problemas principais: i) definir um formato para uma pesquisa clínica que seja capaz de ser executado em todos os tipos de repositórios; ii) executar as pesquisas anteriormente referidas e devolver todos os resultados que preservem a cadeia contextual clínica.

Nesta tese propomos um novo método de PEI para combater o problema de obter dados clínicos que sejam relevantes dentro de um aglomerado de repositórios, quando a codificação possa ser desconhecida ou diversificada. No trabalho de dissertação foi desenvolvido o CMIID, uma ferramenta de harmonização de pesquisa que utiliza um formato inovador e uma arquitetura para pesquisar e devolver dinamicamente a informação dos repositórios baseado nas suas características e na natureza da pesquisa. A metodologia funciona da seguinte forma. Uma pesquisa usando o formato desenvolvido e contendo o conhecimento clínico pretendido, é submetida à arquitetura. O formato desenvolvido usa UMLS como uma fundação para suportar a máxima quantidade de ontologias atualizadas. A arquitetura, usando os serviços da UMLS para validar clinicamente os códigos e as relações entre contextos definidos pelo utilizador, processa a pesquisa e constrói uma representação de conhecimento (grafo) para que indexação, gestão de propriedades e pesquisa

interna constituam operações rápidas e otimizadas. Este grafo é consequentemente usado pelo módulo de Pesquisa, para construir as pesquisas inerentes a cada repositório tendo por base as suas características - tipo (p.e. web-service ou base de dados), contextos clínicos existentes, etc. As leis de Álgebra Booleana são aplicadas para transpor as propriedades definidas para os operadores de pesquisa condicionais suportados pelos repositórios. Os resultados das variadas pesquisas são agregadas com base em características comuns (p.e. contextos clínicos), devolvendo o máximo de variáveis possíveis para permitir uma melhor compreensão e análise dos resultados.

Um protótipo foi desenvolvido e testado num cenário real, usando os serviços web da FDA e a base de dados de cuidados críticos MIMIC-III. Estas fontes foram escolhidas cuidadosamente porque são publicamente acessíveis, possuem dados reais atualizados e são usadas por centros de investigação e instituições de saúde para fins científicos e profissionais. Adicionalmente, foi iniciada uma colaboração para integrar o método desenvolvido numa Rede Nacional de Tuberculose do Brasil, com o objetivo de melhorar a maneira como os investigadores podem aceder aos dados.

Duas avaliações diferentes com especialistas também são apresentadas nesta tese: um "Focus Group" (com 15 participantes) foi realizado com investigadores e médicos (de diferentes domínios) de um centro português de tecnologia e pesquisa em saúde; um questionário "System Usability Scale" visando outro grupo distinto e diversificado (10 participantes), a fim de avaliar a usabilidade do esquema de codificação híbrido na prática clínica. Este grupo de pessoas continha: investigadores do referido local, médicos de vários hospitais de Portugal e especialistas nacionais em prática clínica focados em pesquisa e desenvolvimento. No geral, os participantes avaliaram o âmbito e as funcionalidades do nosso esquema de codificação híbrido, considerando que é realmente importante para a pesquisa e investigação clínica. Além disso, eles identificaram-no como uma ferramenta simples e importante para a rotina profissional, valorizando a natureza dos dados e facilitando a pesquisa em vários repositórios usando uma linguagem universal e comum, independente dos repositórios. Em termos de limitações, mencionaram que o uso do UMLS como um dicionário médico principal não garante necessariamente que termos semelhantes sejam logicamente equivalentes e que não providencia uma representação a 100% entre domínios de variadas terminologias. Desta forma, os contextos podem não ser detalhados conforme o esperado. Além disso, a falta de um protocolo para registo de repositórios com suporte para metadados complementares às limitações do UMLS, também foi uma lacuna identificada. A pontuação do "System Usability Scale" posiciona a avaliação da framework dentro de um intervalo que se situa entre "OK" e "GOOD", com um valor mediano de 72,5.



# Acknowledgments

At the end of my thesis I would like to thank all the ones who made this thesis possible and an unforgettable experience for me.

It is with deep gratitude that I acknowledge my supervisors: Prof. Rui Camacho and Prof. Pedro Pereira Rodrigues, whom I genuinely respect and admire. I would like to thank Prof. Rui Camacho, for the constant support, motivation and driving force to explore further challenging details and paths. I also would like to express my gratitude to Prof. Pedro Pereira Rodrigues for the constant support, guidance and invaluable help during the course of this research work.

Many thanks to all CINTESIS researchers, physicians from Porto and Lisbon Hospitals and national health experts, who dedicated their time during distinct phases of this thesis to collaborate and provide insights on the work developed. Your availability, criticism and willingness to help was a remark.

A special thank you to Prof. Niels Peek (University of Manchester) and Prof. Luís Antunes (FCUP) for being part of the Qualifying Committee and their contributions.

Last but not least, I am infinitely grateful to my Mother and Father, for their unconditional love and support. A very special thank you to them, for being present in all moments of my life, and for those warm fully-embracing hugs that mean a lot and encourage me to aim high.

A word to my sweet nephews - Sofia, Matias, Santiago and Victória, that naturally and on their own way transmit love and peacefulness. Hope one day I can help you make a difference in life.

A word to my everlasting friend Bernardo, who is always there, constantly fighting with me and sharing all up and downs. You truly demonstrated once again the care and affection you put into our friendship and that "You'll never walk alone" is something you live up for. Thank you so much for being who you are, my fellow comrade.

Of course that my most deep word of care goes to the love of my life, my wife Diana, who supported me from the very first pitch of my goal until the very last end. Your comprehension, enthusiasm and dedication to us during these times was tremendous and I cannot find words to express my gratitude. I could not have accomplished it without you. Love you so much.



*'Man cannot discover new oceans unless he has the courage to lose sight of the shore.'*

André Gide



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	3
1.2	Goals and contributions . . . . .	5
1.3	Dissertation structure . . . . .	7
<b>2</b>	<b>State of the art</b>	<b>9</b>
2.1	Data safe havens . . . . .	9
2.2	Health information systems . . . . .	12
2.3	Standardized protocols and terminologies for medical data exchange . . . . .	15
2.4	Federated databases . . . . .	21
2.5	Medical data harmonization . . . . .	22
2.6	Federated queries . . . . .	29
2.7	Summary . . . . .	31
<b>3</b>	<b>Comprehensive Medical Information Identifier framework</b>	<b>33</b>
3.1	Clinical practice . . . . .	34
3.2	Data source . . . . .	37
3.3	Hybrid scheme . . . . .	37
3.3.1	Conceptualization . . . . .	37
3.3.2	Taxonomy scheme . . . . .	38
3.3.3	Taxonomy context management . . . . .	41
3.3.4	Taxonomy coding and representation . . . . .	42
3.4	Comprehensive medical information identifier framework . . . . .	46
3.4.1	Architecture overview . . . . .	47
3.4.2	Query validation . . . . .	47
3.4.3	Search engine . . . . .	48
3.4.4	Context validation . . . . .	50
3.5	Search harmonization technique . . . . .	51
3.5.1	Repositories registration . . . . .	52
3.5.2	Query builder . . . . .	55
3.5.3	Query execution . . . . .	58
3.6	Framework integration in a (cross-)research setting . . . . .	59
3.7	Summary . . . . .	60
<b>4</b>	<b>Discussion</b>	<b>61</b>
4.1	Clinical queries . . . . .	62
4.2	Hybrid coding scheme . . . . .	63
4.3	Framework architecture . . . . .	67

4.4	Framework performance . . . . .	71
4.5	Solution positioning . . . . .	75
4.6	Summary . . . . .	77
<b>5</b>	<b>Conclusions and Further Research</b>	<b>79</b>
5.1	Thesis overview . . . . .	79
5.2	Further Research . . . . .	81
	<b>Appendix A</b>	<b>85</b>
	<b>Appendix B</b>	<b>85</b>

# List of Figures

2.1	DSH layers of accountability during every data input and transformation. . . . .	10
2.2	Clinical workflow and likely issues to happen while using a Health Information System. . . . .	20
2.3	Example of a harmonization process that embraces queries transformations accordingly to the data repositories is associated to. . . . .	23
2.4	SAIL system architecture. . . . .	27
2.5	Observational Medical Outcomes Partnership with an example of the Common Data Model. . . . .	28
2.6	Applicability of RDF in Healthcare systems nowadays. . . . .	29
3.1	Structure of the taxonomy with relationships between context-concept-codes. . .	38
3.2	<i>CMIID</i> core representation - high level view in terms of relationships and blocks. . .	39
3.3	<i>CMIID</i> scheme concept based on the representation from Figure 3.2, with a main context called "core" and the possibility to have individual contexts per each code. . .	39
3.4	<i>CMIID</i> scheme concept using John Doe's example with a core context formed by three contexts and one individual context. . . . .	40
3.5	The code representation scheme in the taxonomy. It is defined by the coding system (type) and the context, i.e., type of bound to other contexts and the codes supporting that relationship. . . . .	40
3.6	Example of a <i>CMIID</i> query with a core context formed by Diagnosis, Procedures and Pharmaceutical contexts and containing various thesaurus. . . . .	42
3.7	Example of the code representation scheme, based on a core context formed by Diagnosis and Procedures contexts and with the relationships between codes. . .	43
3.8	Clinical question coded with <i>CMIID</i> , containing a core context about Pharmaceutical, Procedures and Diagnosis ( <i>ctx1</i> , <i>ctx2</i> and <i>ctx3</i> respectively). Additionally <i>ctx4</i> is an individual context section with ICD10CM and CPT codes having a particular context with ICD9CM codes. This allows an advanced refinement of the results. . . . .	43
3.9	Taxonomy query from John Doe case-study, represented using a connected graph. Code relationships are illustrated: both core and individual contexts. For simplicity purposes, not all <i>related_to</i> connection are represented. . . . .	45
3.10	High-level structure of the <i>CMIID</i> framework architecture, using the UMLS services for contexts validation. Search engine is responsible for segmenting queries based on the repositories characteristics and aggregating the information afterwards. . .	47
3.11	Knowledge graph representation of an example of a <i>CMIID</i> clinical query, using a circular layout and with relationship properties illustrated using colored edges. .	48

3.12	Search engine of <i>CMIID</i> framework using a CDR service to extract properties and identify matching sources. The UMLS services are used to retrieve the clinical contexts per each code. . . . .	49
3.13	Example of a <i>MIMIC-III</i> SQL query coded from a <i>CMIID</i> scenario. On the right is shown the template containing the rules for the translation. . . . .	56
3.14	Example of a web service query coded from a <i>CMIID</i> scenario. On the right is shown the templates containing the rules for the translation. . . . .	57
3.15	High-level scheme of a <i>CMIID</i> query, with a multi-contexts perspective. The scheme supports multiple contexts as the core relation and also diverse individual contexts. . . . .	58
3.16	Example of a journey associated to the execution of a multi-context query in CDRs, with partial context match. Each CDR query is adjusted according to the matching contexts. . . . .	59
4.1	Box plot of the SUS results (scores between 0 and 100 represented in the x-axis), from a total of 10 participants evaluating the <i>CMIID</i> query scheme. Fifty percent of the population scored between 60 and 80. Two distinct outliers are identifiable: i) a lower outlier of 30 and ii) a upper max score of 87.5. According to the system evaluation, with median value of 72.5, the scheme is classified in-between "OK" and "GOOD". . . . .	65
4.2	Average participant's scores per question in the SUS study. Participants rated each question with a score from 1 to 5. Scores were then converted to a score between 0 and 100, using a system formula. Question 4 and 9 target the easiness of the self-use of the scheme, which users gave neutral scores. This means users would like to test the scheme along with the framework for better understanding. . . . .	66
4.3	<i>CMIID</i> mechanism to link search results that share common contexts, using cursors based on search identifiers. If there are shared contexts between sub-queries and matching results in the repositories, the framework implements a linkage between records that belong to the same context. . . . .	68
4.4	Benchmark of <i>CMIID</i> framework varying the number of contexts (C) and siblings (S). For a given context C, all $C_{N-1}$ consider 2 siblings. Results indicate there is a cumulative cost over new additions of new codes. . . . .	72
4.5	Average cost (processing time) of each sibling addition with the increase of the number of contexts. Average cost of each sibling varies between 2.5 seconds and almost 4 seconds . . . . .	73
4.6	Time spent performing UMLS requests, varying the number of contexts (C) and siblings (S). With an increase of the number of siblings and contexts, the framework performance suffers a degradation mostly on UMLS requests, one third of the time. . . . .	74
5.1	Integration of Brazilian National Tuberculosis Network with <i>CMIID</i> . Repository has 3 different sources and use an aggregated key to identify individuals. Repositories <b>gal_*</b> , <b>sinan_*</b> and <b>sitetb_*</b> contain laboratory, disease and health cases and treatment records, respectively. . . . .	83



# List of Tables

2.1	Overview of some medical code sets and standards in terms of models, messaging and vocabularies. . . . .	17
2.2	Overview of existing vocabularies and theirs usage depending on the clinical data category. . . . .	18
3.1	Relationship properties supported between codes in the <i>CMIID</i> query scheme. . .	41
3.2	Code relationships based on knowledge graph from Figure 3.9. All <i>sibling_of</i> and <i>related_to</i> properties represent two-way relationships. . . . .	46
3.3	Properties required in the registration of a <i>CDR</i> , assuming execution is inside a DSH. . . . .	49
4.1	Example of clinical questions translated into <i>CMIID</i> queries using UMLS Root Source Abbreviations as code prefixes. . . . .	62
4.2	Usability Survey - descriptive characteristics of participants . . . . .	64
4.3	<i>CMIID</i> framework analysis in terms of solution acceptance - integration and complexity. . . . .	70
4.4	<i>CMIID</i> framework analysis in terms of solution acceptance - maintainability and scalability. . . . .	70
4.5	Framework average processing time (in seconds) varying number of contexts (C) and siblings (S), for all scenarios described in Section 4.1. . . . .	71
4.6	Average time (in seconds) to process each new sibling within the same context. Values for each S column represent the average time each one of the new siblings took. . . . .	73
4.7	Comparison between <i>CMIID</i> framework and an advanced solution in the research field - SAIL Databank. . . . .	76
A.1	Additional clinical queries used on the evaluation of the <i>CMIID</i> framework. . . .	85



# List of Listings

1	Example of contexts configuration in a web service repository. Resource identification and main identifier for the coding attribute are necessary. . . . .	52
2	Example of contexts configuration in a database repository. Identifying the tables and the main identifier for the coding attribute is necessary. . . . .	53
3	Sample configuration file for data sources registration. Required fields are identified within the placeholders "«»". . . . .	55
4	Sample configuration file for the UMLS connection. Required fields are identified within the placeholders "«»". . . . .	60



# Acronyms

**AFL** Anonymous Linking Field

**AHIMA DQM** American Health Information Management Association Data Quality Management model

**AHRQ** Agency for Health and Research quality

**AMA** American Medical Association

**API** Application Programming Interface

**ATC** Anatomical Therapeutic Chemical Code

**BIRN** Biomedical Informatics Research Network

**BPI** Brief Pain Inventory

**caBIG®** cancer Biomedical Informatics Grid

**CBI** Code Binding Interface

**CCI** Charleston Comorbidity Index

**CDM** Common Data Model

**CDR** Clinical Data Repository

**CDRs** Clinical Data Repositories

**CMIID** Comprehensive Medical Information Identifier

**CORBEL** Coordinated Research Infrastructures Building Enduring Life-science

**CPT** Current Procedural Terminology

**CRL** Clerical Record Linkage

**CSV** Comma Separate Value

**DRG** Diagnosis Related Group

**DRL** Deterministic Record Linkage

**DSH** Data Safe Haven

**EHR** Electronic Health Records

**ER** Emergency Room

**ETL** Extract, Transform and Load

**FDA** Food, Drugs and Administration

**FHIR** Fast Healthcare Interoperability Resources

**FURTHeR** Federated Utah Research and Translational Health e-Repository

**HHS** Department of Health and Human Services

**HIPAA** Health Insurance Portability and Accountability Act

**HIS** Health Information Systems

**HL7** Health Level-7

**HMN** Health Metrics Network

**I-MAGIC** Interactive Map-Assisted Generation of ICD Codes

**i2b2** Informatics for Integrating Biology and the Bedside

**ICD** International Classification of Diseases

**ICD-10-CM** International Classification of Diseases, Tenth Revision, Clinical Modification

**ICD-9-CM** International Classification of Diseases, Ninth Revision, Clinical Modification

**ICPC** International Classification of Primary Care

**ICU** Intensive Care Unit

**IP** Internet Protocol

**IR** Information Retrieval

**IRB** Institutional Review Board

**IS** Information System

**LOINC** Logical Observation Identifiers Names and Codes

**MACRAL** Matching Algorithm for Consistent Results in Anonymised Linkage

**MCC** Major Clinical Category

**MedDRA** Medical Dictionary of Regulatory Activities

**mhss** Ministry of Health Shared Services

**MILA** Multi-Institutional Linkage and Anonymisation

**MIMIC-III** Medical Information Mart for Intensive Care

<b>NCHS</b>	National Center for Health Statistics
<b>NER</b>	Named-Entity Recognition
<b>NHS</b>	National Health Service
<b>NLM</b>	National Library of Medicine
<b>NYSIIS</b>	New York State Intelligence Information System
<b>OHDSI</b>	Observational Health Data Sciences and Informatics
<b>OMOP</b>	Observational Medical Outcomes Partnership
<b>OXMIS</b>	Oxford Medical Information Systems
<b>PEI</b>	Pesquisa e Extração de Informação
<b>PHI</b>	Protected Health Information
<b>PII</b>	Prior-informed imputation
<b>PRISM</b>	Performance of Routine Information Systems Management
<b>PRL</b>	Probabilistic Record Linkage
<b>RDF</b>	Resource Data Framework
<b>RSAB</b>	Root Source Abbreviation
<b>RSE</b>	Registos de Saúde Eletrónicos
<b>SAIL</b>	Secure Anonymised Information Linkage
<b>SHRINE</b>	Shared Health Research Information Network
<b>SIF</b>	Service-Oriented Interoperability Framework
<b>SNOMED-CT</b>	Systematized Nomenclature of Medicine-Clinical Terms
<b>SOA</b>	Service-Oriented Architecture
<b>SOP</b>	Standard Operation Procedures
<b>SPARQL</b>	SPARQL Protocol and RDF Query Language
<b>SUS</b>	System Usability Scale
<b>TOPS</b>	Treatment Outcomes of Pain Survey
<b>UID</b>	Unique Identification Number
<b>UMLS</b>	Unified Medical Language System
<b>VPR</b>	Virtual Patient Records
<b>WHO</b>	World Health Organization





# Chapter 1

## Introduction

Over the last decades the amount of clinical data increased significantly because of the promoted use of HIS and EHR, that enable collecting and storing patient data in an efficient and large scale way. However, re-using this data intra/inter institutions has raised several other problems, new healthcare systems have appeared that are able to use patient data without exposing their identity (Lyons et al., 2009a).

The situation we are at now is chaotic. Many directions have been taken with research, industry and healthcare moving forward in different directions. This self-care from each of the parties involved has made a scalable and common agreement very difficult due to its complexity of getting it right. A simple understanding of this situation is easily explained with people's routines. The healthcare plan governments are establishing, the private and public institutions investment and the easy accessibility for the population, creates an obvious expectation of whoever needs to, can use those facilities in any place of the civilized world.

The starting point is the admission process, with the collection of personal and clinical information. From this moment on, any update to the patient dossier in all specialization areas, inputs more information into the system. This data has multiple purposes: i) as historical data for future appointments and clinical care assistance; ii) for research (when legal terms and consents are accepted); iii) for the improvement of the healthcare unit and national healthcare plans (e.g. which are the most needed surgeries from the population, expenses report, etc.). Another perfectly natural routine considers the occurrence of the previous episodes multiple times in multiple units, for the same subject.

An optimistic analysis of these routines would agree that the whole ecosystem can work seamlessly and everything that is needed are an adequate HIS and well-defined protocols. Research studies, carried out so far, support how this is not in fact so simple and many issues are commonly found on new and existing systems, and in the interoperability protocols between them (Rodrigues et al., 2014; Stearns et al., 2001; Alakrawi, 2016; Bodenreider, 2004).

Putting all of this into perspective with the reality, enables the enumeration of the following issues:

- Clinical history being dispersed in multiple locations;

- Information stored in different electronic formats, either inside of the same institution or dispersed in many institutions;
- Patient information still available in physical documents;
- Codification follows different principles and standards, varying with the institution, department and the physician;
- High rate of errors in the records - misleading, incomplete and erroneous information.

Research environments, and clinical practice, often struggle with many barriers strongly related with these issues. New developments target the following three stages: i) harmonization of information; ii) access to information and iii) applicability of data retrieval and mining procedures. Institutions and governments state rigid security measures defining that the data must have no Protected Health Information (PHI), access needs to be limited, contained and controlled requiring a plausible justification - supervisory and review boards may still not allow it even being under compliance. In a later stage, retrieving the necessary information also challenges the researchers. As previously mentioned, due to the diversity of storage formats the access to the information can be disparate (data provided via Comma Separate Value (CSV), email, physical documents, etc.) or even impossible (no external access from a public allowed Internet Protocol (IP) address). Naturally, harmonization results do not come for free. Identifying common attributes to link the same individuals across all sources is a demanding operation, allocating significant resources in normalizing and mapping the data structures and codification.

There are situations where experts do not agree on the harmonization, situations where it is even not possible to harmonize - data has been collected with incompatible methods or some variables are wrongly codified in some repositories. Lately, in other circumstances highly complex statistical analysis is required prior to the harmonization.

Interoperability becomes very laborious, targeting challenges that if achieved, could revolutionize healthcare. Looking at the current situation broad and important restrictions can be denoted; i) data can not be accessed or transferred; ii) records can not be combined or used along with other sources because there are no linkage contexts or attributes; iii) HIS and national healthcare networks lack security features with robust protocols; iv) absence of codification policies that ensure all healthcare facilities and physicians can follow standards and good practices in such a way that all parties enforce data quality; v) existence of many systems and frameworks that continuously degenerate the routines.

Nowadays governments and healthcare entities are more aware of the benefits of accessing dispersed patient data (e.g. the Scandinavian HIS) but even so there are strong oppositions. Ethical, financial, security and government constraints are still constituting obstacles to an improved system where accountability and security policies need to be enforced severely, to a clean and stable healthcare architecture.

With this in mind a paradigm emerged recently in the UK, called DSH, promoting a safe network where researches and clinicians can query all Clinical Data Repositories (CDRs) (inside

the same network) for information and gather clinical data. This access is according to secure authentication measures and anonymization rules so privacy is ensured from the start.

As previously mentioned, safe network demands several solid principles and assurances from the national and international entities, the government and each institution (e.g. hospitals, research centers, etc.) that become a node of the network. Moreover, the HIS used by the researcher needs to be compliant with such network. Such interoperability is crucial for data harmonization, by supporting different CDRs schema as well as distinct data coding. It may also require extra complexity on retrieving information from the system.

Ensuring faithfulness of the trustworthy research environments and the reliability of DSH is by itself demanding and an evolutionary process. Harmonizing the data given disparate data templates in order to achieve the best matching and query results accuracy possible has been also one of the major concerns within this paradigm.

Record linkage algorithms embrace numerous techniques in order to identify matching attributes to create an Unique Identification Number (UID) to select results based on that identifier. Some of them are common to find across CDRs, for example gender, age, date of birth, country, city, social security number. Due to the lack of enough specificity to be used alone several studies are using new techniques and methodologies to improve accuracy.

## 1.1 Motivation

The possibility of integrating CDRs into a DSH network taking advantage of its characteristics has been in focus over the last years (Burton et al., 2015). One of the strengths relies on this harmonization capability and how separated but related content can be reached using an unique search entrypoint. This understanding of the query scope and the categorization of relevant clinical data is important to achieve a better harmonization but it is affected by what are the type of sources and the thesaurus standards used in each source and how can we correlate them to the same data context.

Researchers are severely affected by the clinical issues explained previously. Apart from the difficulty of being granted access through security layers governed by institutions or boards, they also struggle with searching and harmonizing the information. Thus, different areas are hereby present: information retrieval, data harmonization, clinical data codification and HIS.

Developments in every of the above mentioned areas come with many barriers and issues that represent obstacles to innovation. Various have been mitigated over the years whilst others, vanished due to technology evolution. A common observation of studies focused on these areas is that results are very specific to the use-case. None solve the large-scale issue which obviously would require severe amounts of time and resources.

Attempts to solve such impactful issues would take years of work with high chances of achieving no results - many have tried unsuccessfully. Studies have demonstrated that several initiatives are currently taking place in:

1. Understanding how healthcare policies and governance is evolving (Braithwaite et al., 2017; Vuokko et al., 2014);
2. Building mapping layers between thesaurus in order to enrich meaning (Rector, 1999; Kuo et al., 2011);
3. Developing the DSH paradigm, spreading it to real-life healthcare facilities (Lea et al., 2016; Burton et al., 2015; Robertson et al., 2016; Ford et al., 2009);
4. Improve medical data harmonization techniques (Witham et al., 2015b; Randall et al., 2013; Abolhassani et al., 2016a; Mulder et al., 2014).

Experts and researchers have been working together towards solving particular aspects of the beforementioned issues and the evolution of their results is contributing to new solutions proving that awareness and alignment is definitely happening.

Having a way to submit clinical queries to a data cluster and automatically collect all results via a search harmonization process, that interprets the registered sources and executes the suitable sub-queries per source, would invigorate information retrieval mechanisms.

Outlining a strategy that would try to propose an ideal system across all of the 4 areas would be unrealistic and yet another unmeasurable framework. Instead, specific problems were considered:

1. Necessity of having mapping layers between CDRs with different codifications or versions of the same thesaurus;
2. Search mechanism bias to HIS or to repositories implementation;
3. Repositories registration into the network is very dependent on the import of the whole schema;
4. Lack of a framework for research environments that can potentiate search and harmonization.

Base pre-assumptions were also considered:

1. Framework execution under a DSH environment - no security, access and policies restrictions are violated;
2. Use of UMLS as a source of truth - all mapping conditions and definitions are considered as set in the metathesaurus;
3. Clinical records that are made available online, i.e., via internet access.

## 1.2 Goals and contributions

The search for information in a clinical research environment can be divided into three separate moments: properly gathering and describing the data making sure access is permitted, formulating the search query having in consideration the repository characteristics and mapping layers in-between, and the harmonization process required to ensure that the records can be linked towards meaningful insights with individual representations. Each one of the moments come with barriers in a small and large scale, specific to distinct domains. We focus in clinical environments and the workflow as a whole, making sure less manual intervention is required. To achieve this, we propose making the search mechanism support clinical characteristics in a way it remains simple to formulate queries and at the same time repository agnostic for the user.

We want to make it easier for researchers to describe their queries in a more autonomous manner, reducing the dependency on the repositories characteristics and allowing them to enter more clinical specificity in the main query. This would allow not only to execute searches that are executed against a pool of resources which implementation we're not dependent on but also enable harmonization criteria using clinical context. Additionally, we aim to aid researchers and improve their routine tasks when searching for clinical records in distinct data sources.

**Thesis statement.** *Using a hybrid thesaurus coding scheme embracing a multi-terminology approach, we are able to supply a search solution that harmonizes queries across multiple distinct clinical data sources.*

One of the main purposes of this work is to prove this hypothesis. We argue that is possible to take advantage of the search query formulated using multiple thesaurus to enhance records retrieval meaning with a dynamic repository detection via clinical context.

To pursue the main question, the following research questions were derived from the main one, through the course of the thesis research:

1. Which properties do thesaurus have to enable a network of clinical relationships to be built?
2. Which data repository characteristics are required to be possible the execution of multi-terminology queries?
3. Is it possible to link search results from distinct repositories based on known characteristics?

These research questions are a separation of the hypothesis into sub-problems. Research question 1 aims to determine if existing thesaurus share common characteristics that enable the definition of clinical context between them. Despite similar thesaurus don't present the same level of granularity in what comes to clinical meaning and coverage, they tend to follow descriptive patterns that enable some guidance in its use.

Many steps towards answering research question 1 focused on how mapping systems could be developed in order to embrace the fullness of all parties and not necessarily how they can be used to form a network of thesaurus with a core definition that supports linking between contexts. Answering this meant understanding the state of the art of mapping systems and how they have

been evolving to tackle the use of joint thesaurus. Moreover it was important to understand how DSH are being adopted and what challenges appeared in this matter. Research question 2 covers the means by which we want to enable a pool of repositories and those characteristics be agnostic to the way researchers build and execute the query.

A first approach towards answering research question 2 led us to understand how DSH manage complex source infrastructures, dealing with new registrations, their adoption to the query language and continuous updates on schema changes. The tendency of building mapping layers to integrate systems has been revealing how important is to set requirements that satisfy the integration. Nonetheless, in a clinical environment changes occur often and systems need to be designed in a way they can have some distance from the domain and the business layers. Figuring out examples of such requirements was important and meaningful to start performing some experiments and perceive how new sources could be registered having in consideration a series of set of requirements and also which preparatory routines were necessary.

Previous questions led us to formulate the research question 3, which aims to draw conclusions if the application of a scheme definition that all experts understand and are familiar with and capable of supporting clinical questions as an input to a search harmonization process, would work. In this matter, the distance between the query scheme and the repositories definition we were seeking for could severely destroy the idea of harmonization and several issues that could be solved with this.

Perceiving the thesaurus definitions, repositories requirements and well-defined foundations we concluded it was important to set 2 specific objectives of this thesis, from the main general ones:

- To develop a hybrid coding scheme capable of holding advanced clinical questions using solely international standards;
- To develop a clinical search harmonization process using distinct data sources such as databases and web services.

In the quest for the fulfillment of those objectives, our thesis work led to the main contributions that can be summarized as follows:

- **Contribution 1** A hybrid coding scheme that can be used to code clinical questions with context awareness, without the need for new mapping systems, *OWL* techniques or others;  
We have developed a new scheme that supports the majority of known clinical standards as well as proprietary. We used UMLS services as foundation and in that way CMIID can embed all key terminology, classification and coding standards, scaling and being updated automatically over each release. At the same time provides researchers a way to build questions with the expertise they already have and directly related to the data codification characteristics;

- **Contribution 2** We have developed a search harmonization technique using *CMIID* that is able to search for records in distinct registered data sources (e.g. database and web services). Linkage between sub-queries, if possible, is context-based, i.e., uses the contexts identified from the query which relate to UMLS semantic groups. Sources registration requires the identification of these groups based on the data it holds - for example drugs, procedures, etc..

### 1.3 Dissertation structure

This dissertation is organized in four chapters.

Chapter 2 presents the state-of-the-art in the main fields of the thesis, with : Data Safe Havens, Health Information Systems, Standardized protocols and terminologies for medical data exchange, Federated Databases, Medical data harmonization and Federated queries.

Chapter 3 describes in detail the development of the solution: the definition of the hybrid coding scheme, the *CMIID* framework and its architecture and the construction of a search harmonization technique capable of using *CMIID* on distinct data sources

Chapter 4 presents the evaluation of *CMIID* focusing on four areas: hybrid coding scheme, framework architecture, solution performance and positioning, when comparing with other leading edge solutions. The first two topics of the evaluation were conducted within a Portuguese health technology and research center.

Finally, Chapter 5 presents the thesis' conclusions and points out some further research.





## **Chapter 2**

# **State of the art**

In this chapter we present an analysis of existing related work in distinct clinical fields towards the understanding of concepts, frameworks and techniques, international healthcare entities and their scopes as also, federated system and databases.

### **2.1 Data safe havens**

Clinical research is becoming more enthusiastic, demanding and re-shaped to a new dimension of possibilities. The significant impulsion given by EHR contributed to a new era of big data in healthcare where storage, scalability, anonymization and data retrieval, have become some of the concerns (Fang, 2015).

Even with this optimistic revolution the availability of data for research is restricted by ethical, legal and social issues. In order to take advantage of this empowerment and medical enhancements, a new concept is already emerging.

Many institutions and healthcare providers are collaborating in the development of a framework to support the secure handling of health care information used for clinical research. Balancing compliance with legal and regulatory controls and ethical requirements, while engaging with the public as a partner in its governance (see Figure 2.1). This safe and trustworthy model for conducting clinical research, DSH, aims to yield a solid architecture that easily accepts new data buckets and an accessible agnostic PHI query method for data retrieval.

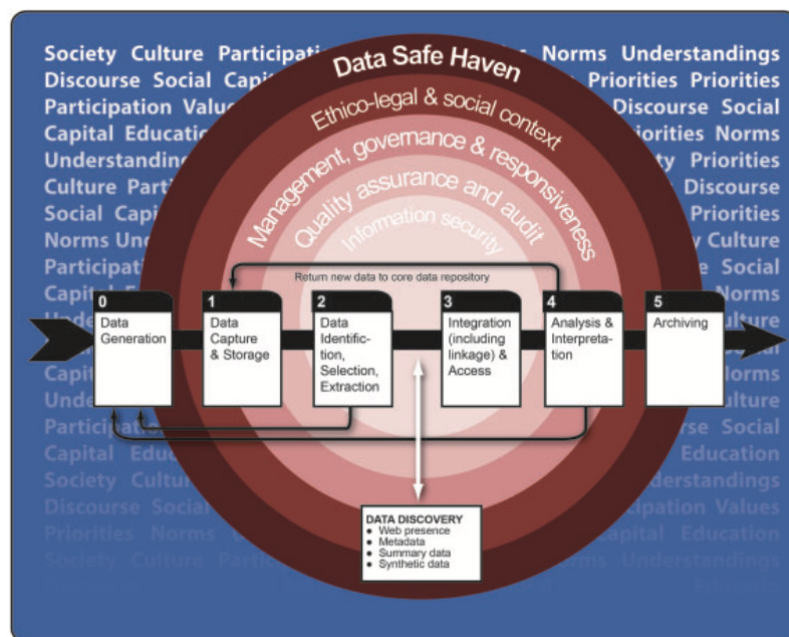


Figure 2.1: DSH layers of accountability during every data input and transformation.  
(Burton et al., 2015)

DSHs are established as secure working environments with different levels of accreditation for researchers and for institutions that provide security and data handling solutions for end systems. These environments have been studied at various levels dealing with legislation issues, certification standards, risk and data management, and others. Moreover the acceptance of an architectural model capable of searching localized patient data, de-identify it, execute algorithmic analysis and surrogate it for the exterior, needs also to be seen as healthcare improvements. It intends to mark a step forward to provide solutions to dilemmas that have been on hold for so many years. It must emphasize the importance of the benefits of information sharing in spite of the risks of re-identification.

Lea et al. (2016) discussed the implementation of the DSH paradigm in research platforms across three jurisdictions in the United Kingdom (England, Wales and Scotland) with reference to a series of case studies across 4 nodes of health informatics research. Besides the strong assets needed for a leading concept as this (core framework, independent ethical review, certification, user accreditation, data management, security, etc.), one of the key points in a network of safe havens with a single entry point, is the acceptance and trust by the participants, funders, the academic research community, and the wider public (Pavis and Morris, 2015).

The safe haven concept is moreover focused on mitigating risks, whether to participants and their re-identification and risks to organizations who process the data. More risks to consider are organizations with control and responsibility for the data, or risks to continuing research and public appetite for the support of research (Lea et al., 2016). This assurance is already published and accepted by researchers and clinical enforcers in the Secure Anonymised Information Linkage (SAIL) Databank model introduced by Lea et al. (2016) at one of the nodes of their study. Being

developed in partnership with several entities, it ensured important metrics as the ones listed:

- Secure data transportation;
- Reliable data matching between datasets;
- Robust anonymization and encryption;
- Disclosure control;
- Data access controls;
- Scrutiny of data utilization proposals;
- External verification of compliance with information governance.

Trying to scrutinize even further these topics, their relevance and definition into a DSH approach, Burton et al. (2015) and Knoppers and Chadwick (2015) introduced a key analysis. Their contribution contains a meaningful starting-point knowledge-base criteria to take in consideration when estimating/evaluating DSH features and the definition of trust, to include the wider public and their trust in security, respectively.

The 12 criteria introduced by Burton et al. (2015) seek to define the faithfulness of the trustworthy research environments and the DSH capability to store and release data faithfully and effectively. Being seen as safe and trustworthy by all key stakeholders: focus on trustworthiness and reliability of the data that is provided, on upholding legal and ethical requirements, and on managing and releasing data within the bounds of social acceptability. Knoppers and Chadwick (2015) developed an understanding of the ethics involved in this area and expanded the scope of trustworthiness to include the public and their views on the security of safe havens, emphasizing governance, security, empowerment, transparency and globalization.

These principles and base considerations that every DSH should take in consideration, are supported by several studies, researchers and external entities that appraise the outcome of such project. Robertson et al. (2016) underlined key principles for an architecture of this kind (using acceptability, usability, sustainability, flexibility, diversity, scalability and validity) presenting a logical architecture for DSH contingent. Authors focus on methods and assumptions, arguing on a formal contract for data sharing, which acts as an overall plan clarifying the roles and tasks of different parties. Witham et al. (2015a) refers to Multi-Institutional Linkage and Anonymisation (MILA) as an important data access and integration procedure, used in several studies, with vital safeguards when sharing data through collaboration.

Such specification provides transparency for review by governance bodies automating as well the data sharing process. Consequently, turns possible a cooperation between data controllers and third-party data integrators, maintaining clearly separated responsibilities that are consistent with governance principles.

To support and improve research based on these concepts and data release paradigm, Burton et al. (2015) acknowledged three ways to access individual-level data (microdata), as a starting point of sharing biomedical data:

- Store the data in repositories and release to potential users, with or without governance controls on that release (e.g. NHS Health and Social Care Information Centre, CARTaGENE, dbGaP, European Genome-phenome Archive, ICGC Data Portal, UK Biobank and UK Data Archive);
- Users access data stored in a repository to analyze it without being able to see or extract any type of information using software with specific surrogates and restricted routines (e.g. UK Data Service Secure Lab, Coordinated Research Infrastructures Building Enduring Life-science (CORBEL), DataSHIELD, ViPAR);
- Data is released directly to data applicants but in modified form that mitigates disclosure risk (e.g. random noise, lesser degrees of data collapse, synthetic data, release of study-level summary statistics).

## 2.2 Health information systems

The amount of medical information captured by healthcare providers is growing at a fast pace. A secure and reliable access to the data must be provided either on role based approaches (doctor, nurse, researcher, etc.) or with a fitter solution that enables the growth of the data management system without compromising patient data (Berg, 2001).

For some decades regulators, policymakers, researchers and clinicians have endeavoured to improve the quality of healthcare by designing and applying patterns of collecting and sharing medical data. Additionally, most of the countries have several advanced and integrated data systems capable of linking social, finance and medical data into one platform for greater effectiveness, efficiency, safety and quality (Braithwaite et al., 2017).

Countries such as the United States, England, Canada, Finland and Denmark have been applying procedures and adopting performance indicators and performance frameworks to make them available at national, regional or institutional level. However there remains the difficulty of linking practice performance to outcomes because of limitations in data availability and poor capabilities to link data (Braithwaite et al., 2017; Vuokko et al., 2014).

An evolutionary work is being done in order to integrate patient data, using differing standards and data architectures so healthcare professionals can access patient records information that may be distributed over many medical databases, under different ethical and legal constraints and of course, different access policies. An integration system responsible for providing access interoperability and homogeneity demands an exhaustive approach of issues handling and law and technical rules compliance (Lippeveld et al., 2000; Bansler and Havn, 2010):

- Architectural demands such as resources systems, hardware, messaging, operating systems, etc;
- Semantic match between data sources;

- Data quality assurance;
- Anonymity and security.

There have been several efforts based on a controlled and uniform platform solution and architecture for accessing medical data driven by healthcare purposes, research and national government infrastructures improvement. Cruz-Correia (2010) addressed the applicability of Virtual Patient Records (VPR) to assess some of the fragile topics already mentioned.

VPR are systems that aggregate medical patient data from different Information System (IS) in real-time. The project goal was to link federated databases allowing therefore clinical documents retrieval into a helpful and integrated overview of patient data. This VPR project identified valuable assets concerning security, data harmonization, access control and data quality. The interpretation of data from different sources is prone to inconsistencies, the lack of thesaurus compliance between all sources requires supplementary measure, with applicability of standard coding systems in semantics (e.g. Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) and International Classification of Diseases (ICD)), in messaging (e.g. Health Level-7 (HL7))). Afterwards it is required anonymization and access roles.

Studies over the last 6 years monitoring significant amounts of information, concluded that the usage of patients past information is correlated to the settings of healthcare. Usage of web services to allow clinical data access by an IS, multi-agents systems (acting as the integration engine ensuring communication between all data suppliers) or direct access to legacy databases, and have a big impact on the outcomes.

With this in mind Li et al. (2012) focused their work in developing a data management system based on metadata. The need for integrating clinical research data and clinical care information, or even reuse clinical trial data can contribute to improve quality, postmarket drug surveillance, clinical research, and public health.

*ClinData Express* was developed to facilitate data collection, storage and management in clinical trials and researches. It is made of a metadata definition algorithm and a data warehouse system: the metadata needs to be specified by the researchers and it is converted into a signed code using a dictionary such as SNOMED-CT for document transformation and sharing. Moreover it comprises access control lists and a custom standardization code system so researchers be able to code their metadata at will, making possible the reuse and data sharing of data collected during the research.

With more acceptance and demand for health information systems, the variety of approaches and strategies will also follow. An inherent development and application of good practices is necessary but resource-intensive and complex for all entities enfolded in. The nature of the HIS along with the integration capabilities of national or abroad services make things even more complex.

Over the years significant HIS frameworks have been developed to encompass solutions for problematic issues in the field of healthcare and evolving to meet patient needs (Cruz-Correia et al., 2007). Marcelo (2010) focused his work in the literature review of theoretical frameworks with conceptualized well-defined strategies that can assure success with minimal costs, specially

in shaping models in various areas: technology, human, organization, etc. Almeida (2016) developed one of the most advanced novel solutions in the world that constantly collects and correlates massive amounts of patient data scattered across hospital systems to automatically identify critical clinical evidences in a time-window ahead of the event. Additionally it is a fundamental tool for hospital management and a precious aid for the medical decision.

Marcelo (2010) mentioned four major frameworks (developed by Aqil, Yusof, Killingsworth and Heeks) that proposed architectural solutions with a complete undertaking of new concepts and abilities. Aqil proposed Performance of Routine Information Systems Management (PRISM) which offered a paradigm shift focused on the routine information systems and its internal organization, technical and behavioral determinants. Additionally, it embeds real-world templates to document health information systems performance as opposed to simply publishing a generic, abstract theoretical framework.

Yusof suggested a qualitative framework based on Human, Organizational and Technology providing a restrict and solid guide with system development that allows adoption factors within:

- The technology acceptance (ease of use, system usefulness, system flexibility, time efficiency, etc);
- The human approachability (user perception, user roles, user skills, clarity of system purpose, user involvement, etc.);
- The organization (support, clinical process, user involvement, internal communication, inter-organizational system, etc.).

Killingsworth et al. proposed a process based on strategic information systems incorporating four dimensions (theory, organization, analysis and management) that have the ability to recognize challenges to each implementation and respond to diverse external calls. Heeks on other hand developed his “design-reality gap” model considering the existing gaps on HIS, highlighting seven dimensions and the gaps that exist between current reality and the proposed design for new systems.

Most recently, the Health Metrics Network (HMN) proposed an overarching framework with solid standards on HIS components and data sources with a clear goal towards harmonization. These theoretical frameworks among all the others, seek to define and set a common development good-practice of every components belonging to a HIS and exposing the concerns driven by the integration of services and evolution of technology.

Healthcare organisations are complex and under some pressure to integrate technology into their domain. Evidences of such use in healthcare has proven to be a path of risks and dangers with more incidence over failures than success stories (Berg, 2001; Sligo et al., 2017). Many authors have studied the factors around information systems development and which markers (e.g. stakeholders, technology, ethics, etc.) really interfere in success and why some have lead to failures. Although many have focused on the individual level of these topics, Berg (2001) labels the

implementation of these systems as a mutual transformation between the organization and technology and not individual adjustments.

Understanding the HIS surroundings (issues, complexities, pitfalls to be aware of, end-users, organizational aspects, etc.), a top down vision and a framework for the implementation is crucial for the correct steering of the solution. Sligo et al. (2017) and Andargoli et al. (2017) addressed literature reviews for HIS evaluation, enumerating distinct studies about the impact factors that lead healthcare systems to failure and available frameworks to assist on the evaluation. On a data quality and research point-of-view, many reasons can be found responsible for affecting the use of HIS for enhanced research outcomes: low financial support, undefined application domains, ethical reasons and lack of interests of extending the scope of health care (beyond its geographical and conceptual boundaries).

## **2.3 Standardized protocols and terminologies for medical data exchange**

The interoperability between several IS has been in focus due to the advantages it offers, improving the efficiency of healthcare delivery while reducing costs and time. Although some problems appear with the variation of hardware, software and terminologies/nomenclatures between healthcare systems.

Kuo et al. (2011) categorized interoperability into three models that define how IS of health organizations can communicate: point-to-point oriented, standard and common-gateway model. The first one forces all entities to commonly agree in coding terminologies, messaging protocol and business process. This is a demanding integration for IS and that is why the standard model considers beforehand an unique standard for data exchange. Alternatively, the common-gateway model holds the idea of independent protocols and infra-structures exploiting a standard message structure between parties.

As mentioned before, we can use HL7 standards to design and develop interfaces for querying and exchanging data from different sources. These methodologies for formulating a Code Binding Interface (CBI) are "syntactic" and concerned with whether data structures can be processed. They are not concerned about how accurate or correct is the information reaching there. Rector et al. (2009) mentioned this focus on the data structures rather than on the meaning itself suggesting coding systems and standards altercations derive in part from lack of clarity about this distinction between validity and accuracy.

Standards on terminology, security and data exchange play a vital role in the integration of health information systems. Before controlled vocabularies can be label standard, requirements of its purpose need to be articulated. The issue of developing and maintaining shareable, multipurpose, high-quality vocabularies has been under heavy study with many requirements annotations. Cimino (1998) scrutinized this desiderata in 12 items which address some priority topics on his

perspective: redundancy, evolution, context representation, multiple consistent views and granularities, polyhierarchy, formal definitions, concept orientation and permanence. The author underlined some innovative changes on nowadays vocabularies such as using systematic approaches for vocabulary updates and conceptual graphs as a transformation approach between different synonymous. Nonetheless, he acknowledges that solutions for all necessary requirements vary on *"technical to political, from simple adoption to basic shifts in philosophy, and from those currently in use to areas ripe for research"*.

Despite years of research and work no clinical terminology has yet been demonstrated in widespread use. Many plausible interpretations may be presented and explained in detail however Rector (1999) put forward 10 reasons why it has been hard. By overseeing healthcare and clinical practices as a whole, stated reasons are bounded to patient centered systems (e.g. EHR), knowledge representation and clinical pragmatics. In other words, separating language and concept representation is difficult, terminologies must be co-ordinated and coherent with medical records, with messaging models and standards and rigorous support for information submission. Likewise search and retrieval systems and conflicts between user needs and requirements are important considerations as well.

Using a common standard for data exchange (see Table 2.1) has been used in several studies to trigger a reduced cost solution that can query social, legal and medical data from different locations, and therefore promote the development of research of clinical data analysis. A real-time data retrieval and harmonization approach is leading to new communication standards and protocols such as the Fast Healthcare Interoperability Resources (FHIR), which defines a set of resources that represent granular clinical concepts, and that can be managed in isolation, or aggregated into complex documents (Kuo et al., 2011).



Table 2.1: Overview of some medical code sets and standards in terms of models, messaging and vocabularies.

Standards for Clinical Research and Pharmaceutical Product development					Standards for Healthcare	
Data Models	CDISC					HL7RIM HL7CDA
	SDTM	ODM	LAB	Define.XML		Templates
	Protocol (SCTP)			ADam		Order sets
Messaging	HL7RPS, Clinical Genomics					HL7 v2.x and v3.0
	E2B (for safety reports)					NCPDP (for Rx)
	DICOM (for images)					DICOM
						IEE (Beside instruments, MIB)
						X12N (for financial data/HIPAA)
Vocabularies	MedDRA (for drug safety)					
	WHODrug (for drug safety)					
	VA/KP/SNOMED (for SPL)					SNOMED CT (for clinical data)
	NCI Thesaurus (for SPL)					ICD9CM (for billing diagnoses)
	LOINC (for SPL)					CPT (for billing procedures)
	NDF-RT (for SPL)					LOINC (for lab)
	FDA DRLS, FDA SRS (for SPL)					NDF-RT, RxNorm for drugs
	CDISC/RCRIM terminology (for CRF)					HCPCS/APC’s (add l claims data)
	HUGN (genomic data)					HUGN (genomic data)

Many of the data management systems available use version 2 or version 3 of the HL7 but the features of the latter are not backwards compatible. V3 provides more of a true standard and less of a customizable framework with application roles that are defined, fewer message options, and less expensive options to build and maintain mid to long term interface (Kuo and Kuo, 2017). With this lack of business logic customization, FHIR proposed a REST architecture making it easier to implement and use by organizations and developers with cost savings and greater integration.

Standard clinical terminologies have many advantages and usages, specially for research. However some HIS database coding schemes can be proprietary and not under a standard that fulfills scalable data operations and linkage. Moreover using outdated thesaurus not according anymore to the recent international medical revisions, represents other known constraint (Schulz et al., 2010; Saitwal et al., 2012; Stuart-Buttle et al., 1996).

Implementing an unified mapping standard with the latest terminologies guidelines and existing coding solutions (see Table 2.2 with some examples) implies a trade-off between risks and gains. With that in mind is important to assess the applicability of the standard in the target data

repositories. Firstly, it signs a semantic incongruence, with an increase of semantic space (mapping from a restrictive definition to a more inclusive one). Secondly, a hierarchy incongruence as result of having a more complete concepts tree, leading to the identification of many patients (Cardillo, 2015; Garla and Brandt, 2012; Rodrigues et al., 2015; Saitwal et al., 2012). Furthermore this approach is vulnerable to false positives (requiring additional resources to assess associations) and false negatives (missing a real association) requiring extra costs to re-assess the clinical validity of the definitions (Reich et al., 2012).

Table 2.2: Overview of existing vocabularies and theirs usage depending on the clinical data category.

Quality Data Model Category	Quality Data Model Data Type	Quality Data Model Attribute	Clinical Vocabulary Standards	Transition Vocabulary
Condition/ Diagnosis/ Problem	Condition/ Diagnosis/ Problem	N/A	SNOMED CT	ICD9-CM ICD10-CM
Encounter (any provider interaction)	Encounter	N/A	SNOMED CT	CPT, HCPCS, ICD9-CM Procedures, ICD10-PCS
Laboratory test (names)	Laboratory test	N/A	LOINC	N/A
Laboratory test (results)	Laboratory test	Result	SNOMED CT	N/A
Diagnostic study test names	Diagnostic study	N/A	LOINC	HCPCS
Diagnostic study test results	Diagnostic study	Result	SNOMED CT	N/A
Procedure	Procedure	N/A	SNOMED CT	CPT, HCPCS, ICD9-CM Procedures, ICD10-PCS

A standard vocabulary is needed to improve the efficiency and reproducibility of analytic methods when applied across a network of disparate CDRs. Aiming to maximize sensitivity and improve specificity there are several vocabularies to represent medical information:

- International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) - its scheme is outdated (latest version International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) - more comprehensive and complex set of codes designed to address some of the issues of previous version) and is used for insurance claims and reimbursement systems. It is often used as a primary thesaurus for recording diagnoses and is developed, monitored, and copyrighted by the World Health Organization (WHO);
- SNOMED-CT - comprehensive, computerized healthcare terminology – containing more than 311,000 active concepts – with the purpose of providing a common language across different providers and sites of care. This thesaurus can be mapped to other coding systems, such as ICD-9 and ICD-10, which helps facilitate semantic interoperability;
- Medical Dictionary of Regulatory Activities (MedDRA) - used to classify adverse events in clinical trials and spontaneous adverse events reporting systems;
- RxNorm - standardized nomenclature for clinical drugs;

- WHO-DDE - WHO Drug Dictionary Enhanced standardized nomenclature is the most comprehensive and actively used drug coding reference work in the world;
- UMLS - developed by the National Library of Medicine (NLM), consisting of a metathesaurus and allied resources to facilitate mapping of medical terms across different vocabularies.

Likewise other customized thesaurus are used in specific countries (e.g.: United States, United Kingdom) such as (Reich et al., 2012; Regenstrief Institute, Inc, 2017; Dotson, 2013; McGinnis et al., 2011):

- Oxford Medical Information Systems (OXMIS);
- Read codes;
- International Classification of Primary Care (ICPC);
- Current Procedural Terminology (CPT) - U.S. standard for coding medical procedures, maintained and copyrighted by the American Medical Association (AMA);
- Logical Observation Identifiers Names and Codes (LOINC) - universal standard for laboratory and clinical observations, and to enable exchange of health information across different systems. Where ICD records diagnoses and CPT services, LOINC is a code system used to identify test observations. This coding codes are often more specific than CPT, and one CPT code can have multiple LOINC codes associated with it.

Reich et al. (2012) evaluated the feasibility of mapping clinical conditions recorded in disparate data sources to a standardized vocabulary, using mapping tables to convert ICD-9-CM diagnosis codes to SNOMED-CT and MedDRA using UMLS as an auxiliary thesaurus. Being able to maintain the integrity as well as the reliability of analytic methods, they highlighted as limitation, the constant need for maintenance of the data source, destination vocabularies and the mapping tables, the coarse-grained ICD-9-CM concepts which explicitly retain less of the original clinical information when mapping from ICD-9-CM and also the semantic precision that lays on the mapping tables created by experienced medical coders (Bodenreider, 2004).

Within the same reasoning, NLM along with National Center for Health Statistics (NCHS) are working in a project to map SNOMED-CT concepts to ICD-10-CM codes called Interactive Map-Assisted Generation of ICD Codes (I-MAGIC), to automatically generate codes from one to another to fulfill the requirements of healthcare and therefore try to serve as a standard data infrastructure for clinical application (Alakrawi, 2016; Campbell et al., 2013).

Independent of how good and complete classification/terminologies systems are, it is important to assess their applicability and the data they will handle. Clinical classification systems as the ICD and clinical terminologies as SNOMED-CT, represent two distinct sets of coding schemes that are used in healthcare, both with divergent purposes and specifications. A combination of both

helps achieving the maximum benefits of information technology in healthcare (Rodrigues et al., 2014, 2013).

However using one of these systems demands *a priori* knowledge of the target healthcare infrastructure and the capabilities of the chosen scheme to accommodate healthcare needs and data structure. The American Health Information Management Association Data Quality Management model (AHIMA DQM) provides the foundation of data and information governance through 10 key principles to monitor data quality in four different domains: data application, collection, warehousing, and analysis - see Figure 2.2 (Alakrawi, 2016).

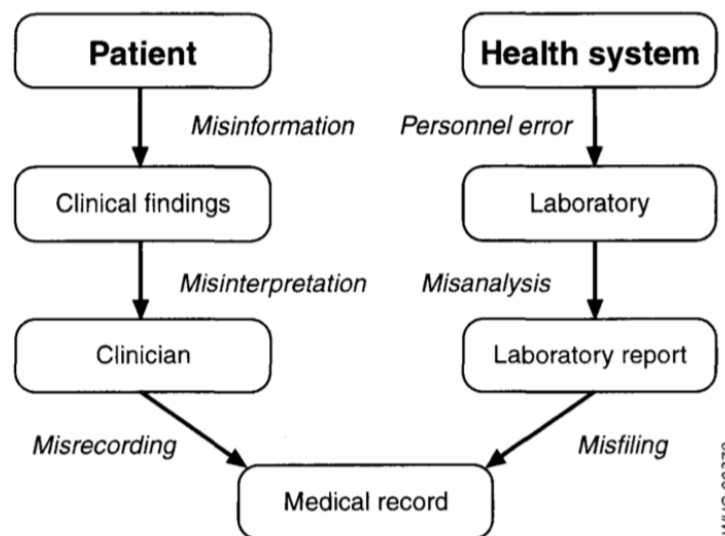


Figure 2.2: Clinical workflow and likely issues to happen while using a Health Information System.

(Lippeveld et al. (2000))

These principles pretend to check the characteristics of data integrity that should be applied in each domain regarding accessibility, accuracy, granularity, precision, timeliness and help identifying the system strengths and weaknesses.

Such domains comprise indicators of how clinical use of schemes can be evaluated, deliberating the harmonization with other existing solutions, and the incorporation of the outcomes into common frameworks for all the healthcare beneficiaries. Similar to these, other functional characteristics (for scheme models structure, maintenance, administration and general adoption) were studied and growth towards less broaden definitions dependent on time, schemes availability, national and international support by healthcare entities. Those that practice, research, legislate and define common terminologies and common guidelines development (Chute et al., 1998).

Nevertheless SNOMED-CT has been one of the most advanced clinical systems in use by consumers, healthcare providers, quality and utilization management personnel, researchers, and other administrative staff, providing high interoperability and clinical coverage. Some other conveniences include extensibility feature that allows users to extend the thesaurus, standardized logical structure enhancing high-level information sharing and retrieval and a fully automated scheme

(Rodrigues et al., 2014; Stearns et al., 2001).

## 2.4 Federated databases

Clinical researchers find themselves struggling to have access to real data sets with the purpose of assessing clinical reasoning to improve the data quality in healthcare. Access to the medical databases is always under restricted admittance policies and anonymization rules set by international organizations and standards for health safety: FDA Protection of Human Subjects Regulations, Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, Department of Health and Human Services (HHS) Federal Policy for the Protection of Human Subjects and others (He et al., 2010).

The interest of clinical research on federated databases is increasing significantly either for research evaluation as also for collecting a bigger train set to fulfill requirements. This is why the number of requests to access these resources is increasing and being restrained by the Privacy Rule. An approval process requires researches to present appropriate documentation to justify the usage of PHI. Managing and monitoring these authorizations and accesses is a responsibility of an Institutional Review Board (IRB) and privacy boards.

Understanding the constraints in-between researchers and authorities, and the importance of federated databases content and architecture safety, is a priority when developing systems to assist on these medical requirements. He et al. (2010) proposed an approach with a federated data repository fed by disparate data sources and with a unified query interface that returns harmonized data. This system (Federated Utah Research and Translational Health e-Repository (FUR-THeR)) considers communication channels to as many IRB as necessary, to make it possible the researcher role negotiation and the query validation with on-demand performance. Additionally, Zhang et al. (2011) addressed the creation of a federated database system that provides a unified access to disparate, geographically distributed data sources, agnostic and platform independent, called BioMart. It has several levels of query optimization to efficiently manage large data sets and various application programming interfaces to ensure that queries can be performed in whatever manner is most convenient for the user.

In another data access perspective, and including academic and industrial research, quality improvement initiatives and higher education coursework, the MIMIC-III critical care database is a remarkable example. It is a large, single-center database comprising information relating to patients admitted to critical care units (Intensive Care Unit (ICU)) at a large tertiary care hospital. With an extent long time support and updates, this project is trying to diminish the restrictions to the data by incorporating digital health records acquired directly during routine hospital care (Johnson et al., 2016; Saeed et al., 2011a,b).

Systems assembled on federated databases should include: message encryption; digital signature to make sure health information is not modified; compliance with interoperability standards

like HL7; role-based access control and patient consent-based access control; federated authentication to authenticate users from different institutions without creating a central user registry; compliance with technical and semantic interoperability; management of heterogeneous data quality; and federated authorization (He et al., 2010; Kuo and Kuo, 2017; Teodoroa et al., 2009).

In a federated database architecture, data sources are independent, but one source can call on others to supply information. Depending on the nature of the system, data warehousing architecture can be employed having data from several sources extracted and combined into a global schema. Many advantages come with this but in particular the uniformity in semantics, and traceability back to individual data sources (Fox et al., 2013).

Even though systems are developed and operating under these assumptions, national and international authorities don't allow centralizing the patient raw data in a permanent location. Teodoroa et al. (2009) explored an architecture considering three main components: wrappers, local CDR and a central virtual CDR (federated database instance). Wrappers are responsible for extracting data from all databases and perform the Extract, Transform and Load (ETL) process loading the data into a local CDR. Sequential data manipulation steps are followed applying a unified schema normalization process, building the information model, drafting the core ontologies and applying data mining methods.

The delegation of responsibilities without exposing the CDR policies and contracts, pushes HIS to implement supplementary federated layers of privacy and anonymization. Shared Health Research Information Network (SHRINE) tool, developed by Weber (2013), holds a federated architecture querying full patient populations of multiple hospitals, without sharing any patient information, just the aggregate count of the number of patients that match the query. This approach despite the fact it allows hospitals to retain control over their infrastructures without exposing security flaws, has strong limitations regarding the aggregate counts. A countable measure within a federated system is not imperative the same as what the result would be if run against a combined central database. Adding partitioning and sampling methods improved results diminishing the aggregation count error.

## 2.5 Medical data harmonization

Gathering all possible patient medical information to assist clinical decision has been in focus in the last years specially from many medical sources possible. This additional peek goes with the intention to have richer information about patients and their medical background for helping health care providers either in manually analysis tasks or with an assisted IS.

Searching for and collecting additional medical data into a query system to help providers achieve better medical aid care, rises legal and ethical questions. Anonimization procedures and security protocols to avoid breaches, patients identification and traceability are two major worries about medical databases linkage. Nonetheless, unambiguous identification of patients is a critical success factor for healthcare reform and for the provision of speedy, safe, high quality, comprehensive and efficient health care.

In order to favor records linkage every medical information should be correctly linked to the patient and identified within the domain by a UID and a minimum information profile which as a whole, represents the identity of the subject (Sauleau et al., 2005). This process is non-trivial when unique person identifiers do not exist and when linkage is based in probabilistic techniques that consider as many valuable variables as those that exist, in an attempt of common identification (Randall et al., 2013; Waïen, 1997).

Unfortunately the medical data input is prone to errors turning the linkage even harder, and requiring additional steps. A standardization (see Figure 2.3), also called data cleaning, step is needed to minimize the impact of mistakes resulting from not adhering to data entry guidelines (i.e. abbreviations, accented letters, date format, etc.) that appear when linking records from different data sources (heterogeneous and non-compatible data-entry guidelines) (Sauleau et al., 2005).

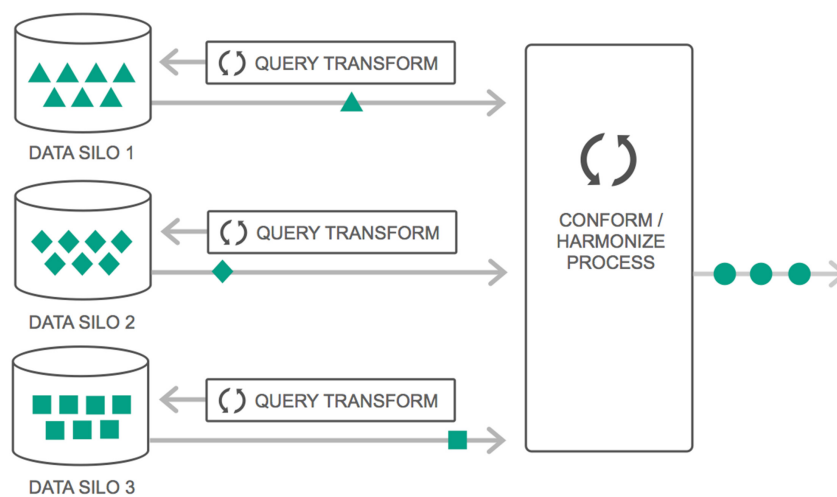


Figure 2.3: Example of a harmonization process that embraces queries transformations accordingly to the data repositories is associated to.

(MarkLogic, 2016)

Avillach et al. (2012) used UMLS as the common terminological system to map events across different terminologies, combining data across databases of various origins. A common database model was used to share and pool data and verify the semantic basis of the event extraction queries (to deal with database heterogeneity). The authors concluded the knowledge described in the various terminologies, which are included in the UMLS, was inadequate to define all of the clinical aspects of an event and so expert knowledge and experience from the database holders were necessary to build a more comprehensive definition of the event.

Witham et al. (2015b) developed a linked health and social care database resource with significant barriers around process, content and cultural aspects. They mentioned that apart from the difficulty of linking records in a technical, confidential and access-controlled way, organizational structures are the second major potential pitfall - to ensure this data was transferred to a Safe Haven.

Gathering all the necessary knowledge hosted in experts (either clinical or HIS) is time-consuming and building relationships to understand the culture and priorities of the sharing organizations, should not be underestimated. Although both aspects are connected, the authors consider it the foundations for: 1) a better research environments and outcomes so that 2) improved healthcare service developments are tangible.

Some standardization techniques tend to increase the number of variables by splitting apart free text fields while others transform variables into a distinct representation without changing the information. Inflicting this separation may require data cleaning techniques in order to maximize the understanding and quality of the data (Randall et al., 2013; Ferrante and Boyd, 2010; Rahm and Do, 2000):

- Reformatting values: data can be changed to a new format without creating or deleting information. New format can help in matching fields;
- Removing punctuation: unusual characters can be misrepresented and cause misleading when matching (spaces, hyphens, apostrophes, etc.);
- Removing alternative missing values and uninformative values: data sets can contain default values for unfilled fields. As so it may not be relevant and even prejudice the matching score in the linkage process. Missing or blank values are harmless;
- Phonetic encoding: creating an encoding of the phonetic information wrapped in an alphabetic variable helps in disambiguation. Several algorithms are used - Soundex, New York State Intelligence Information System (NYSIIS) and Metaphone;
- Name and address standardization: by splitting names and addresses in several parts/categories diminishes the impact of different text representations. The process of splitting can be done using a set of rules or applying statistical methods;
- Nickname lookups: use of nicknames to bring together records which contains different names for the same person;
- Sex imputation: a missing sex field value can be automatically filled by interpreting the person first name;
- Variable and field consistency: records containing inconsistent variables can be edited to remove the inconsistency if it is easy to discover and fix.

There are 3 types of record linkage methodologies: Clerical Record Linkage (CRL), Deterministic Record Linkage (DRL) and Probabilistic Record Linkage (PRL) (Waijen, 1997). Clerical method is the most time consuming because it's a manually process prone to errors but remains the criterion standard, not ideal for large data sets. The deterministic linkage requires a common unique identifier between the data sets: it can be the social security number, the national medical identification number, etc. A significant constraint happens when there is an error in the unique



variable which invalidates the linkage producing unlinked records and therefore missing information.

The probabilistic linkage considers a set of common link variables instead of just one (e.g.: name, age, date of birth, etc). In pursue of an accurate and correct linkage this method requires cut-off probabilities and a weight system based on the data sets being linked. These polishing steps strengthen the criteria to match the records to the same identifier and are included in available harmonization software such as AUTOMATCH, QualityStage and BioSHaRE (DuVall et al., 2012; OBIBA, 2017; Doiron et al., 2013). PRL is more robust against errors, more adaptable when linking large amounts of data and results in better linkage quality than other methods, having higher sensitivity than DRL but lower specificity (Randall et al., 2013; Lyons et al., 2009b).

In 2002 Grannis demonstrated a way of creating these linkages using data that has been de-identified with an one-way hash function, increasing specificity (up to 100%) and sensitivity by generating multiple hash values for each patient using different combination of variables (Weber, 2013).

Alternatively there are other linkage methods depending on the linkage scenarios. The EM algorithm provides accurate estimates of the probabilities and true matches, when the amount of typographical error in the identifiers is minimal (Dusetzina et al., 2014).

Linking several data sets require additional care when understanding in what conditions we can have an accurate match and what produces different output. Firstly, it is important to determine the order of linkage and secondly to develop an algorithm for each linkage process which includes blocking variables and matching variables. Blocking variables are used to partition data into mutually exclusive blocks where matches between data sets are limited to the pairs within those blocks. The contribution of the chosen matching variables to the overall matching process within each linkage (weights), is calculated based on the probability either when the variables agree the pair examined is a true match and when it's not (Dusetzina et al., 2014; Baxter et al., 2003). The composition of these weights is then examined against an interval of confidence based on an upper and lower threshold. Above are true matches and those below are considered unmatched. The others ones, in the interval, are often tagged for a clerical review.

Baxter et al. (2003) described four different blocking methods: standard blocking (clusters records that share an identical blocking key composed of one or more attributes of each record), sorted neighborhood (sorts the records based on a sorting key and compares sequentially all records within a moving fixed size window of all records), bigram indexing (blocking key values converted into a list of bigrams and sub-lists of possible permutations using a threshold) and the canopy clustering (creates overlapping subsets composed for each record of all records within a loose threshold distance).

Independently of the blocking technique that is chosen, identifying and qualifying two different records as duplicates is also hard. It is possible to compare easily, the fields of both records directly as a boolean operation but key-stroke mistakes and spelling changes are ignored even with the standardization process. Some techniques consider the use of similarity measures that rank the

records based on string comparison. Edit-Distance and Levenhstein's distance techniques have effectively been used for a while but the Smith-Waterman algorithm proved to be more reliable due to its ability to introduce gaps in records (Sauleau et al., 2005). Surrogate measure using Major Clinical Category (MCC) improved accuracy and the use of ICD as a finer tool for discriminating between a true match and an inaccurate match (when variables are not available, e.g. name) also has proven to help MCC results.

Measuring the linkage quality can be achieved by understanding how many false positives (two records designated as belonging to the same person when they should not be) and false negatives (two records not designated as belonging to the same person when they should be) the system detects. Subsequently the F-measure is calculated: the harmonic mean between the proportion of correct matches found (precision) and the proportion of correct matches not founded (recall) (Randall et al., 2013). This evaluation approaches consider also true negatives when there is no link present and true positives if we have a correct match (Lyons et al., 2009b).

An additional quality measure is to consider a single variable which has nearly always the same value for the records belonging to the same person, but has always a different value than all records belonging to other people. Taking this into consideration, a high precision is the proportion of times that two variables which have the same value belong to the same person, and a high recall is the proportion of times two records matching each other, have the same value of that variable.

Dusetzina et al. (2014) also indicated an initial assessment of linkage quality by plotting the match scores in a histogram. In ideal conditions the plot will show a bimodal distribution of scores, with a large peak at the smaller scores illustrating the large proportion of likely non-matches and a second base shorter peak for the smaller set of likely matches.

The success of data linkage between electronic health databases depends on data quality, linkage methods and the purpose of the linked data, therefore it is important to evaluate the impact of errors in a linkage system. This is a difficult task because the separation of linkage and analysis due to confidentiality motives, leads to lack of information for researchers to assess the impact of errors in the results. Moreover, the measures of detecting linkage errors (sensitivity, specificity and match rate) are not sufficient (Harron et al., 2014).

As already mentioned, data can be insufficient and prone to errors causing either false positives or false negatives. Although PRL likelihood technique based on matching weights is more robust against errors, the choice of a threshold as an acceptance criteria opens the uncertainty of the cut-off on each situation and can easily manifest to be flawed. Prior-informed imputation (PII) technique combines PRL with information in unequivocally linked records, avoiding errors associated with accepting the wrong record as a link or failing to accept any record as a link (Harron et al., 2014). It transfers values of variables of interest from the linking file to a primary analysis file, rather than linking to a complete record.

Harmonization in DSH is also a work in progress: methods in-use are dependent on the data warehouse characteristics and the own nature of the data. SAIL databank, as mentioned before, is one of the references regarding DSH paradigm and it uses an advanced matching algorithm (Matching Algorithm for Consistent Results in Anonymised Linkage (MACRAL)) that provides

consistently efficient matching results with specificity and sensitivity rates over 95% (Ford et al., 2009).

Lyons et al. (2009b) presented an accurate matching process to enable the assignment of an Anonymous Linking Field (AFL) to person-based records making the SAIL databank ready for record-linkage studies. SAIL has been updated with disparate datasets from multiple health and social care services providers, and represents a research-ready platform for record-linkage studies. MACRAL solution, SQL-based algorithm, proved to be very accurate when using a hybrid approach with DRL and PRL. Besides considering forename, surname, gender, postcode of residence and date of birth as matching variables it also took advantage of PRL methods (lexicon and Soundex - anonymised phonetic code - techniques for example) through a Fuzzy matching process with a 50% threshold.

As seen in Figure 2.4, the approach from the SAIL system uses the HIRU and the HSW layer to ensure that an anonymised identifier is assigned constantly to each individual in the data file - it can be an existing field or a generated one explicitly for the purpose. These two layers also certify the removal of the commonly-recognized identifiers from the datasets by first, organizing the data files into demographic and clinical categories and then generating the AFL (Ford et al., 2009).

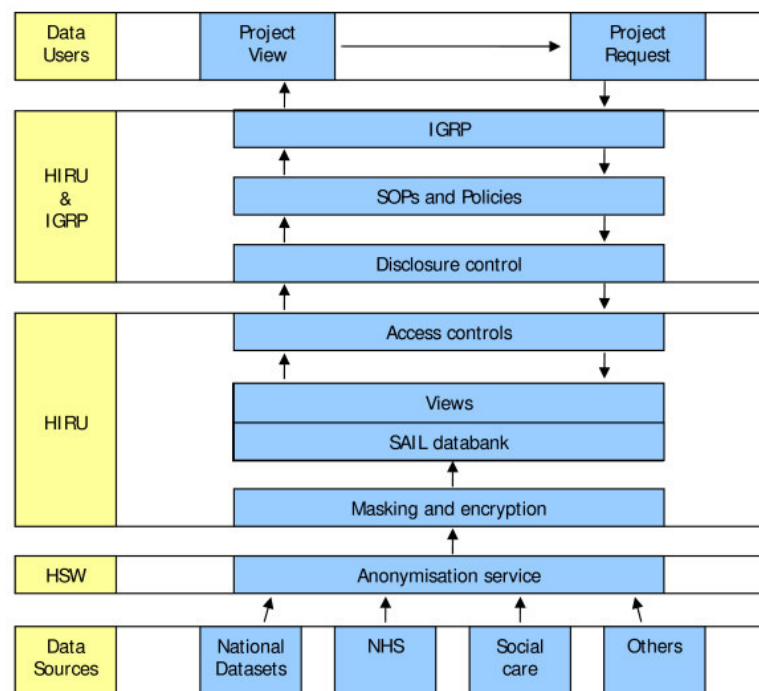


Figure 2.4: SAIL system architecture.

(Ford et al. (2009))

Similar to SAIL, Observational Health Data Sciences and Informatics (OHDSI) is an initiative to enable the analysis and sharing of real world health data (or observation data) between different institutes and companies (Hripcsak et al., 2015). Requires the mapping of the data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) which allows for

the systematic analysis of disparate observation databases. The concept behind this approach is to transform data contained within those databases into a common format (data model - see Figure 2.5) as well as a common representation (terminologies, vocabularies, coding schemes), and then perform systematic analyses using a library of standard analytic routines that have been written based on the common format (Codd, 1970).

OHDSI use this model to provide data mapping services and a dedicated ETL pipeline, so data can be transformed - OMOP includes support for the major commonly used ontologies (SNOMED, Loinc, RxNorm, etc.). Using this framework requires an extensive knowledge of the data set and a good overview of the OMOP data model

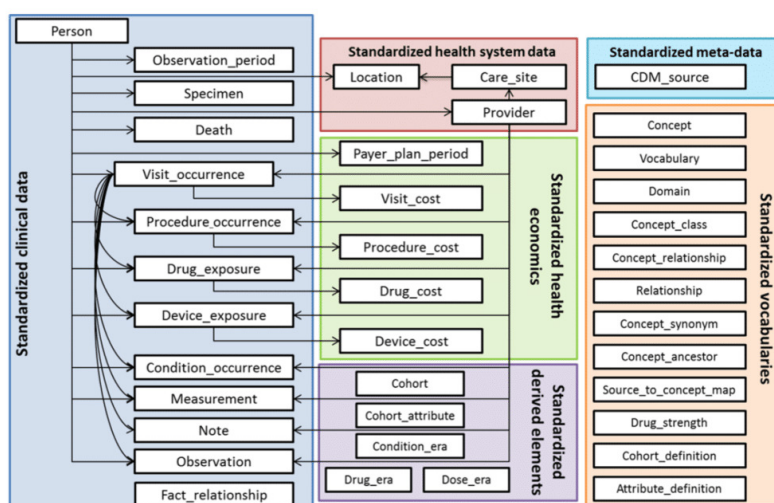


Figure 2.5: Observational Medical Outcomes Partnership with an example of the Common Data Model.

Other federated approach includes Opal as a comprehensive software infrastructure, facilitating data harmonization from multiple and heterogeneous sources as well as seamless and secure data-sharing amongst CDRs. Setting up a network of Opals we gain data access control across Opal servers and data and individual-data hosting by the Biobank they belong to (OBIBA, 2017).

Opal implements DataSHIELD methods which enables individual-level data analysis across Opal instances and integrates Mica web interface, allowing authenticated researchers perform distributed queries on the content of each individual Biobank data collection hosted by Opal.

A different approach accommodates data virtualization using Resource Data Framework (RDF) and a query language such as *SPARQL* (see Figure 2.6) whereas a knowledge graph is used for representing data by creating mappings between all data sources (Stroetman et al., 2009; Ko et al., 2006). Data Virtualization lacks knowledge about inter-database relations. Objects need to be identified and stored in the databases manually (in the application level) while queries are going through the virtualization tools - aggravates maintenance and update (Abolhassani et al., 2016a).

Linked Data paradigm appeared as a solution to make smarter queries providing various benefits such as identification, access integration coherence, provenance, governance and agility. Is a

bottom-up approach for publishing structured data so that it can be interlinked and become more useful through semantic queries, using URIs and RDF.

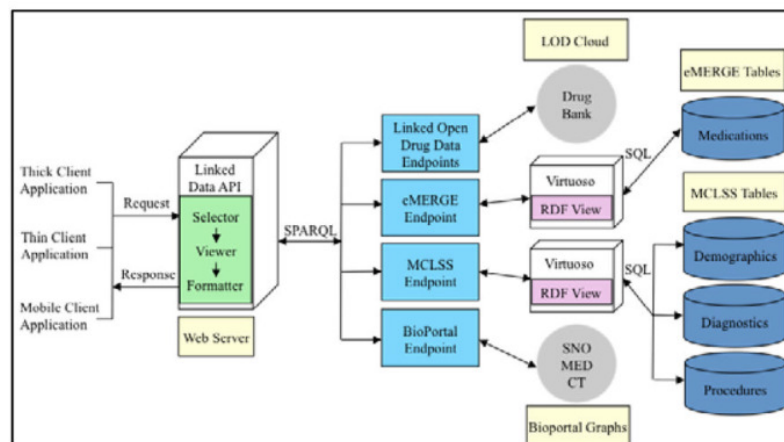


Figure 2.6: Applicability of RDF in Healthcare systems nowadays.  
(Pathak et al. (2013))

Particularly known in ontologies creation, this approach has several advantages in healthcare:

- Captures information content, not syntax;
- Allows data models and vocabularies to evolve;
- Multi-schema friendly;
- Good for model transformation;
- Allows distinct data to be connected and harmonized.
- Support inference

The generation of mappings for all data repositories is a heavy process and as mentioned by Tao et al. (2013), there are relevant studies focused on this subject including guidelines, mapping quality algorithms, semantic evaluation and descriptors for classes creation.

## 2.6 Federated queries

HIS and federated-systems have their own search mechanisms to solve clinical queries and linking records. Most of these integrations are in the data and presentation layer, and consequently not sharing core functionalities. These systems are numerous and can be part of a many-to-one relationship with the healthcare facilities where they are installed (e.g. hospitals). Therefore integrations are the most common practice (Ribeiro et al., 2010).

The capability of querying heterogeneous databases with different schema relies on methods such as Data Virtualization, integrating data from diverse sources, locations and structures into a

common interface that hides the technical details of the stored data (Abolhassani et al., 2016b; Mulder et al., 2014). Existing tools lack the knowledge about inter-database relations which therefore demands the insertion of them between objects and database identification.

The adoptive evolution of EHR is increasing the availability of electronic information and changing the way care systems organize their data and define policies to integrate it in shared query systems, promoting research clinical assistance.

Federated networks of clinical research data repositories are embracing constantly new challenges with the increased interest of researchers promoting federated query tools on demand for patient information in restricted and site-specific repositories. The way these networks and tools are built and used afterwards, dictate how feasible it will be to gather reliable results from multiple queries in multiple repositories laid on growing data cores.

Establishing criteria and definitions in the architecture of federated networks, helps achieving more reliable query results. With this in mind Weber (2015) explored 8 specific properties that define critical points of view for federated queries comprising ontologically equivalent data repositories, system availability, data access restrictions, semantic discernibility and others.

A query on a federated network requires individual resolution techniques in each site repository. It may target characteristics and local ontologies that may not match other repositories. Determining the judgment of the result relies in the query translation mechanism for the system, either if we consider an adaptive query process to each site based on the premises it uses (type of data, ontology and semantic in use, availability, etc.) or an overall approach for all sites (Ciccarelli, 1999; Betawadkar-Norwood et al., 2013; Cragun et al., 2007). More important it helps dealing with scenarios when the researcher is using a query based on a coding scheme that differs from the one in use in some CDRs. This is an advantage if there is no need to sacrifice functionality or lose semantic specificity by forcing sites to use the same software and a common ontology (Weber, 2015).

Weber (2015) mentioned some approaches that could be based on a two-stage process. Firstly, the researcher asks for the properties of each site so it can be understood the knowledge distribution among all cores and secondly, doing a batch of queries optimized to each repository converging as much linked records as possible. Considered an optimization to the architecture, this methodology does not sacrifice functionality or lose semantic specificity, it takes advantage of each site specification optimizing the query and above all is in compliance with access policies good practices.

To go beyond this approach, and to enable a knowledge graph for data harmonization in the data virtualization middleware, RDF is commonly used. It offers a simple semantic model based on a directed acyclic graph structure. Each graph database can have its own specialized query language (e.g. SQL, Neo4J and SPARQL). Besides easing the data representation and capability of integration with many query languages, integrity rules are based on its graph constraints rather than from an imposed relational schema (Mulder et al., 2014).

Ensuring a virtual boundary between CDRs and the federated query system, is possible and starts with isolating the metadata, schema and individual ontologies. This will then have the

responsibility to handle different heterogeneous data sources from different institutions, joining them through common indexes. Indexes are based on query definitions that coalesce federated sub-queries targeting individual data sources. Other solutions may use systems acting as middle-ware to facilitate the integrations of several HIS, not needing to reproduce common components (Shahmoradi and Habibi-Koolaei, 2016).

Several federated data solutions were developed and continually studied, addressing a Service-Oriented Architecture (SOA) advance . They exploit federated queries techniques to prevent data sources exposure focused on integrity and interoperability: Biomedical Informatics Research Network (BIRN), NCI-sponsored cancer Biomedical Informatics Grid (caBIG<sup>®</sup>), Informatics for Integrating Biology and the Bedside (i2b2), Service-Oriented Interoperability Framework (SIF), Mica and FURTHeR (OBIBA, 2017; Livne et al., 2011).

## 2.7 Summary

This chapter presented the state-of-the-art in the fields of Data Safe Havens, Health Information Systems, Clinical data coding, Data harmonization and Information Retrieval. We have presented various harmonization techniques and data coding thesaurus used in clinical research, as well as their advantages and weaknesses. We have also presented a set of general issues and concerns regarding all of the fields previously mentioned. This chapter ends with a brief analysis of the state-of-the-art techniques and solutions regarding federated queries which *CMIID* also pretends to enhance.





## Chapter 3

# Comprehensive Medical Information Identifier framework

We have seen in Chapter 2, the relevance about the DSH paradigm and how research groups and big organizations strive to improve healthcare, given the benefits that can arise from such role-model architecture. Additionally, using multiple clinical data repositories with distinct coding schemes, has been one of the most important focus areas of research - how to improve records linkage using data stored using different standards.

Part of existent studies already embrace these concepts on diverse clinical subjects using mapping classification systems and the harmonization of distinct CDRs (e.g. DuVall et al. (2012); Cruz-Correia (2010); Kuo et al. (2011)). So far none, (to the best of our knowledge) has focused on how we can improve data quality and research simultaneously, by minimizing the problems in data interpretation, coding, re-use and search.

In one hand, some studies developed mapping techniques for the most accepted thesaurus trying to promote a comparable interface. Others, developed advanced query solutions that facilitate clinical questions on information systems. Taking into consideration the constraints on both scenarios and so far to our knowledge, none tried to perceive a relationship between every domain specific thesaurus in order to propose a taxonomy that:

- Embeds more than one standard to provide deeper knowledge on a concept;
- Allows experts to code a clinical question based on their expertise in known vocabularies, available on the target data repository (either public or proprietary);
- Is of easy understanding and scalable - incorporation in federated systems using corresponding mappings and query languages (e.g. SPARQL Protocol and RDF Query Language (SPARQL)).

In this chapter, we will present a novel framework addressing:

- Hybrid Coding Scheme - how to build clinical questions with the new format;

- Architecture of the framework - how the search harmonization process is done having as input a clinical question coded with *CMIID*;
- Search harmonization technique - how a clinical query is translated into technical repository queries based on the *CMIID* subject.

### 3.1 Clinical practice

Understanding the advantages of this framework in terms of applicability of search harmonization techniques and data quality measures in HIS, is also an important field study. Moreover is also relevant assessing if, taking into consideration a settled healthcare infrastructure, is possible to reuse and extract data easily without the constraints already mentioned.

With this in mind and to develop a better understanding on the key topics mentioned in the research questions, a series of interviews about clinical research were performed. This session was conducted in a Portuguese health technology and research center responsible for working with several regional and national healthcare institutions, and the use of advanced techniques to produce valuable outcomes to healthcare.

Twelve researchers from different research groups participated in the interview process, with the goal of understanding the following key points:

- Source of each database in use and the variables that make part of them (how they are collected, used, etc.);
- Security policies roles when accessing data;
- HIS in use;
- Anonymization techniques;
- Harmonization techniques;
- Data coding standards in use;
- Data manipulation layers used in the flow;
- Record linkage facilities.

It was asked each interviewee to describe in detail, for each topic, the existing procedures, solutions and techniques in use, known difficulties including the most impactful ones, and external dependencies they have required to conduct their work.

Research groups had distinct contexts: health complexity sciences applied to physiological systems; clinically relevant temporal abstractions from medical data streams; indicators for data quality and hospital performance measurement; health technology assessment; environmental-related exacerbations of airways diseases and adverse drug events discovery and assessment.

The research projects within these groups were working with distinct databases:

- Hospitalization;
- ICU;
- Emergency Room (ER);
- Appointments and daily patient care treatments;
- Mortality;
- National drug adverse events;
- Self-created databases based on the National Health Service (NHS), population inquiries and hospital administration, and clinical databases;
- Other national databases for analytics drill-down (e.g.: national statistics, Ministry of Health Shared Services (mhss), National Health Service).

Access to the previous databases required an authorization consent which varies depending on the study and the research context. Generically it required a form submission to the competent authorities (e.g. national entities regulators like data protection and regulatory ethic entities, population statistics centers, etc.) and/or to IRB (e.g. hospital administration, hospital departments supervisors, data farm owners, data access facilitators, etc.). Accessing other research sources could also require similar authorization policies or new ones depending on the data regulation and privacy range, for that specific domain.

Upon validation, data was usually provided in tabular format such as CSV and Excel comprising several fields. Some strongly domain specific which may not be self-explanatory and consequently needing further guidance for its interpretation. The inclusion of specialists in the understanding process was a time consuming step and similar to the coding process, there could be divergent opinions and incorrect afterwards analysis. One specific research group had a different data access approach sharing a copy of the raw database with the hospital entity. Once the relevant variables for the study goal were extracted, the copy was eliminated to avoid unwanted leaks. This protocol also obligates the research group to proceed with records anonymization (name, address, social security number, etc.) eliminating any possibilities of tracing.

In all research projects, data records possessed coding systems that best suite the clinical domain - MedDRA, ICD, Anatomical Therapeutic Chemical Code (ATC) and ICPC - as also other systems from non-clinical sources, that have a high impact on advanced studies.

Coded fields were not 100% reliable because they live with a significant coding error rate: data not coded, considerable amount of records in natural language text which are not taken into consideration, and also *upcoding* evidences. Moreover, the amount of "*missings*" in the incoming data sets prejudice also the quality leading to necessary efforts from the researchers to clean up for further use. Unfortunately, the uncertain and ambiguous way patients describe symptoms, the lack of precision on the temporal occurrence of the events, poor interoperability between physicians reporting and the HIS limitations, contribute to a high data variability affecting the data re-use

and quality. Some national entities like *INFARMED* implemented Standard Operation Procedures (SOP) to enable research centers to fix data issues and submit the records back, for approval and replacement.

Researchers often call for clinical expert guidance (doctors, nurses, psychologists and other health technicians) that were responsible for the data input or with the appropriate knowledge to understand the variables meaning, and how the values were collected. Their help is meaningful in every study, specially when there is no protocol and metadata description capable of endorsing self-intuitive critical investigation. The process of combining the variables with external ones, was highlighted as a demanding data analysis process (time and effort). A similar model was also studied by (Krishnankutty et al., 2012), showing that this process of collection, cleaning, and management of clinical data in compliance with regulatory standards is demanding and requires many parties.

Studying variables with multiple origins, lead researchers to standardize as much as possible. The goal is to ease data mining processes and with that perspective in mind, quality-driven approaches like the Charleston Comorbidity Index (CCI) and the ones provided by Agency for Health and Research quality (AHRQ), are taken into consideration. Moreover, proprietary forms have been used for medical surveys among the population with goal-oriented customized templates, comprehending precise evaluation indexes and classification instruments (e.g.: for pain measure). These templates are already validated by the medical scientific community ensuring a higher level of confidence (e.g. Brief Pain Inventory (BPI) and Treatment Outcomes of Pain Survey (TOPS)).

In terms of the usage of federated systems, some research groups used the HIS from the health-care providers (e.g. *JOne* and *Alert*) to take advantage of the integrated features and the efficient management of complex raw data. Others built their own interface which aggregates the desired metrics and variables.

Each research group was focused on different healthcare subjects dealing with distinct entities and data providers. Such responsibility demanded mandatory actions to use the data accordingly without compromising the sensibility of it. Nonetheless all groups suffered from:

- Bad data quality - wasting a significant amount of time processing the incoming records to an usable basis;
- Jurisdictions and heavy procedures that institutions put in practice for healthcare;
- Time delay in getting answers and the data into theirs projects;
- Lack of interoperability between HIS intra and inter institutions;
- Mismatch of protocols (SOP) for data storage and understanding.

## 3.2 Data source

Two distinct CDRs were used during the development of the framework: a database (*MIMIC-III* running in a localhost) and an API (US Food and Drug Administration).

*MIMIC-III* critical care database is a large, single-center database comprising deidentified health-related data relating to patients (forty thousand) admitted to critical care units at a large tertiary care hospital. With an extent long time support and updates, this project is trying to diminish the restrictions to healthcare data by incorporating digital health records acquired directly during routine hospital care (Johnson et al., 2016; Saeed et al., 2011a). It uses a relational database and includes information such as demographics, vital sign measurements made at the bedside (1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (available thesaurus: *ICD9*, *CPT*, *LOINC* and *Diagnosis Related Group (DRG)*).

US FDA API (openFDA project) was created to provide easy access to public data, to create a new level of openness and accountability, and to ensure the privacy and security of public FDA data. For the goal of this study we used it to access the adverse events database containing drug records (available thesaurus: *RxNorm*).

These sources were carefully chosen with the intent of representing a cluster of repositories with contexts in-common and others differing (e.g. UMLS Chemicals & Drugs context). The outcomes from Section 3.1 along with having a cluster with these configurations, was helpful to design and evaluate the framework.

## 3.3 Hybrid scheme

Issues and studies mentioned in Chapter 2 concerning coding standards, search and harmonization of information, mentioned it would be important that any new development within DSH (focused on these areas) should not demand additional layers of knowledge and processing (e.g. no new mapping solutions and/or new terminologies). Nonetheless, should simultaneously be easy to integrate, use and master in a clinical research environment.

With the knowledge acquired from this study we found a leading opportunity to propose a new way to harmonize clinical searches on distinct repositories (e.g. databases and/or web services), using solely the expert knowledge on broadly known major clinical thesaurus (and proprietary ones) to formulate clinical questions.

The following sections detail the organization of *CMIID* framework, from a conceptual level to the description of the solution.

### 3.3.1 Conceptualization

The outcomes from the research field study constituted an important assessment for a technical approach on the taxonomy scheme. For an easy understanding, let us consider the following scenario:

Researcher John Doe wants to query a data repository based on a clinical question about a cardiac event. He does not know how data is coded in that repository. However, his expertise encourages him to code his question using 7 thesaurus: SNOMED CT US, RxNorm, MedDRA, LOINC, CPT, ICD9-CM, ICD10-CM.

He is only interested in getting the records based on 3 clinical contexts: diagnosis, procedures and pharmaceutical.

### 3.3.2 Taxonomy scheme

The previous type of question is very common among researchers and relates different contexts. In fact, John Doe query capabilities would benefit if a relationship between contexts was taken into account, acting as a need to refine for events relatedness. Bearing this thinking, the taxonomy core should lay on the structure shown in the Figure 3.1.

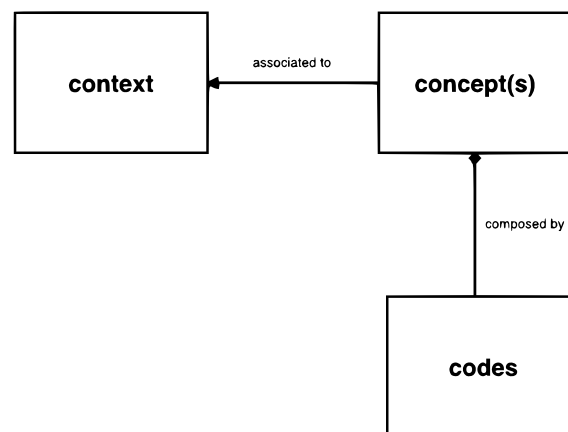


Figure 3.1: Structure of the taxonomy with relationships between context-concept-codes.

For each particular clinical context, the taxonomy allows the association of multiple codes from distinct ontologies/thesaurus to classify the concepts that relate with the context. For example, in the context of a diagnosis of a cardiac event, there can be multiple concepts inherent and consequently, several codes can be used. Then, they all have a bound (relationship) based on the overarching context they represent.

Multiple contexts can share a relationship, which represents the relatedness between contexts (e.g. diagnosis of a cardiac arrest and a specific surgery). Thus, a relationship between codes can be achieved through the definition of the property the researcher wants to bound records on.

Contexts can then be grouped into a bigger context called core - all contexts together should represent the subject of the query. Additionally, individual contexts can be added - allows filtering capabilities on the results. Figure 3.2 shows a high-level representation of the *CMIID* skeleton while Figure 3.3 introduces more detail.

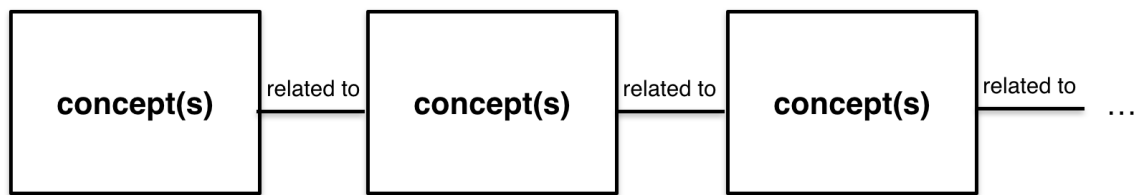


Figure 3.2: *CMIID* core representation - high level view in terms of relationships and blocks.

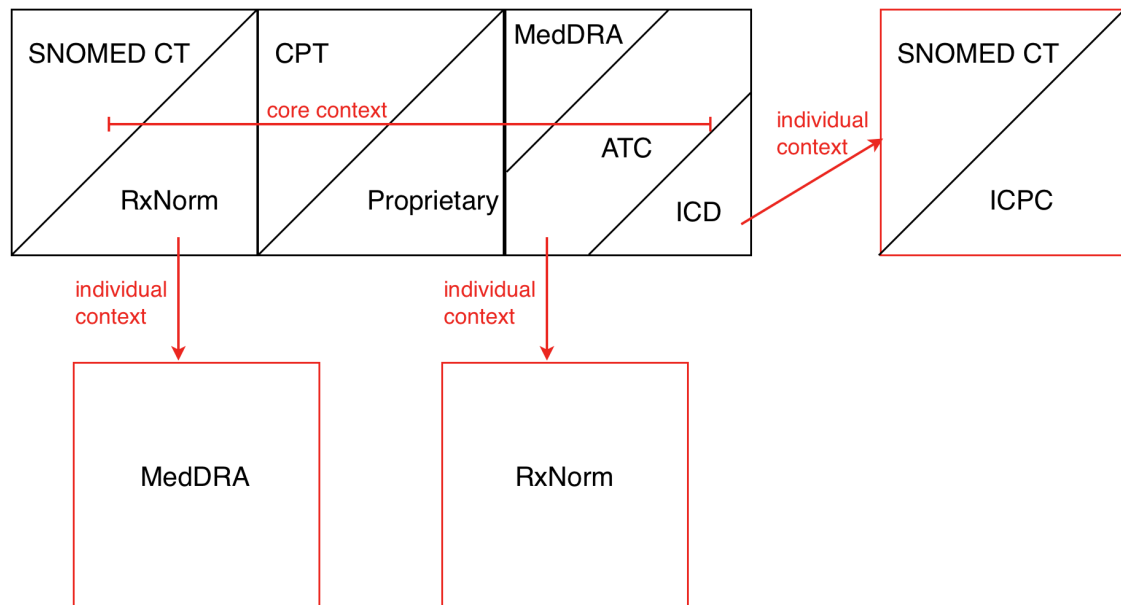


Figure 3.3: *CMIID* scheme concept based on the representation from Figure 3.2, with a main context called "core" and the possibility to have individual contexts per each code.

Having in mind John Doe's question, all concepts represented in the 3 clinical contexts mentioned (diagnosis, procedures and pharmaceutical) define the core context. The 3 set of codes that are related to each others, formulate the question. In the process of revising the results of the executed query with the core context, John Doe could perceive that he wants to restrict the results: when a pharmaceutical code from the core context has occurrences, additional filtering should be applied to also look for occurrences of other code (e.g. of a different domain). This would filter for all records that have both codes, which can be helpful to understand clinical events. See Figure 3.4.

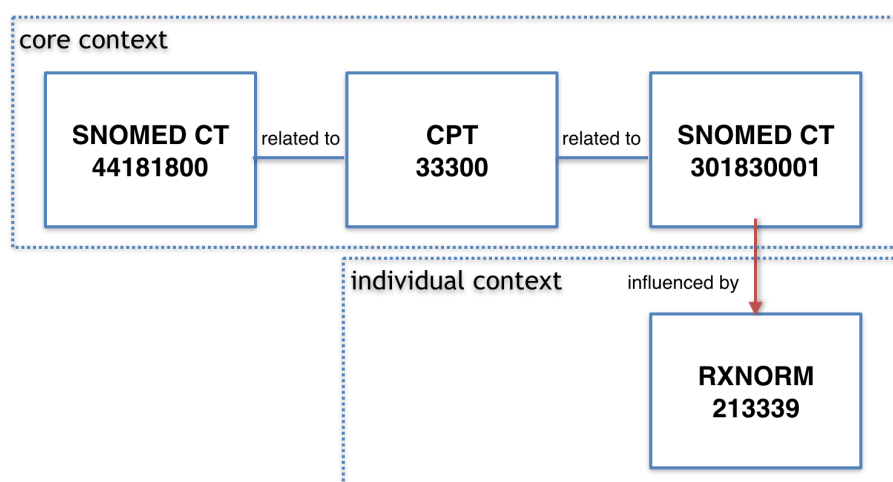


Figure 3.4: *CMIID* scheme concept using John Doe's example with a core context formed by three contexts and one individual context.

So far we have described the hybrid scheme concept and what it supports. Individually, each code holds the characteristics to support the details so far presented. In Figure 3.5 a code representation is shown.

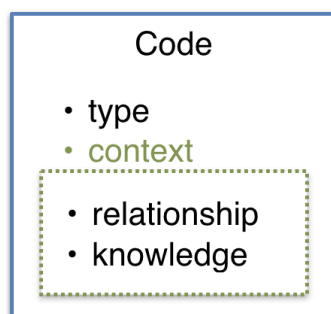


Figure 3.5: The code representation scheme in the taxonomy. It is defined by the coding system (type) and the context, i.e., type of bound to other contexts and the codes supporting that relationship.

A code representation includes:

- **type** - the coding system for the intended context (e.g. **icd10cm\_I10**). Each concept is properly structured with a thesaurus prefix ID followed by the corresponding code. In order to be a prefix universally supported, understood and not imposed by this solution, *CMIID* uses the Root Source Abbreviation (RSAB) of each ontology available in the active release (e.g. US Edition of SNOMED CT - snomedct\_us) (NLM, 2017). This prefix suffers a cleanup process by the framework removing characters such as space and underscore, which could not be correctly parsed during the process;
- **context** - characteristics (**relationship** and **knowledge**) that define it:



- **relationship** - property that identifies the type of context bound (see Table 3.1 and Section 3.3.3 for further explanation);
- **knowledge** - list of codes with which this code has a **relationship** with - useful to index and map a knowledge graph.

Table 3.1: Relationship properties supported between codes in the *CMIID* query scheme.

Property
sibling_of
related_to
defined_by
enforced_by

### 3.3.3 Taxonomy context management

As shown in Figure 3.5, a code context sustains relationships from those listed in Table 3.1. This property can be defined in different ways:

- *sibling\_of* - two codes used within the same context, are siblings. They all help defining in a granular way a better concept classification (e.g. see Figure 3.8: *ctx1* refers to drugs, Ketanserin or Amoxicillin);
- *related\_to* - Two codes from distinct contexts are mutually related. This relationship allows a clinical connection clause between contexts (e.g. researcher seeks to relate a diagnosis context with a procedures one. See Figure 3.8: *ctx1* with *ctx2* and *ctx3*);
- *defined\_by* - a specific code may not exist in the CDR and other codes may have been used instead for several reasons (e.g. upcoding). With this property, the researcher can define other codes that allow the definition he is seeking for, using for example downstream codes or even deprecated codes. It helps creating that way an agnostic interpretation of the version in use. This relationship targets solely the refinement of the results from the core context, without the need to change the main query. For example, John Doe built the main query and runs it periodically against several clusters but in a particular one, there are some irregularities, which he wants to filter. This structure allows him to easily execute that. In summary, it specifies other codes and respective contexts we also want to consider in the search process as additional records to the main core context (e.g. see *ctx4* - code **icd10cm\_I50** - in Figure 3.8: acts as a conditional operator *OR*);
- *enforced\_by* - a code may have been used incorrectly and in order to filter these outliers to avoid influencing the results, a set of codes can be enforced to ensure when that code was used in a specific context, all the other codes are available. This relationship targets solely the refinement of the results from the core context, without the need to change the main

query. For example, John Doe built the main query and runs it periodically against several clusters but in a particular one, there are some irregularities he wants to filter on the results. This structure allows him to easily understand that. In summary, it acts as a mechanism of validation and its usage relies on the knowledge the researcher has over the data or how clinically safe is to say when A happens B and C also happens on each particular situation (e.g. see *ctx4* - code **cpt\_33300** - in Figure 3.8: acts as a conditional operator *AND*).

### 3.3.4 Taxonomy coding and representation

With the initial premise in mind, now we consider the following question representation as a hypothetical one for John Doe's initial subject. The skeleton of the taxonomy would be as shown in Figure 3.6.

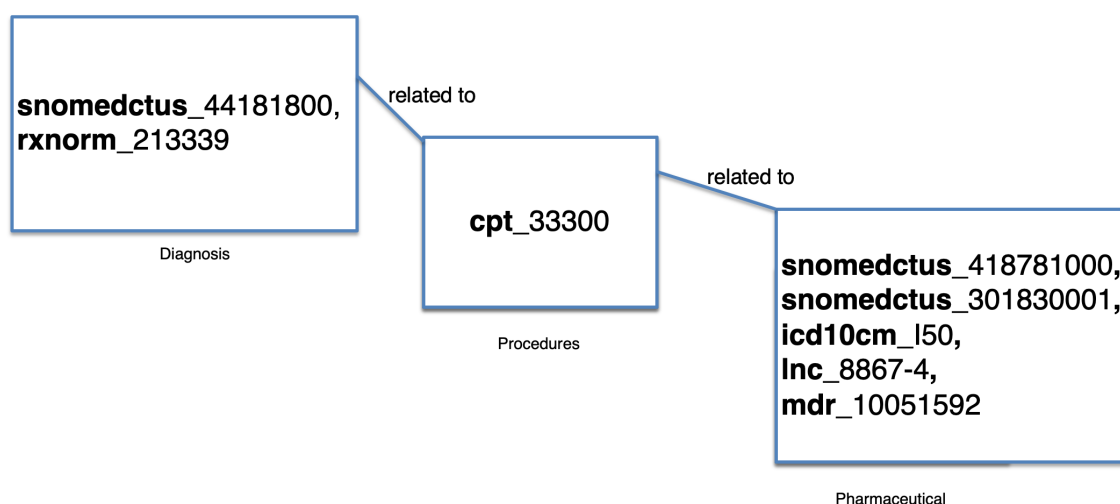


Figure 3.6: Example of a *CMIID* query with a core context formed by Diagnosis, Procedures and Pharmaceutical contexts and containing various thesaurus.

In order to improve the understanding of codes definition and relationship to others, in Figure 3.7 is shown a slice of the diagnosis and procedures contexts where is possible to perceive the definitions, atomically. Each code possesses the type used to code the concept and also a clear description of the context properties to others codes of the taxonomy.

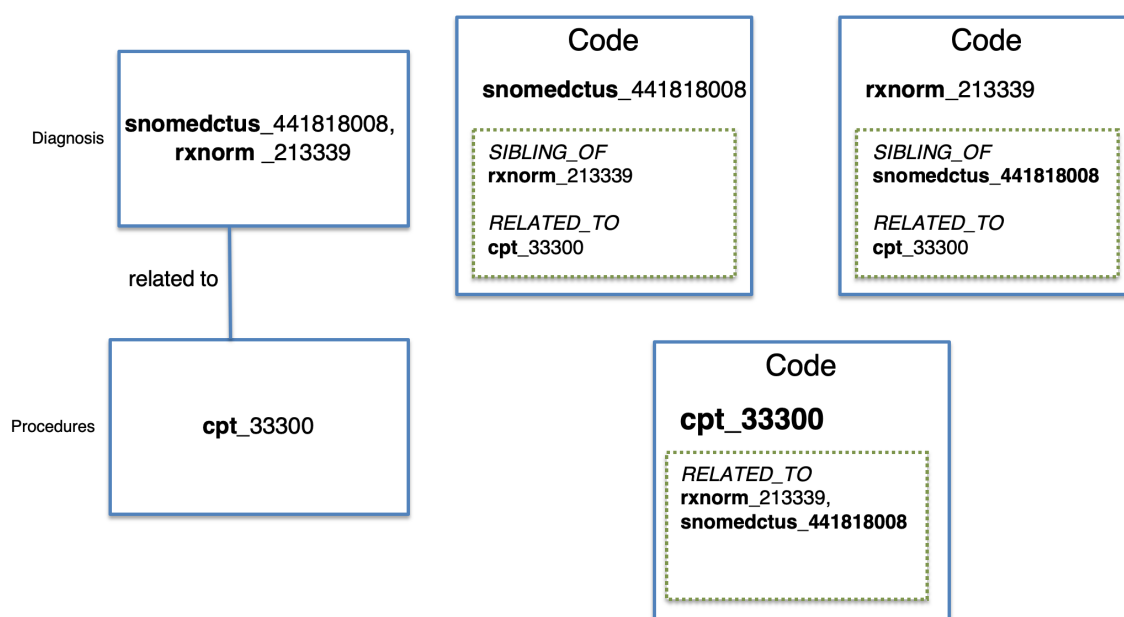


Figure 3.7: Example of the code representation scheme, based on a core context formed by Diagnosis and Procedures contexts and with the relationships between codes.

With the intention of applying advanced knowledge on the query for a more effective search (use of individual contexts stated in Section 3.3.3), a new clinical question can then be performed as show in Figure 3.8. Several contexts can live together and the way those are related matters in the search engine.

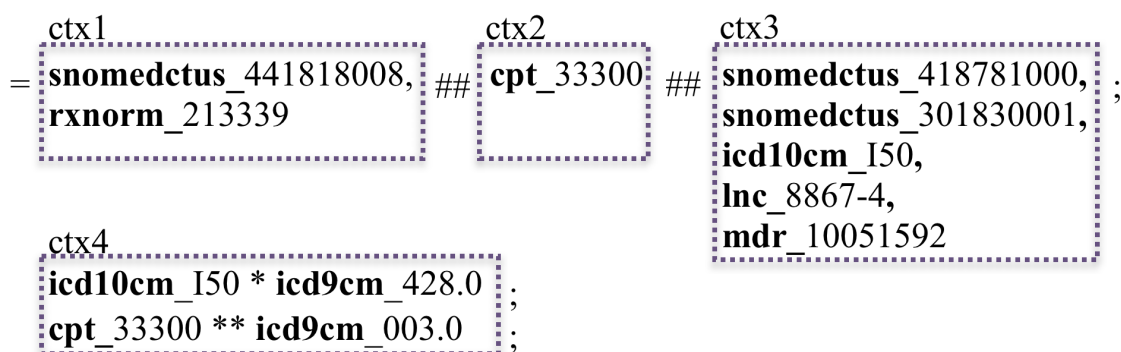


Figure 3.8: Clinical question coded with *CMIIID*, containing a core context about Pharmaceutical, Procedures and Diagnosis (*ctx1*, *ctx2* and *ctx3* respectively). Additionally *ctx4* is an individual context section with ICD10CM and CPT codes having a particular context with ICD9CM codes. This allows an advanced refinement of the results.

To better retain code properties, relationships and manage contexts, Graph Theory is applied on the taxonomy (see Section 3.4 to understand usage) providing layers to validate contexts, and indexing each code for a quicker access.

The information is represented using a connected graph whereas from each code is possible to reach any other code of the query, within the same context or between contexts (Stevanovic,

2014). Using a graph of this kind and adopting the graph database principles, benefits from several advantages (Vicknair et al., 2010; Angles, 2012; Ehrlinger and Wöß, 2016):

- More reliable insights - connecting all codes helps providing a complete and contextually relevant perspective of the state of the query. With the ability of adding more information to each node can help improve the relationships and the search engine by having access to all properties;
- Better performance - each node maintains its neighbors information only, and no global indexes about other node connections are kept. This allows constant performance while data size grows;
- Flexible schema - provides a flexible solution while serving the query. We can add and drop nodes or their attributes to extend or shrink the data model;
- Representation of higher-order relations - useful for modeling data of other areas of knowledge representation (e.g. bioinformatics);
- Enhanced engine to generate new knowledge - contributes with a reasoning engine to generate new knowledge and the possibility to integrate or be integrated in one or more information sources.

Using a graph with these properties allow us to evolve our framework to support more metadata, either from the repositories or from the queries. This metadata can be used to promote better data quality by adding characteristics related to how the data was coded, specificities of the domain, or even enable the researchers to tag inconsistencies (e.g. invalid code).

Figure 3.9 illustrates how the clinical question from John Doe is represented.

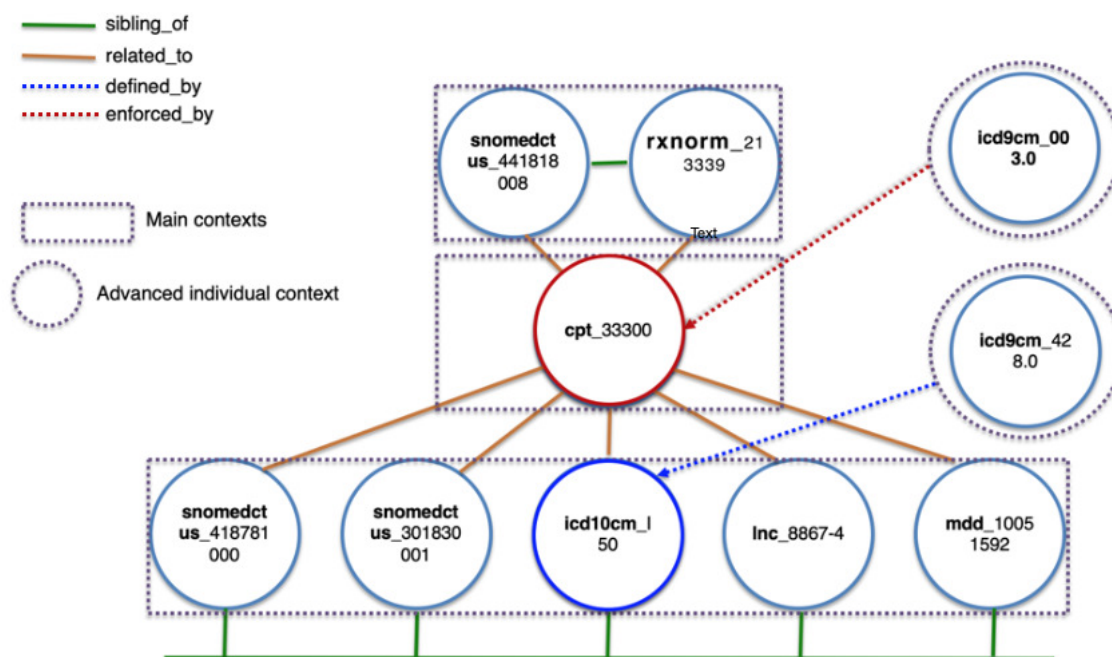


Figure 3.9: Taxonomy query from John Doe case-study, represented using a connected graph. Code relationships are illustrated: both core and individual contexts. For simplicity purposes, not all *related\_to* connection are represented.

Each edge is colored according to the context relationship. The *sibling\_of* and *related\_to* represent a two-way relationship, which means each code has a bound to all the other ones with the same property and vice-versa (see Table 3.2).

John Doe case-study, comprises all possible four context relationships (see Section 3.3.3). Its representation is a single graph containing the main contexts illustrated in Figure 3.9. For simplicity purposes, not all *related\_to* connections are represented - between the top 2 codes (SNOMED CT and RxNorm), and the 5 lower ones.

In the individual contexts, each one of those graphs has a direct edge on an individual code in the main graph. Codes **cpt\_33300** and **icd10cm\_I50** are the target codes because they hold the relationship, i.e., they are defined and enforced by an **icd9cm** code, accordingly. This clearly indicates there is the attachment of specific properties to that code. Having all 4 relationships in-use we get a taxonomy representation that aggregated 3 graphs in total.

With a graph representation several advantages are inherent:

- Easy validation of the context constraints;
- Easy visual representation of the taxonomy and coding standards associated;
- Ground floor to further studies on how coding standards are used across clinical contexts;
- Graph capabilities such as search and edges management;
- Knowledge retrieval and integration into other sources (e.g. federated systems and HIS).

Table 3.2: Code relationships based on knowledge graph from Figure 3.9. All *sibling\_of* and *related\_to* properties represent two-way relationships.

Code	Code	Relationship
icd10cm_I50	snomedctus_418781000	<i>sibling_of</i>
icd10cm_I50	snomedctus_301830001	<i>sibling_of</i>
icd10cm_I50	lnc_8867-4	<i>sibling_of</i>
icd10cm_I50	mdr_10051592	<i>sibling_of</i>
icd9cm_428.0	icd10cm_I50	<i>defined_by</i>
icd9cm_003.0	cpt_33300	<i>enforced_by</i>
snomedctus_441818008	rxnorm_213339	<i>sibling_of</i>
snomedctus_441818008	cpt_33300	<i>related_to</i>
rxnorm_213339	cpt_33300	<i>related_to</i>
snomedctus_441818008	snomedctus_418781000	<i>related_to</i>
snomedctus_441818008	snomedctus_301830001	<i>related_to</i>
snomedctus_441818008	icd10cm_I50	<i>related_to</i>
snomedctus_441818008	lnc_8867-4	<i>related_to</i>
snomedctus_441818008	mdd_10051592	<i>related_to</i>
rxnorm_213339	snomedctus_418781000	<i>related_to</i>
rxnorm_213339	snomedctus_301830001	<i>related_to</i>
rxnorm_213339	icd10cm_I50	<i>related_to</i>
rxnorm_213339	lnc_8867-4	<i>related_to</i>
rxnorm_213339	mdd_10051592	<i>related_to</i>

### 3.4 Comprehensive medical information identifier framework

We introduced and explained *CMIID*: scheme definition, codification of clinical questions, context management and its scalability.

Even though this scheme introduces a revolutionary way to formulate questions, a system that is capable of understanding and processing the scheme is still required.

Existing information and research systems characteristics challenges the integration of the scheme in an effective way. Therefore, we developed a framework that acts together with the scheme and then can ease that integration, without conceding functionalities. The developed framework helped us exploit *CMIID* acceptance (in terms of integration, scalability, maintainability and complexity) as well the search harmonization performance.

This way an expert could formulate an advanced question using the respective codes from any source without having to build a query adjusted to the repository or to the search tool. Additionally, any new source just requires a minimum of information to make it completely available in the framework, without the need to (re-)build mapping layers or other solutions.

The framework, developed in Python, was designed to be easily integrated in healthcare systems and requires the registration of the data sources and its characteristics. Additionally, it uses the UMLS core services to ensure in an automatic way a continuous up to date knowledge base (with a wide range of vocabularies of different domains) imposing a strong integrity validation on the query understanding and execution.

### 3.4.1 Architecture overview

*CMIID* framework comprises five main modules (see Figure 3.10):

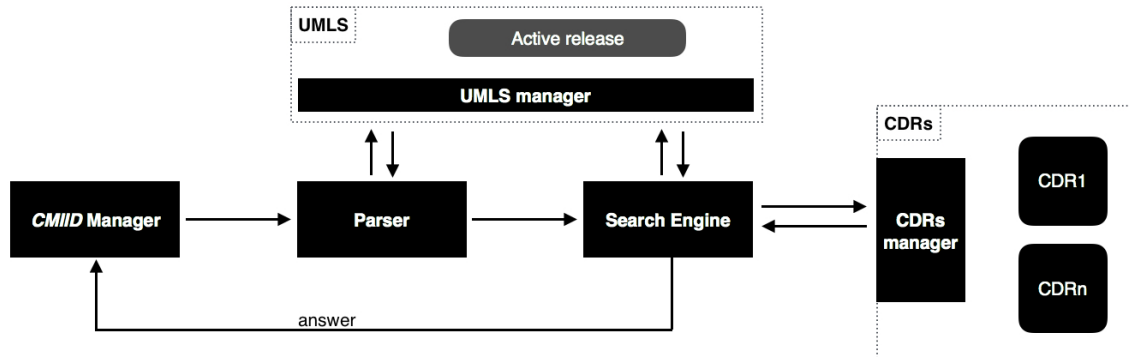


Figure 3.10: High-level structure of the *CMIID* framework architecture, using the UMLS services for contexts validation. Search engine is responsible for segmenting queries based on the repositories characteristics and aggregating the information afterwards.

- **Manager** - Responsible for all incoming connections and requests. This entrypoint ensures we can have security and request validations in-place;
- **Parser** - Validates syntactically, semantically and clinically (if codes exist) the requested query and builds a knowledge graph for an optimized manipulation in the search engine;
- **Search engine** - Given registered CDRs and theirs characteristics, builds dynamic and optimized queries depending on the target data source and on the contexts and relationships retrieved from the query, properly identified and validated using UMLS manager;
- **UMLS - Manager** is used to identify and validate codes taxonomy and context, transposing that info into the query builder mechanism inside the search engine. It uses the active release to get continuous official updates and to be simultaneously compliant with numerous ontologies - no need for an intervention on the framework side;
- **CDRs** - Pool of data sources registered as targets - can be databases, web services or others. Manager is responsible for the registration of each *CDR* along with managing the query execution in each data source;

### 3.4.2 Query validation

As displayed in Figure 3.11, the knowledge graph is a representation of the codes taking into consideration the syntax mentioned in Section 3.3.4. Using Graph Theory we have a faster way to index and search contexts, extract relationships, employ strict validations and set properties for the query builder process. Such allows constituting a base line for further data mining developments (Riaz and Ali, 2011; Khan et al., 2016; Balamurugan and Zubar, 2018; Berge, 2001; Gross and Yellen, 2004).

Example of a clinical query:

```
=icd9cm_003.0,icd9cm_003.1##icd9cm_9910,cpt_99291
##rxnorm_866513,rxnorm_0452;
icd9cm_003.0*icd9cm_96.6;
cpt_99291**icd9cm_43.7;
```

Legend:

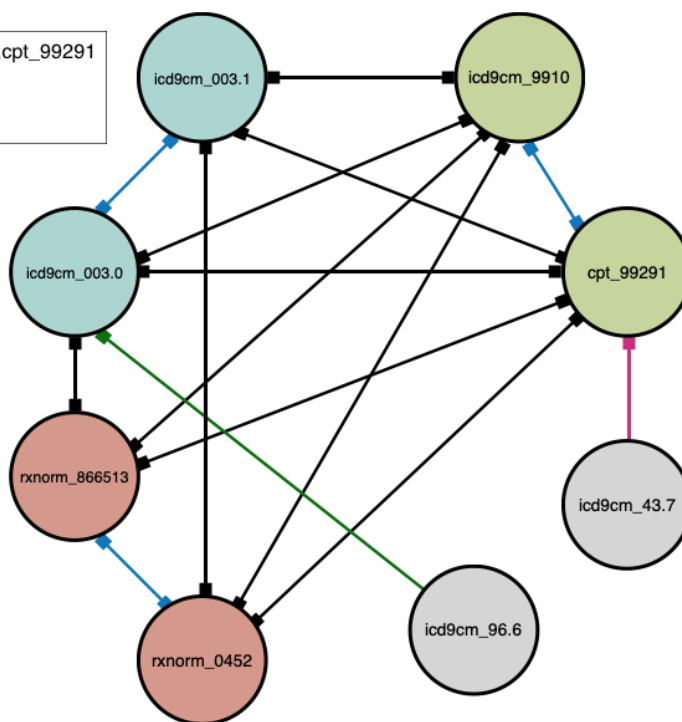
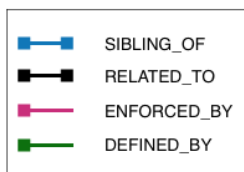


Figure 3.11: Knowledge graph representation of an example of a *CMIID* clinical query, using a circular layout and with relationship properties illustrated using colored edges.

Furthermore, contexts that share a "Relation" relationship, have codes mutually related. This bidirectional affiliation improves graph search efficiency and it is one important property in the search harmonization process, either because of the clinical significance it has on the results and also the efficiency on the query builder mechanism.

Using a graph also enables the framework to provide advanced analytical analysis between different queries. Having an optimized knowledge representation and a history of usage is possible to further develop studies: contexts most queried, associations between codes (e.g. siblings), contexts relationships, etc.

Apart from being able to manage this history, the *CMIID* Manager permits to create sessions based on requirements (by department, research lab, user, etc.) where this system is going to be used. Withal, we can retain more information of each code at anytime, from the query input or by the *CMIID* Manager. This information can be helpful to enrich the search engine or to develop data quality measures.

### 3.4.3 Search engine

The search engine module (see Figure 3.12) is responsible for building dynamic queries based on the relationships defined in the knowledge graph. These queries are built accordingly to each registered CDR and its main characteristics (e.g. available contexts).



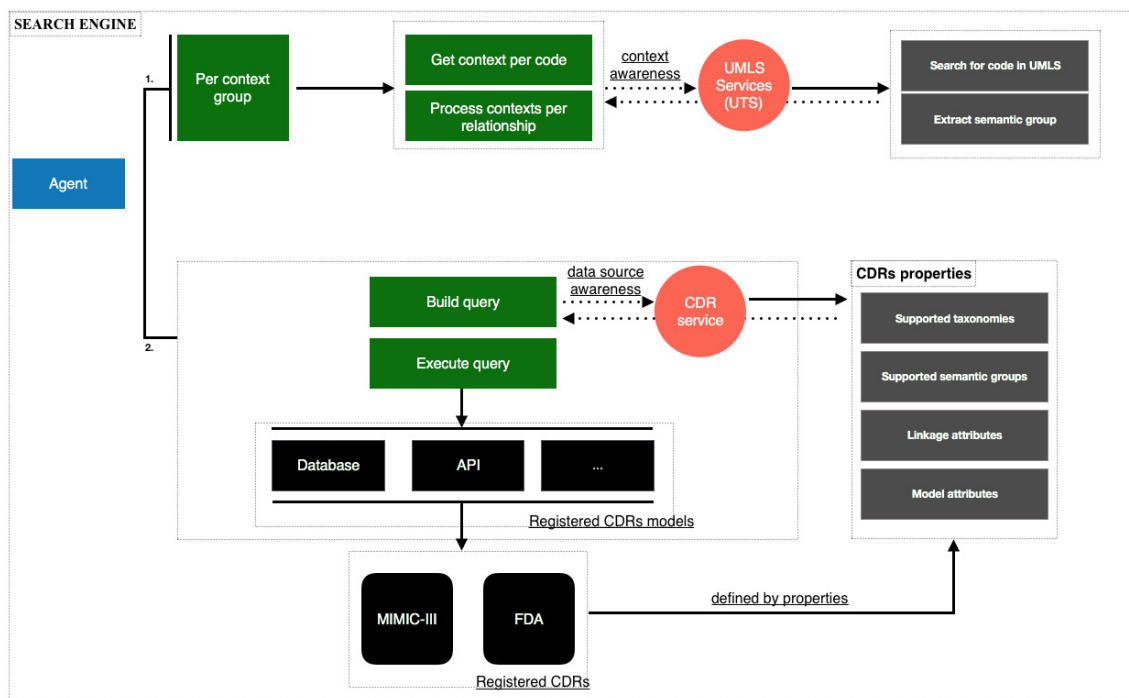


Figure 3.12: Search engine of *CMII* framework using a CDR service to extract properties and identify matching sources. The UMLS services are used to retrieve the clinical contexts per each code.

This engine has two responsibilities: 1) extracting and processing contexts per relationship; 2) build and execution of queries using the registered CDRs properties.

Apart from the available contexts, other characteristics such as linkage and model attributes are also important to properly build the query, specially for database sources (see Table 3.3). This behaves as a mapping source configuration whereas it is expected an identification of the tables that have codes, its corresponding semantic group, table attribute which represents the code, and other properties useful for connecting to the sources.

As mentioned, the property **subject\_key** is an intra-database linkage attribute that identifies the relationship key between tables - this is most of the times a patient identifier. Having a value set, means the resource is a relational database (e.g. MySQL and Postgres), and consequently that field is used in JOIN operations. When inexistent a NoSQL database resource is considered.

Table 3.3: Properties required in the registration of a *CDR*, assuming execution is inside a DSH.

CDR	Property	Description	Details
API	type	Type of CDR: API or Database	
	url	Target base url	
	token	Authorization token	
	taxonomies	Supported taxonomies (from UMLS active release or proprietary)	
	contexts	UMLS semantic groups the CDR holds	
Database	type	Type of CDR: API or Database	
	taxonomies	Supported taxonomies (from UMLS active release or proprietary)	
	contexts	UMLS semantic groups the CDR holds. Specify per each one, the table name and column to identify that has the codes.	[[{"table": "procedures_icd", "id": "icd9_code"}, {"table": "cptevents", "id": "cpt_cd"}]] or [{"table": "d_labitems", "id": "loinc_code", "table_fk_id": "itemid", "fk_table": "labevents", "fk_id": "itemid"}]]
	subject_key	Supports 2-level tables referencing using foreign keys. Intra-database linkage attribute that represents the subject	
	db_type	If Postgres, MySQL or others	
	configs	Host, user, password, database name, schema	

These properties, apart from being compliant with the DSH paradigm through the consideration of security and reachability of all sources, require additional configurations: identification of the type of the source (meaningful for building advanced queries) and the contexts (UMLS semantic groups), defined in there.

These are the minimum requirements to register a new CDR into the framework and to be automatically detected in a new search. The work developed had in consideration the investigation from Section 3.1 along with the analysis of the sources used on this thesis, mentioned in Section 3.2. By connecting more distinct sources we expect to enhance the registration module and how it is used with the Search engine.

#### **3.4.4 Context validation**

To properly identify contexts, the UMLS services are used to extract the categories. The UMLS semantic network provides a consistent categorization of all concepts represented in the UMLS Metathesaurus and consists of a set of broad subject categories called Semantic Types (NLM, 2018; McCray et al., 2001). Since this is applied to all concepts, there are more than 130 categories and such specificity would not bring a solid and clear leverage because of two key reasons:

1. Queries would be more complex, with an increase number of contexts. Knowledge graph would be more complex as well the query building module.
2. The CDRs may not have such detailed contexts translated into the domain and therefore significant query relationships would be loss.

With this in mind a high-level categorization approach was used. Within the same scope, UMLS provides a smaller and coarser-grained set of semantic type groupings that fulfill these requirements. These contexts are called semantic groups and can be (NLM, 2018):

- Chemicals & Drugs;
- Procedures;
- Activities & Behaviors;
- Anatomy;
- Concepts & Ideas;
- Devices;
- Disorders;
- Genes & Molecular Sequences;
- Geographic Areas;

- Living Beings;
- Objects;
- Occupations;
- Organizations;
- Phenomena;
- Physiology.

This approach provides a way to a) validate if a set of codes belong to the same context (a code may have more than one context association) and b) validate the relationships to the remaining groups: this is important in the query builder. In this step, the CDRs manager service uses registration information to assess which data sources are compliant with the semantic groups identified in the knowledge graph.

According to the UMLS strategy, more semantic groups are added if it brings more definition and allows deeper relationship within the ontology (McCray, 1989). For the definition and development of this work, these types are enough.

As shown previously, the framework supports linkage between records (intra-CDR) either by using a patient reference or other reliable linkage property implemented in the repository schema. Harmonization inter-CDRs is onerous and it may be impossible if we can't identify common subjects across sources. Knowing this limitation and the work that has been developed in the literature (see Section 2.5), we implemented a harmonization technique living on the definition of the clinical contexts.

In the next section, we will introduce and explain the practical terms of this implementation.

### 3.5 Search harmonization technique

In Section 3.4, we introduced the architecture of the *CMIID* framework, explaining how the query builder process is done having as input a clinical question coded with the *CMIID* syntax. As shown in Figure 3.12, the search engine validates the contexts at first - builds a knowledge graph using the UMLS services. Secondly, queries are built, executed and harmonized.

A correct registration of the repositories is fundamental to ensure the well-functioning of the framework, either in terms of interpreting the *CMIID* question into multiple queries and the execution itself.

We will be also presenting the search technique, explaining how a clinical query is translated into technical repository queries retrieving all related information while explaining the harmonization of search results by clinical context.

### 3.5.1 Repositories registration

As explained in Section 3.4, the search harmonization solution is focused on sources such as databases and web services requiring the configuration of certain fields. Some of those fields enumerate the supported contexts and the respective objects (database tables or service objects) that hold the coded information.

In the Listing 1 and in the Listing 2 is presented a snippet containing the contexts configuration for a web service and a database repository, respectively.

```
1  {  
2      "CONTEXTS": {  
3          "drugs":  
4              [  
5                  {  
6                      "resource": "patient.drug.openfda",  
7                      "id": "rxcul"  
8                  }  
9              ]  
10     }  
11 }
```

Listing 1: Example of contexts configuration in a web service repository. Resource identification and main identifier for the coding attribute are necessary.

For both of the scenarios, the context definition is mandatory and shares a common structure. In the database snippet, **Disorders** and **Procedures** represent the 2 contexts available and for each one of those there is more than 1 resource identifying the table and attribute name which retains the clinical coding. Depending on the nature of the source, it may exist more than 1 object for the same context.

```

1  {
2      "CONTEXTS": {
3          "disorders":
4              [
5                  {
6                      "resource": "diagnoses_icd",
7                      "id": "icd9_code"
8                  },
9                  {
10                     "resource": "drgcodes",
11                     "id": "drg_code"
12                 }
13             ],
14         "procedures": [
15             {
16                 "resource": "procedures_icd",
17                 "id": "icd9_code"
18             },
19             {
20                 "resource": "cptevents",
21                 "id": "cpt_cd"
22             },
23             {
24                 "resource": "drgcodes",
25                 "id": "drg_code"
26             }
27         ]
28     }
29 }

```

Listing 2: Example of contexts configuration in a database repository. Identifying the tables and the main identifier for the coding attribute is necessary.

The configurations for both repositories are similar differing in some annotations specific to the source type and the authentication method. On both, "SUPPORTED\_TAXONOMIES" and "CONTEXTS" refer to the identification of the taxonomies and contexts there represented. The accepted syntax is based on the UMLS active release definition (as explained in Section 3.3.1:

1. Supported taxonomies - RSAB available in the U.S. National Library of Medicine (NLM, 2017). A lowercase version without special characters is required;
2. Contexts - semantic groups as explained in Section 3.4.4. A lowercase version is required.

Additional mandatory fields ensure the correct source registration. Source type (identified by the field "TYPE"), allows to set the source as "webservice" or "database". Field "PATIENT\_KEY",

refers to the internal attribute that links a subject across all tables. If there is none, must be left empty.

When attempting to register a database server that hosts multiple databases, each one of those needs to be configured as a separate source in the configuration file.

Supporting distinct database management systems is meaningful. Specifying the database type enables the *CMIID* search engine to flex the execution accordingly, using the correct driver - the value can be set as "mysql" or "postgres".

```

1  {
2      {
3          "<<SOURCE_1>>": {
4              "TYPE": "<<SOURCE_TYPE>>",
5              "API_KEY": "<<WEB_SERVICE_TOKEN>>",
6              "BASE_URL": "<<BASE_URL>>",
7              "SUPPORTED_TAXONOMIES": [ "<<tax1>>", "<<tax2>>", "..."],
8              "CONTEXTS": {
9                  "<<umls_ctx1>>": [
10                     {
11                         "endpoint": "<<endpoint_subroute>>",
12                         "id": "<<attribute>>",
13                         "keepNotation": false
14                     },
15                     {
16                         "endpoint": "<<endpoint_subroute>>",
17                         "id": "<<attribute>>",
18                         "keepNotation": false
19                     }
20                 ],
21                 "<<umls_ctx2>>": [
22                     {
23                         "endpoint": "<<endpoint_subroute>>",
24                         "id": "<<attribute>>",
25                         "keepNotation": false
26                     },
27                     {
28                         "endpoint": "<<endpoint_subroute>>",
29                         "id": "<<attribute>>",
30                         "keepNotation": false
31                     }
32                 ]
33             },
34         },
35         "<<SOURCE_2>>": {
36             "TYPE": "<<SOURCE_TYPE>>",
37             "SUPPORTED_TAXONOMIES": [ "<<tax1>>", "<<tax2>>", "..."],
38             "CONTEXTS": {
39                 "<<umls_ctx1>>": [
40                     {
41                         "table": "<<table_name>>",
42                         "id": "<<table_attribute>>",
43                         "keepNotation": false
44                     }

```

```

45         {
46             "table": "<<table_name>>",
47             "id": "<<table_attribute>>",
48             "keepNotation": false
49         }
50     ],
51     "<<umls_ctx2>>": [
52         {
53             "table": "<<table_name>>",
54             "id": "<<table_attribute>>",
55             "keepNotation": false
56         },
57         {
58             "table": "<<table_name>>",
59             "id": "<<table_attribute>>",
60             "keepNotation": false
61         }
62     ]
63 },
64 "PATIENT_KEY": "<<internal_>>",
65 "DATABASE": {
66     "CONFIGS": {
67         "host": "<<DB_HOST>>",
68         "dbname": "<<DB_NAME>>",
69         "user": "<<DB_USER>>",
70         "password": "<<DB_PASSWORD>>"
71     },
72     "SCHEMA": "<<DB_SCHEMA_NAME>>",
73     "TYPE": "TYPE_OF_DB"
74 }
75 }
76 }
77 }

```

Listing 3: Sample configuration file for data sources registration. Required fields are identified within the placeholders "<<".

In addition to the previous properties, it can be provided the default schema to be used, in the field "SCHEMA" - if left empty it will use the one set in the database server.

### 3.5.2 Query builder

As mentioned in Section 3.4, the search engine has two responsibilities: 1) extract and processing contexts per relationship; 2) build and execution of queries using the registered CDRs properties.

In each context group, the UMLS services are used to extract the matching contexts. Thereon, siblings represented in each group may be identified with more than one context. This fact, can act as a confirmation for the researcher because either a code may have been wrongly used or it has more than one possible context and for that reason, the query may need a refinement.

Subsequently, Boolean algebra laws are applied (Schmidt and Ströhlein, 2012):

1. Distributive Law - used to map identified contexts to all of the codes within the same group (relationship: siblings). Result is a set of *OR* clauses that target the resource mentioned in the Listings in Section 3.5.1;
2. Associative Law - used to facilitate and manage the overall query once Distributive Law is applied to all groups (relationship: relation, definition and enforcement). Result is a set of *AND* clauses that target the resource mentioned in the Listings in Section 3.5.1.

Results from these algebra operations are correctly adjusted to the CDR type and characteristics. The representation of *AND* and *OR* clauses depends on the nature of the source specially if it is not a database - it involves a more complex logic formulating the query to the correct syntax. In Figure 3.13, is presented an example of how a given query is translated into a repository query (database in this case) having the previous considerations present.

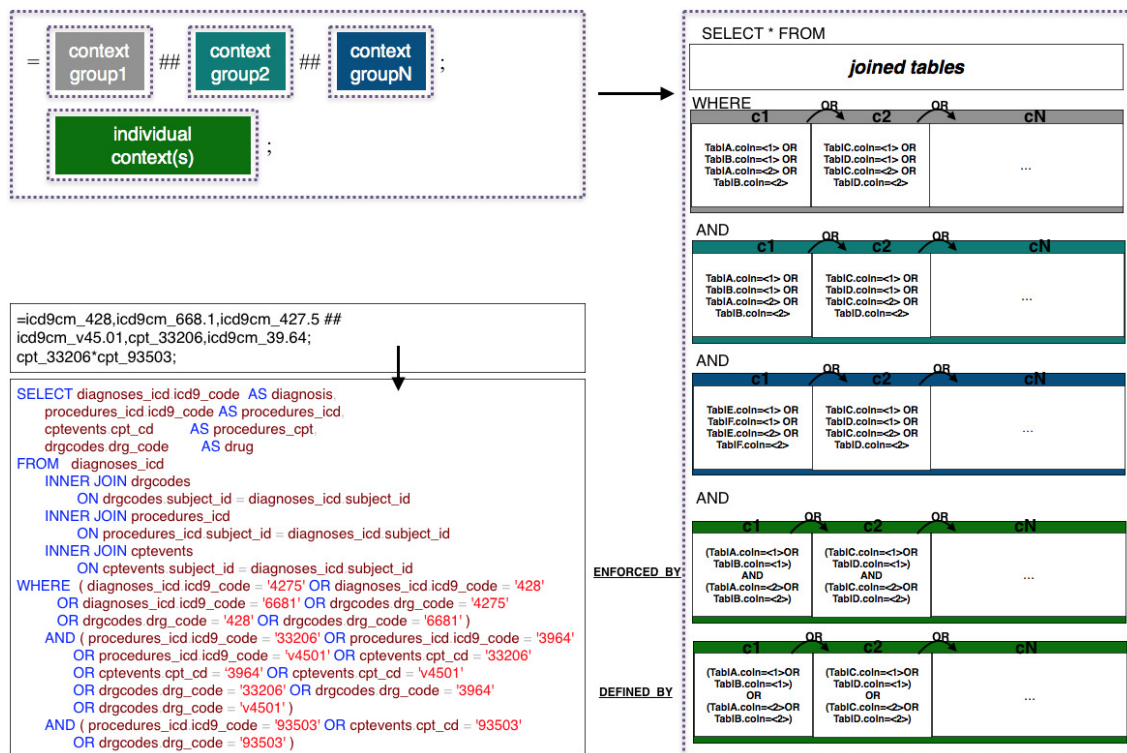


Figure 3.13: Example of a *MIMIC-III* SQL query coded from a *CMIID* scenario. On the right is shown the template containing the rules for the translation.

Joint tables are retrieved from the query contexts and corresponding CDR settings. The **WHERE** clauses depend on the correct interpretation of the contexts and their relationships. In the same way, if multiple tables exist for the same context, they are represented using a *OR* clause for each code of that context. If more than one context is identified for each group a *OR* clause is used (*c1, c2...cN*).

In this example, there are two context groups: first (*icd9cm\_428,icd9cm\_668.1,icd9cm\_427.5*) and second (*icd9cm\_v45.01,cpt\_33206,icd9cm\_39.64*). They are about Disorders and Procedures



(respectively) and each context has multiple resources - hereby Distributive Law is applicable to form all possible conditions.

On a multi-context consideration, the same operations are done for each individual context. In case of uncertainty, the algorithm does not pick a single context for the group of codes and resultantly all scenarios are compiled and left to consideration on the researcher side to improve the query.

Definition and Enforcement relationships are easier to explain once the core context logic (all group contexts) is introduced. They target helping the researcher to apply an advanced level of filtering inasmuch as more codes can be provided as "optional" (behaving like a SQL column filter using *OR* operator) or mandatory (behaving like a SQL column filter using *AND* operator). Both are directly attached to an individual code which helps expanding the clinical knowledge and search impact.

Converting this logic to other repositories is a demanding task. For example, in web services complex queries can be built using URL and/or query parameters with the possibility to impose mentioned operators in one single execution or with a composition of requests (see Figure 3.14).

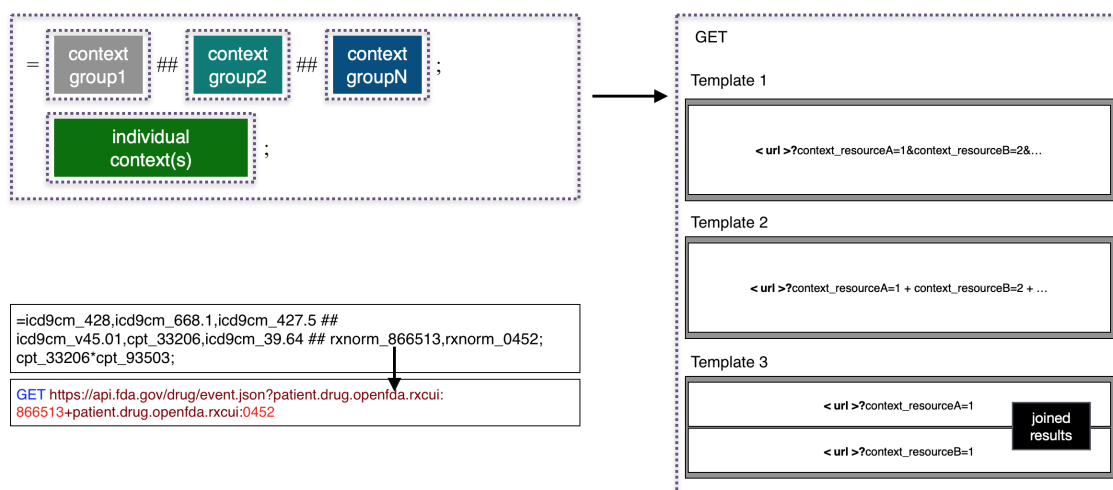


Figure 3.14: Example of a web service query coded from a *CMIIID* scenario. On the right is shown the templates containing the rules for the translation.

Distinct web services templates may be used to query contexts and corresponding codes. This variety depends on how services are setup and how they handle information request. Depending on such conditions, different templates may be used to apply *OR* and *AND* operators. As shown in Figure 3.14, for that specific example the *OR* operator was supported by the service via a plus sign within the same query parameter - could be used to search for multiple context groups. The "Relation" relationships or additional contexts within the same group would require to perform more than one request, joining results and then filtering.

### 3.5.3 Query execution

Using a context-based medical information identifier to execute queries, is a new value proposition which comes with high demands and complexity. At first it is important to understand which is the question we are seeking an answer to. We can start by describing it using natural language text as for instance:

Are there any patients diagnosed with heart attack symptoms that have been administrated with antibiotics during a pacemaker surgery?

Having identified our question we can then translate it into the CMIID syntax, building the contexts that try to satisfy our needs. As shown in Figure 3.15, satisfying means finding all records whose contexts relate to each others on specific logical conditions.

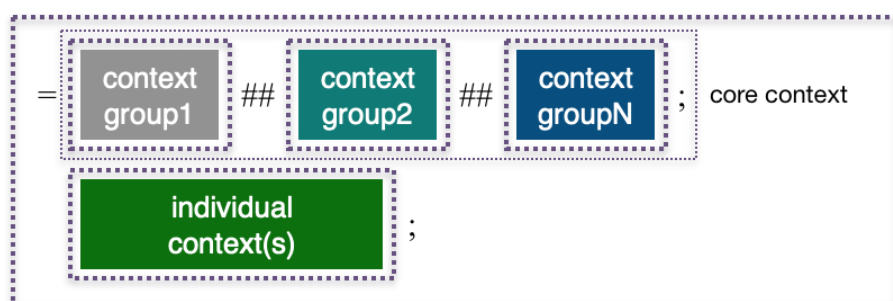


Figure 3.15: High-level scheme of a *CMIID* query, with a multi-contexts perspective. The scheme supports multiple contexts as the core relation and also diverse individual contexts.

The queries resulting from the Search engine interpretation are then executed against the sources. By restricting these queries into a harmonization topic two concerns are relevant to be highlighted:

1. how to link records from repositories that don't share common identifiers (previously mentioned *subject\_id* or "PATIENT\_KEY")?;
2. how to link search results of a multi-context query from repositories that don't share the same contexts (for example a database and a web service)?

The *CMIID* proposal has clinical contexts as its core and the use of subject identifiers is only considered within each repository whenever is possible (to improve the quality of the results). In an inter-repositories point of view, contexts are the linkage identifiers because distinct CDRs may not have properties (in common or not) to identify the same *subject\_id*. The volatility of repositories characteristics can affect significantly the harmonization performance if no common identifier is considered - in a DSH perspective this could be much likely a problem.

On the opposite of other linkage solutions that take a subject identifier to build the virtualization layers, this approach only requires it intra-CDR (when existing: e.g. databases) and uses a

context-based approach. However, CDRs may not have exactly the same contexts and therefore harmonization is not straightforward (see Figure 3.16).

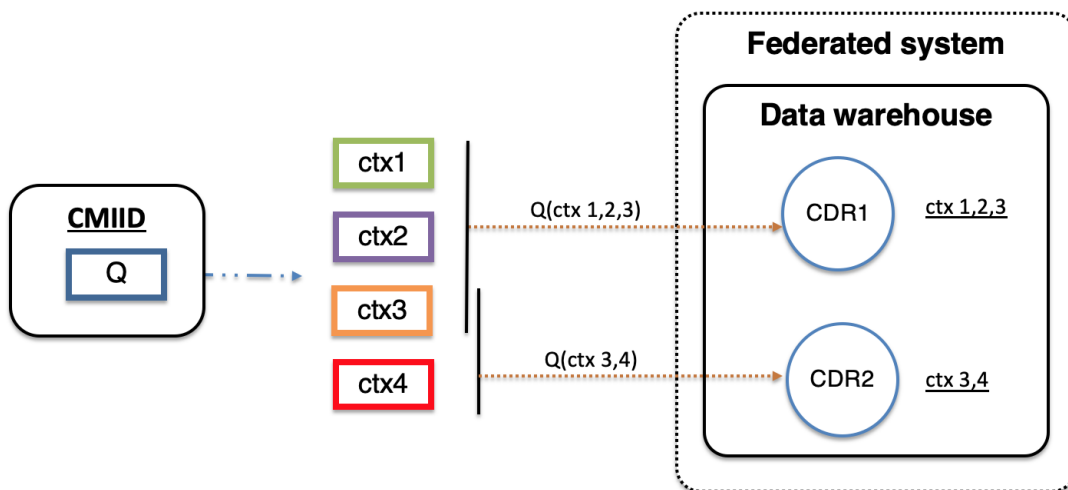


Figure 3.16: Example of a journey associated to the execution of a multi-context query in CDRs, with partial context match. Each CDR query is adjusted according to the matching contexts.

The search harmonization technique we propose, links records based on the context(s) shared with the CDR, and the relationship defined between them. Considering the example shown in Figure 3.16, from a query with 4 contexts, two queries are executed: one for CDR1 containing context 1, 2 and 3 maintaining the relationship between them and another query for CDR2 using the same approach - in this case only context 3 exists in both repositories.

With this approach, results from all the subsequent queries are then joined into one response. When there are shared contexts, the framework defines a cursor in the results to bridge the same context from different sub-queries. Further processing and analysis may be needed, by the researcher, so it can properly study matching criteria in situations like this, with common contexts between repositories.

### 3.6 Framework integration in a (cross-)research setting

Integrating the framework in a research environment can be accomplished by setting the following configurations:

1. Downloading the *CMIID* package;
2. Creating a **cmiid\_ums.json** file with the UMLS account service credentials, as shown in Listing 4;
3. Creating a **cmiid\_sources.json** file with the sources required configurations, as shown in Listing 3. Each intended source must have a single entry in the JSON file.

Configurations files must be placed in the same location as the binary and follow the structured hereby presented.

```

1  {
2      "UTS_SERVICE": {
3          "BASE_URL": "https://utslogin.nlm.nih.gov",
4          "API_KEY": "<<TOKEN>>",
5          "SERVICE": "http://umlsks.nlm.nih.gov",
6          "USERNAME": "<<USER_CREDENTIALS_USER>>",
7          "PASSWORD": "<<USER_CREDENTIALS_PASSWORD>>"
8      },
9      "UTS_CONTENT": {
10         "BASE_URL": "https://uts-ws.nlm.nih.gov"
11     },
12     "UMLS_RSABS_URL": "https://www.nlm.nih.gov/research/umls/
13     knowledge_sources/metathesaurus/release/active_release.html#"
14 }

```

Listing 4: Sample configuration file for the UMLS connection. Required fields are identified within the placeholders "<<>>".

A valid account in UMLS Terminology Services is needed. Otherwise it is necessary to request a UMLS Metathesaurus License and create a UTS account, whose credentials and API token need to be inserted in the configuration file.

Running *CMIID* can be accomplished by executing the following:

```

cd /path/to/cmiid-binary
./cmiid «query»

```

The command argument "query" must be in the same format as illustrated in Figure 3.11.

### 3.7 Summary

In this chapter we have introduced and explained *CMIID*: scheme, framework and search harmonization technique. We have also described the methodology we used to develop the solution and what techniques we applied in order to overcome some literature issues stated in Chapter 2.

## Chapter 4

# Discussion

In the last chapter, we addressed the *CMIID* hybrid scheme, the framework and the search harmonization technique. Our solution uses clinical contexts in the formulation of research queries, making use of the domain knowledge available in the UMLS: conferring validation and trustworthiness layers and a reliable world-class ontology.

In this chapter we discuss the results on four main topics:

1. Definition of the hybrid coding scheme;
2. Framework architecture (integration, complexity, maintainability and scalability);
3. Framework performance;
4. Solution positioning in comparison with an existing solution.

An evaluation of the framework is also presented in this chapter. As 1) and 2) are concerned, a Focus Group (with 15 participants) was assembled with researchers and physicians (from different domains) from the Portuguese health technology and research center mentioned in Chapter 3 (Rabiee, 2004; Morgan, 1997). In this session it was presented the solution and its motivation, explaining the architecture, the formulation of queries and the expected results, using the repositories from this thesis as an example. Several questions were then asked:

1. What do you think about formulating queries using clinical contexts and codes?
2. Do you see any disadvantage of using UMLS as the main ontology?
3. Can you identify other relationships that are not represented, and that you use in your professional routine?
4. What do you think of the full disclosure of the registered repositories?
5. What known problems, from your field of expertise, would be solved using this solution?
6. Would you use this solution in your daily routine? If not, what are the main reasons?

Additionally, a System Usability Scale (SUS) questionnaire (Brooke et al., 1996; Bangor et al., 2008; Martins et al., 2015; Lewis and Sauro, 2009) was performed to another distinct and diverse group (10 participants) in order to evaluate the usability of the hybrid coding scheme in clinical practice. This group of people contained: researchers from the beforementioned site, physicians from various Portuguese hospitals and national experts in clinical practice focused on R&D. As 3) is concerned, the framework performance was evaluated by measuring the variation on the number of siblings, contexts and data sources in order to understand the system response time. On 4), we conducted an evaluative comparison with the reference solution from literature - SAIL databank (Ford et al., 2009).

For simplicity, the Focus Group participants will be referred to as evaluators. All the results are explored in each one of the following sections, related to the topics mentioned before.

## 4.1 Clinical queries

To conduct the evaluation of the framework, 7 base cases of clinical medicine were used from Kumar and Clark (2012) and Baliga (2012) - characterized in Table 4.1. These ones took into consideration day-to-day situations in clinical medicine such as: patients that have been diagnosed with cardiac insufficiency and placed a pacemaker; patients with renal insufficiency and that did hemodialysis; patients diagnosed with Alzheimer's and medicated with Galantamine 8mg extended-release capsules, among several others.

Table 4.1: Example of clinical questions translated into *CMIID* queries using UMLS Root Source Abbreviations as code prefixes.

Description	Query
Subjects that took Ketanserin or Ibuprofen 20 MG/ML Oral Suspension	=snomedctus_441818008,rxnorm_544393;
Subjects diagnosed with Salmonella gastroenteritis and septicemia	=icd9cm_003.0,icd9cm_003.1;
Subjects having Salmonella gastroenteritis and septicemia, being diagnosed with Frostbite of face with tissue necrosis and considered a critically injured patient	=icd9cm_003.0,icd9cm_003.1##icd9cm_991.0,cpt_99291;
Subjects having Salmonella gastroenteritis and septicemia, being diagnosed with Frostbite of face with tissue necrosis and considered a critically injured patient, being medicated with Metoprolol Tartrate 100 MG	=icd9cm_003.0,icd9cm_003.1##icd9cm_991.0,cpt_99291#rxnorm_866513;
Subjects diagnosed with Diabetes and being under surgery for cataract extraction with possible insertion/removal of intraocular lens prosthesis	=icd9cm_250##icd9cm_v45.61,cpt_66982,cpt_66984,cpt_66840,cpt_66850,cpt_66852,cpt_66920;
Subjects with symptoms of malnutrition, nausea and anemia and that were under ventilation assist and management and transfusion	=snomedctus_2492009,snomedctus_276608005,snomedctus_129845004,snomedctus_422587007,icd9cm_776.5,icd9cm_787.02##snomedctus_266700009,cpt_94003;cpt_94003*icd9cm_99.04;
Subjects diagnosed with heart failure and cardiac arrest, prescribed with Ketanserin and Amoxicillin	=icd9cm_787.01,icd9cm_787.02,icd9cm_787.0,icd9cm_787.91,icd9cm_009.3,icd9cm_338,icd9cm_789.0,icd9cm_783.0,icd9cm_263.0,icd9cm_263.1,icd9cm_262.0,icd9cm_263.8,icd9cm_263.9,icd9cm_285.8,icd9cm_285.9##cpt_94003;cpt_94003*icd9cm_99.04;icd9cm_263.8*icd9cm_99.04;

On top of this selection, 12 query variations were built in total, to grant us with more precise measurements - see Appendix A.1. Each one of the scenarios was manually coded using the ontologies (guided by codes description) available in the UMLS active release. Individual contexts (see Section 3.3.2) were also introduced to filter for specific clinical considerations of each scenario - for example drugs that may have been administered under rare circumstances.

## 4.2 Hybrid coding scheme

As seen in Figure 3.13 and Figure 3.14 (from Chapter 3), we were able to build an automated query builder process for both of the sources before mentioned. In order to effectively access the resources that contain the data and use them when building the query, a minimum of meta-information from the repositories is mandatory, as we highlighted in Section 3.6.

As opposite to virtualization solutions, that define several mappings to the repositories attributes in order for an abstraction to be provided between search and storage, *CMIID* considers solely the contexts information. That means, those that are available and the resources that exist per context - see Listing 2. This works seamlessly if repositories have a simple representation of information but for complex structures with elaborated records relationships, it requires the capability to support knowledge linkage within the same repository.

At this point we understood that building effective queries intra-CDR, would also require some knowledge about how resources could link to each other so that, additional integrity could be ensured - this was an easy understanding when defining the meta-information upon studying the *MIMIC-III* relational database. In general, each table contains either a) *id* of the subject which that information is related to or b) a foreign key to other table that has the *id* of the subject. This awareness led us considering the framework should have this "subject linkage" support as a feature.

Knowledge graphs demonstrated to be a powerful mechanism to interpret the query definitions and easily represent all properties in a fast, searchable and linkable way. Therefore, this constitutes one of the most crucial parts of the query builder acting as a consumer of the query (easily adjustable to syntax changes) and as an interface to the query builder part.

Relationship types (see Section 3.3.3) were also considered adequate and valid for the clinical domain but evaluators identified one additional type that is used often during research: the *NOT* operator. Aims to get results based on exclusion of particular contexts, leading to an increased number of results for further detailed analysis.

The proposed method relies on the access to the UMLS and the accuracy of it, to provide contextual information for the *CMIID* processes. However, it serves as a medical thesaurus and does not necessarily guarantee that similar terms are logically equivalent. Furthermore, UMLS does not provide 100% representation between terminology domains. Knowing this, the framework lays the responsibility of terminologies aggregation (context-based) to the users, so they can choose adequately the terms that are similar, depending on the limitations and definitions of each terminology, for their query core context. Nevertheless, in situations where UMLS does not guarantee a full representation between terminologies and codes that should be similar, our framework is able to support supplementary information.

The beforementioned concern, was also supported by the evaluators. They also stated UMLS context mappings are shaped to specific definitions under certain circumstances in the past, and may not be accurate for the codes available in the repositories. Two suggestions are given, regarding improvements to the query scheme:

- Ability to specify for each code or context, the codification variables - the characteristics that define the data (such as place, data collection mechanism, etc.);
- Ability to refer codes of a different domain, that are seemly in terms of equivalence.

The simplicity, openness and limberness of the query scheme were identified as positive characteristics: the support of a large amount of ontologies, domain knowledge to build queries and the possibility to add relationships between codes to achieve richer results.

As already mentioned, an additional evaluation (System Usability) was conducted with experts from distinct areas of healthcare, to study the applicability of the query formulation in their daily routines and clinical practice. It was asked the participants a set of relevant clinical research questions in their domains, which we manually translated into the *CMIID* scheme. Detailed results were shared back to make the evaluation process possible (see Appendix B).

Gathering feedback from different profiles was important, thus including people with distinct awareness, experience and sensitivity for this subject. Table 4.2, describes the characteristics of the participants.

Table 4.2: Usability Survey - descriptive characteristics of participants

Age	Gender	Areas of Activity	Sites	Domains
[30-40]	Female	Researcher	Medical Research Center	Health Information Systems and Electronic Health Records
[30-40]	Male	Principal Investigator; Physician; Lecturer	Medical Research Center; Head of a private outpatient unit; University Medical School	Patient Centered Innovation and Technologies; Clinical trials; Allergology and Clinical Immunology
[40-50]	Male	Medical Manager and Principal Investigator; Physician	Private Contract Research Organisation (CRO); Public Hospital;	Pharmaceutical, biotechnology, and medical device industries; Clinical trials; Clinical Pharmacology and Nephrology;
[20-30]	Female	General Practitioner	Public Hospital	-
[40-50]	Male	Clinical Pharmacology Director; Physician	Private Contract Research Organisation (CRO); University Medical School	Clinical Pharmacology and Pharmaceutical Medicine; Clinical trials; Clinical Research and Drug Development
[20-30]	Female	General Practitioner	Public Hospital	-
[20-30]	Male	Clinical Research Associate (CRA)	Private Contract Research Organisation (CRO)	Clinical trials;
[20-30]	Female	Study Coordinator and Data Manager	Institute of Oncology	Pathology; Clinical trials;
[30-40]	Female	General Practitioner	Public Hospital	-
[30-40]	Female	Physician	Public Hospital	Neurology

Ten participants enrolled in this study, with ages from 20 up to 50 years old, constituting a gender-heterogeneous group. Areas of activity comprehend physicians in public and private hospitals which were devoted to different areas of practice, researchers in medical research centers and principal investigators, research associates and study coordinators with experience in areas such as clinical trials.

One common characteristic in this group was that, everyone was aware of clinical codification and already used it in different opportunities - some coding information and others, making use of HIS and other sources to practice research. Despite the knowledge to interpret ontologies and its codes, only half of the group was coding in their professional routine.



Usability scores results from this study (see Figure 4.1), were represented using a Box plot to assess the distribution of scores and understand the variability outside the upper and lower quartiles.

We got a minimum score of 30 and a maximum of 87.5, with all the rest of the scores being located between the first quartile (score of 60) and the third quartile (score of 80). The minimum score we obtained, detaches from the rest of the distribution. It was from a participant with vast experience in distinct clinical domains and research projects, supporting that, clinical codification should be restricted to specific areas and not the clinical practice in general (using codes we lose significance and context).

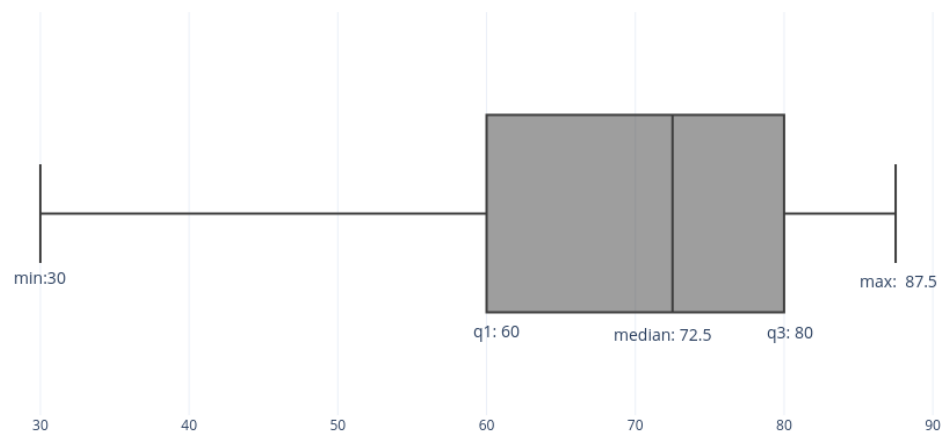


Figure 4.1: Box plot of the SUS results (scores between 0 and 100 represented in the x-axis), from a total of 10 participants evaluating the *CMIID* query scheme. Fifty percent of the population scored between 60 and 80. Two distinct outliers are identifiable: i) a lower outlier of 30 and ii) a upper max score of 87.5. According to the system evaluation, with median value of 72.5, the scheme is classified in-between "OK" and "GOOD".

Additionally, users mentioned that the use of the *CMIID* scheme will need additional vocabularies to overcome the ontological issues and limitations. In terms of using a mechanism to search for information in multiple sources, the participant stated the main solution should be firstly based on how the clinical process occurs in reality (using key phrases, natural language text, etc..) and secondly, how *CMIID* proposes to do which is optimal.

The remaining majority of the participants, positions the scheme concept evaluation within a range that classifies in-between "OK" and "GOOD", with a median value of 72.5 - meaning participants opinion vary between grade C and B (Bangor et al., 2009; Martins et al., 2015). Feedback from participants value the scope and features of our hybrid coding scheme, considering it as really important for research and clinical investigation. Moreover, they identified it as simple and

important tool for their professional routine, valuing the true meaning of data and easing research across multiple sites with an universal and common language.

Analyzing participant's feedback with the scores per question (see Figure 4.2) and the results from Figure 4.1, we understood contributors would like to test the scheme along with the framework to better understand the difficulty using it. Such willingness and due to nature of the study, led them to submit more neutral ratings on questions that targeted the self-use - question four and nine. In consequence, those questions had an impact in the first and third quartile scores, lowering them.

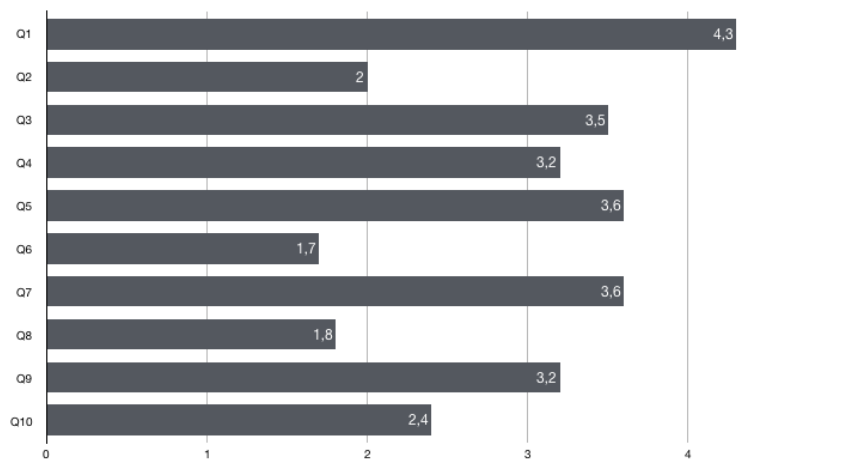


Figure 4.2: Average participant's scores per question in the SUS study. Participants rated each question with a score from 1 to 5. Scores were then converted to a score between 0 and 100, using a system formula. Question 4 and 9 target the easiness of the self-use of the scheme, which users gave neutral scores. This means users would like to test the scheme along with the framework for better understanding.

### 4.3 Framework architecture

The *CMIID* representation, as explained in Section 3.3.2, was designed to provide a comprehensive and domain-oriented way to represent clinical knowledge aiding on research and in the harmonization of search queries. The *CMIID* framework was developed to act as a software that can be installed in distinct environments so that researchers can execute optimized queries.

With the above mentioned features, we highlight several key reasons where *CMIID* is a differentiated solution:

- No need to learn specific query languages (e.g. SPARQL) to build a research question;
- Absence of data control layers, making available all repository attributes;
- Required knowledge comes directly from the domain;
- Automatic and continuous update of the universal coding systems - due to the use of UMLS as a foundation;
- Use of a ontology that can be a bridge to other systems and data representation.

As shown in Table 4.3 and in Table 4.4, several advantages and disadvantages are considered in four levels of study - integration, complexity, maintainability and scalability.

Important outcomes can be highlighted. Security, as we have seen in Chapter 2 remains an issue if a multi-location approach is considered. However, and based on the subject of this thesis, we are studying the applicability within the DSH paradigm and therefore it must not be considered as a barrier. In another perspective, the performance depends on several factors, some on the framework side that can be optimized, others are external dependencies that can be somewhat minimized.

Despite evaluators considered the framework well-defined, robust and trustworthy by using UMLS, they identified two concerning topics:

- Complexity management if new layers are added to the framework - this solution uses the UMLS active release with no restrictions or logic inherent but since each repository has its own characteristics it may be necessary to use: a) target versions and settings of UMLS or b) use simultaneously other vocabularies;
- Acting on top of UMLS (dependency and flexibility factors) - although using such service provides numerous advantages, the solution relies on the quality of the service and known limitations.

Both items are relevant and future work is needed to understand the impact of new changes, that could benefit the most from an integration in a real healthcare scenario.

As presented in Figure 3.16, multiple queries are dispatched by the framework according to the matching contexts between the main query and the CDR characteristics. This translation was successfully implemented and facilitated with a knowledge graph.

For the two data sources under consideration (see Section 3.2), we designed the templates responsible for the translation syntax (SQL and HTTP calls (web service) template). However, we have just covered the standard ones and many others can be defined, depending on the data retrieval mechanism bias to the source (e.g. web service authentication and authorization, response format, etc.).

As mentioned in Section 3.2, there was a context in common in the data sources - "Chemicals & Drugs". Despite this fact, both repositories use different internal linkage methods - *MIMIC-III* uses a subject ID (see Section 3.4.3) and *FDA* holds no linkage between records. No common attribute was found to aid on the bridge. Linking records was solely possible using the clinical context defined in the main query (see Section 4.2). This particularity, led us to implement internal cursors to bridge outcomes from searches aiding the manual analysis meaning that each search result was added with an ID of the other searches results that share common contexts (see Figure 4.3).

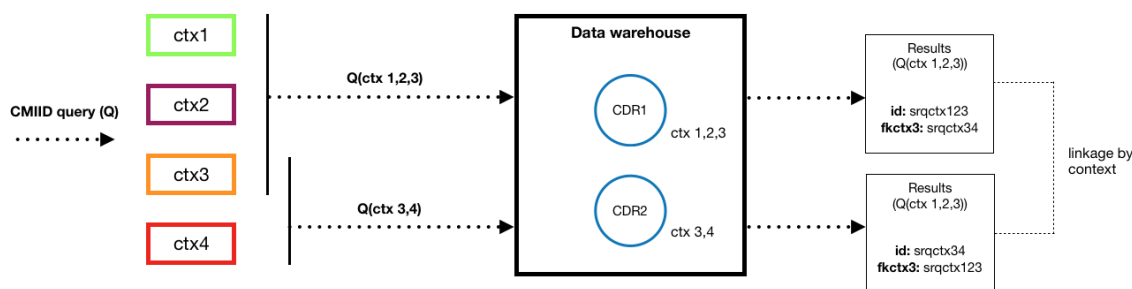


Figure 4.3: CMIID mechanism to link search results that share common contexts, using cursors based on search identifiers. If there are shared contexts between sub-queries and matching results in the repositories, the framework implements a linkage between records that belong to the same context.

Such refinement continues to be a burden process and therefore, evaluators suggested the repositories setup need to include more metadata and not only the contexts they have represented. Additional knowledge would help on establishing linkage patterns between resources of the same context from distinct CDRs, as we have seen in Section 4.2. It would help to understand the circumstances/variables in which the data was coded - to enable the imposition of even more specific details in each code rather than the UMLS definition.

Using a unified medical language system as a central piece of our framework, enables interoperability but also limits our domain of health and biomedical vocabularies. The *CMIID* framework does not support external vocabularies to the UMLS and therefore, adding them in a query (e.g. *MIMIC-III* DRG vocabulary) would impact our solution to ignore those codes during the process (see Section 4.1). There are several research studies focused on building effective mappings between UMLS and other ontologies, but they all underline distinct advantages and issues (Burgun and Bodenreider, 2001; McCray, 2003; Brandt et al., 2011; Falconer et al., 2007). Further investigation is required to study a new layer in the framework that can process codes from vocabularies not supported by UMLS.

The current implementation returns the maximum number of fields possible (from each repository), so researchers can adequately promote a refinement and analysis with better quality. Evaluators agreed this is a good principle and that helps significantly to have as much information as possible.

The previously mentioned improvements, led the evaluators to suggest the definition of a protocol to get data from registered repositories and to enhance linkage process. Acting as a standard it would dictate how new sources should be registered and how data should be made available. In which regards to data quality, they also underlined the framework has a huge potential and can be used to flag erroneous records, providing a mean to submit corrections and see other suggestions already added (e.g. alternative accurate codes).

Table 4.3: *CMIID* framework analysis in terms of solution acceptance - integration and complexity.

	Integration	Complexity
Description	Can the framework be integrated into existing solutions?	How complex is the framework and its dependencies?
Requirements	<ul style="list-style-type: none"> <li>- Granted access to databases, NATs or proxy gateways</li> <li>- CDRs descriptions</li> </ul>	<ul style="list-style-type: none"> <li>- UMLS services with a licensed account</li> </ul>
PROS	<ul style="list-style-type: none"> <li>- Easy to integrate into existing services: can run as an additional module</li> <li>- Can be used as SaaS (software as a service)</li> <li>- Setup time is almost zero</li> <li>- Knowledge require to setup: CDRs description</li> </ul>	<ul style="list-style-type: none"> <li>- Large Metathesaurus</li> <li>- Clinical terms and validation accuracy with the use of UMLS</li> </ul>
CONS	<ul style="list-style-type: none"> <li>- Security concerns (DSH paradigm overcomes this issue)</li> </ul>	<ul style="list-style-type: none"> <li>- Dependency on an external service: in case of failure the framework does not work on offline mode</li> <li>- Overall processing performance lacks optimization</li> <li>- Performance also dependent on UMLS performance and network conditions</li> </ul>

Table 4.4: *CMIID* framework analysis in terms of solution acceptance - maintainability and scalability.

	Maintainability	Scalability
Description	How much effort is required to maintain the framework?	Can the framework scale for extra demands?
Requirements	<ul style="list-style-type: none"> <li>- UMLS services with a licensed account</li> </ul>	<ul style="list-style-type: none"> <li>- To scale the CDRs pool, it is necessary to know the repository descriptive characteristics</li> </ul>
PROS	<ul style="list-style-type: none"> <li>- Framework uses the latest versions of UMLS (active releases)</li> <li>- Continuous updates on active release are done automatically</li> </ul>	<ul style="list-style-type: none"> <li>- New CDRs can be added easily</li> <li>- For high-volume processing, more instances can be added</li> <li>- UMLS also supports high volume of requests</li> <li>- No use of new mapping ontologies</li> </ul>
CONS	<ul style="list-style-type: none"> <li>- Dependency on an external service for complete maintenance</li> </ul>	

## 4.4 Framework performance

There are three variables we have considered when assessing the performance: 1) the number of siblings 2) the number of contexts and 3) the number of data sources (excluding the nature of the source). We have also established the following conditions:

- One context and zero siblings means having only one code in that context (e.g. "snomed-ctus\_441818008;");
- One context and two siblings means having two codes within the same context (e.g. "snomed-ctus\_441818008,rxnorm\_544393;").

The rule follows the same logic for more siblings;

- Siblings variation for more than one context uses two siblings per previous context. This means, with three contexts we varied the number of siblings in only one context and used two siblings per each one of the other two contexts (e.g. three contexts and five siblings):  
(e.g. "icd9cm\_787.01,icd9cm\_787.02,icd9cm\_787.0,icd9cm\_787.91,  
icd9cm\_009.3##cpt\_94003,cpt\_99.04##rxnorm\_866513,rxnorm\_544393;")

We have used this understanding to execute a measurement of the response times having in consideration the scenarios and sources described in Section 3.2.

The results presented in Table 4.5 and plot in Figure 4.4, concern the time spent by the framework processing the query before running it against the source. As upper measurement limits, we considered 10 siblings and 6 contexts because we were able to perceive from the results (above {C=3,S=5}) a growth pattern, practically linear.

Table 4.5: Framework average processing time (in seconds) varying number of contexts (C) and siblings (S), for all scenarios described in Section 4.1.

		# siblings (S)			
		0	2	5	10
# Contexts (C)	1	3s	4s	14s	25s
	2	7s	10s	25s	35s
	3	15s	20s	29s	40s
	4	21s	27s	34s	47s
	5	27s	35s	40s	55s
	6	36s	42s	48s	61s

The results indicate an increasing processing time when adding more contexts and therefore more siblings. With this said, it was important to assess individually, by siblings (see Table 4.6), the degradation of performance.

Looking to the summary results in Table 4.5, it was noticeable that adding more siblings caused a hit on the performance of the framework. Varying the number of contexts for S=2, we can see that for each extra context (two extra siblings) overall processing time increases approximately

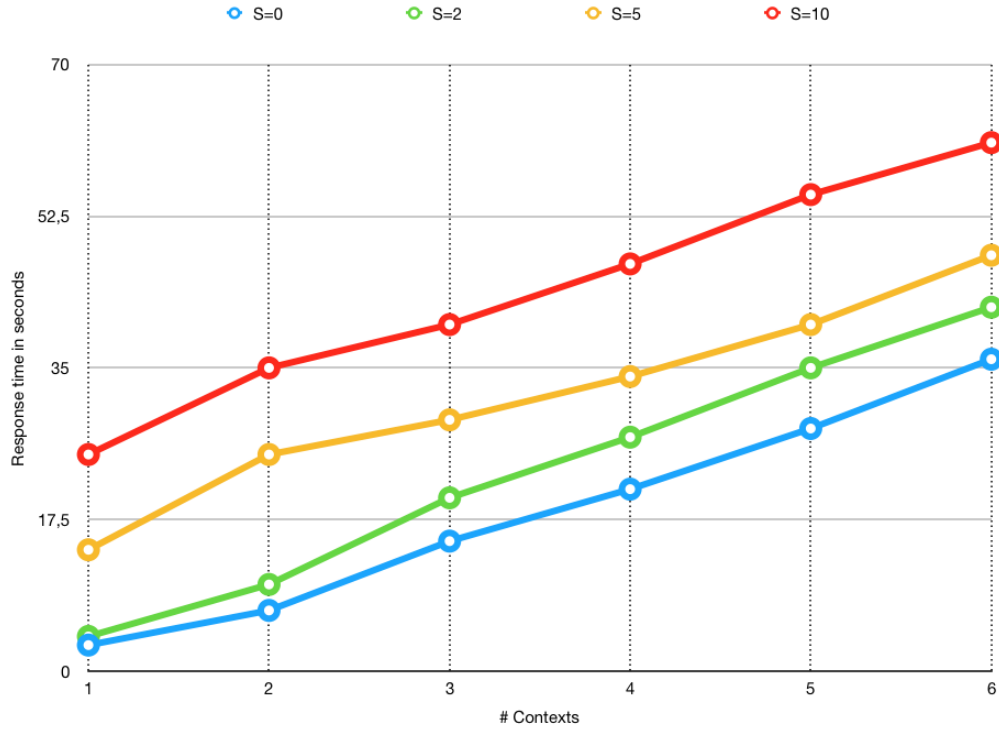


Figure 4.4: Benchmark of *CMIID* framework varying the number of contexts ( $C$ ) and siblings ( $S$ ). For a given context  $C$ , all  $C_{N-1}$  consider 2 siblings. Results indicate there is a cumulative cost over new additions of new codes.

8 seconds. Furthermore, whilst  $\{C=1, S=0\}$  took three seconds,  $\{C=1, S=10\}$  took 25 seconds, meaning that adding 9 extra codes cost 22 seconds. The same rationale is applied to the rest of the simulations being  $\{C=6, S=10\}$  the worst case scenario.

Moreover, from the Figure 4.4 we can understand that:

- We have an increasing cumulative cost, over new additions;
- Up to 3 contexts, there is no behavioral pattern. Adding more contexts, results in an almost linear growth.

Conclusions so far, did not provide us with a correct understanding of the volatility of the framework when adding more siblings per each context. We then measured the average cost of a new sibling per pair  $\{C, S\}$  - see Table 4.6.

Into detail:

- $\{C=1, S=0\}$  - only one code cost 3 seconds;
- $\{C=1, S=2\}$  - one new sibling cost 1 second (comparing with  $S=0$ );
- $\{C=1, S=5\}$  - three new siblings cost in average 4 seconds each (comparing with  $S=2$ );
- $\{C=1, S=10\}$  - five new siblings cost in average 2.2 seconds each (comparing with  $S=5$ ).



Table 4.6: Average time (in seconds) to process each new sibling within the same context. Values for each S column represent the average time each one of the new siblings took.

		# siblings (S)				Average
		0	2	5	10	
# Contexts (C)	1	3s	1s	4s	2.2s	2.55s
	2	3s	3s	5s	2s	3.25s
	3	5s	5s	3s	2.2s	3.8s
	4	1s	6s	2s	2.6s	2.9s
	5	1s	7s	2s	2.3s	3.3s
	6	1s	6s	2s	2.6s	2.9s

The results show that, irrelevant of the number of contexts, the average cost of each sibling varied between 2.5 seconds and almost 4 seconds - see Figure 4.5. In average, each required 3 seconds of processing time. Such variation allows us to conclude that the system response time is not predictable using a linear function. There is a significant dependency to the UMLS response times.

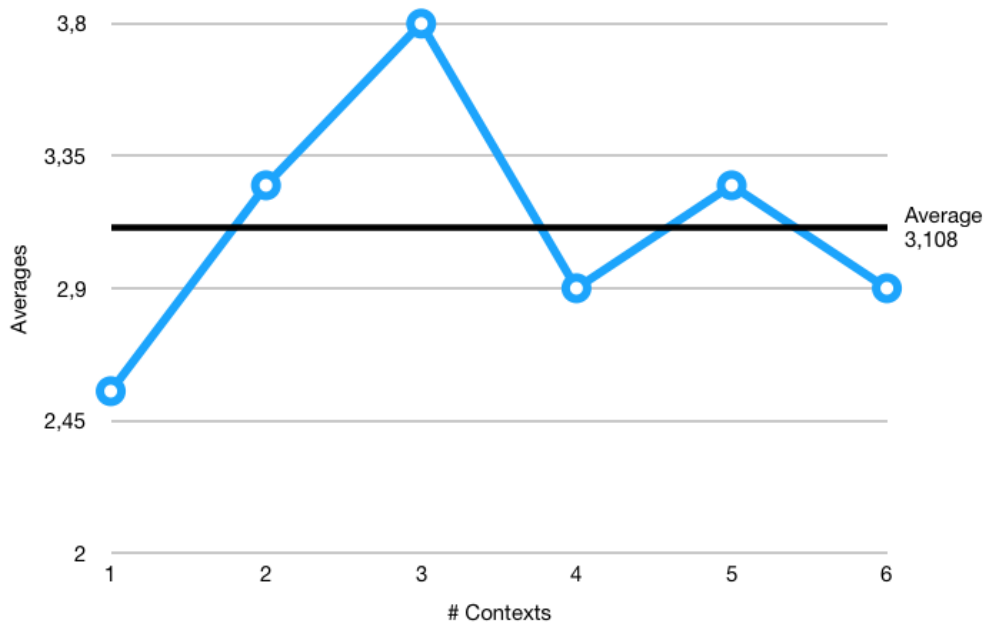


Figure 4.5: Average cost (processing time) of each sibling addition with the increase of the number of contexts. Average cost of each sibling varies between 2.5 seconds and almost 4 seconds

UMLS has a negative performance impact in the architecture. The use of its services requires an application access grant (one every 8 hours) and the use of that grant to get a dedicated token to use in each request. To build the knowledge graph we need to (per code):

1. Search for its information (one request - average of 500 milliseconds reply);
2. Search for its semantic groups (one request - average of 500 milliseconds reply).

With this said, this cost is directly proportional to the number of siblings - see Figure 4.6. As we can see that in the worst case scenario  $\{C=6, S=10\}$ , we spent 20 seconds out of 61 (one third of the time) performing UMLS requests.

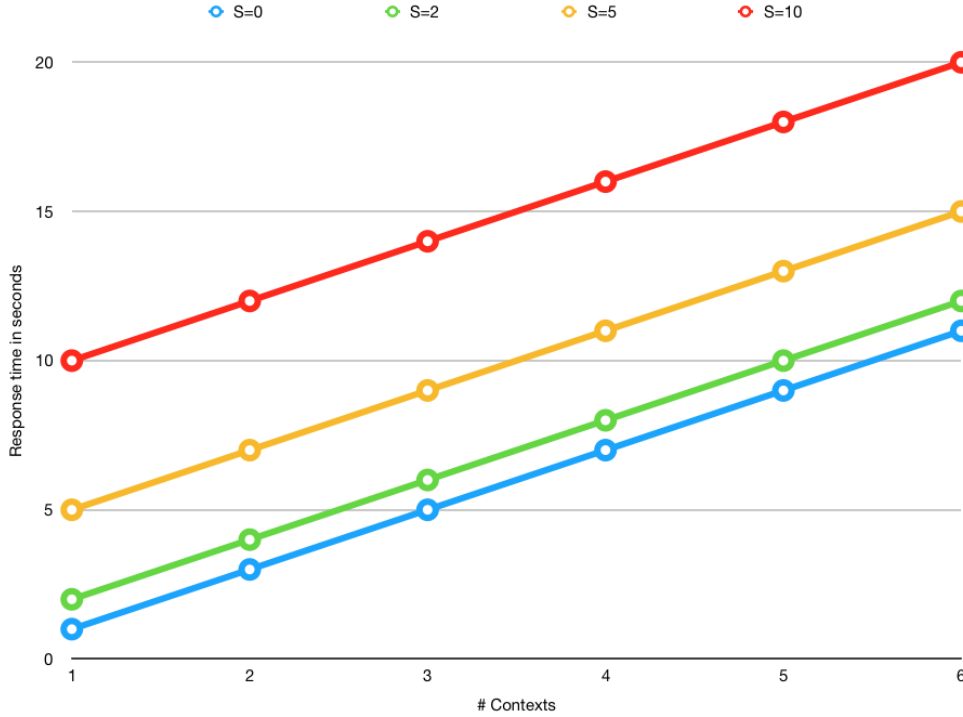


Figure 4.6: Time spent performing UMLS requests, varying the number of contexts (C) and siblings (S). With an increase of the number of siblings and contexts, the framework performance suffers a degradation mostly on UMLS requests, one third of the time.

On top of these results we have previously explained, we need to consider 1) the time to execute the query and return the results to the framework and 2) linkage of results on shared contexts for all data sources.

Sources have distinct characteristics (e.g. web service or database) which impact the response time of the framework in making the results available for the researcher:

- Service location (localhost vs outdoors);
- Configuration of the service and its performance (e.g. caching, indexing, integrity keys);
- Nature (e.g. database vs web service).

For the scenarios mentioned before, we got average response times of 3 seconds for *FDA* and 5 seconds for *MIMIC-III*. Although *MIMIC-III* was running in a localhost server, the *CMID* queries we used joined a significant amount of records, causing higher response times.

Adding more sources activates the framework parallelization mode, running queries in parallel. However, the system waits for the slower thread, i.e., the one that takes more time to answer - in this study *MIMIC-III* was always the slowest service.

The beforementioned results use internal cursors to link records from all intra-searches, and therefore aid the manual analysis, as we explained in Section 4.3. An evaluation of the system performance considering a different type of linkage across search results was not possible and therefore, requires further investigation to understand which metadata should be set on the source registration. This operation did not reveal demanding processing cost, on the opposite.

## 4.5 Solution positioning

A novel solution such as *CMIID*, intends to leverage researchers with better tolling so healthcare research can be promoted further. Such impact analysis must first start with an understanding of the tool itself and which are the contributions, needs and pitfalls - as we did in this chapter. Additionally, is important to compare it with existing solutions that share similar characteristics, domain and intents.

Having this in mind, we have compared (see Table 4.7) our solution with the reference one from the literature, as we have introduced in the Chapter 2 - SAIL databank. The comparison, based on five different topics, seeks to understand the benefits of using one solution in favor of the other, and the most valuable characteristics in this field of expertise.

Table 4.7: Comparison between *CMIID* framework and an advanced solution in the research field - SAIL Databank.

Access and policies	Controlled access to data Application of Policies and Authorities requirements	SAIL	CMIID
		Yes, using authorization grids  Yes	Yes, under the DSH paradigm  Yes, under the DSH paradigm
Data integrity	Use of a Linkage algorithm	Yes, using MACRAL to assign an anonymous linking field to person-based records	Yes, linking records based on clinical contexts
	System to identify and tag data quality issues	No (Ford et al., 2009)	No
Anonymity and disclosure of information	Anonymisation and encryption of commonly-recognized identifiable variables	Yes	No, relies on repositories disclosure layers
	Methods to control risk of disclosure	Yes, based on policies and authorities compliance	Not, relies on DSH compliance
	Additional layers of safeguards to ensure anonymity is protected	Yes	No
System	Plug-in system (system can be integrated into other solutions)	No (Ford et al., 2009)	Yes
	Requires manual intervention in certain steps	Yes, 1. data needs to be provided by Data Providing Organizations 2. Technical intervention on data before making it available (e.g. views)	Yes, on repositories registration
Search	System scalability when new information is added	Low, restricted to the views schema	High, permissive to repositories schema
	Information comprehensiveness	Low, restricted to the information available in the views (format, codification, et.)	High, through multi-terminology and distinct repository types support

Both solutions are compliant (in different ways) with legal requirements, policies and authorities demands, which is really important nowadays for tools that access confidential data. Whilst, one has to fulfill all requirements internally, the other is under the DSH paradigm which oversees all requirements in a high-rank stage.

*SAIL* has an advanced and extensive layered architecture in which regards to anonymity and disclosure of information. In this perspective, our solution does not embrace any mechanism, relying exclusively on the repositories schema available in the Safe Haven. Having methods to control the risk of disclosure and automatic validation for anonymisation and encryption of PHI, would be beneficial ensuring a redundant but still safe architecture in a DSH approach. This particular aspect, is more meaningful when considering using the framework as a plug-in system, meaning it could be hosted in a different system with distinct policies and the same level of privacy and security would remain.

A significant difference between the two solutions is the access to data and how a researcher can search for information. On the contrary of *SAIL*, which has a fully-manual process of loading sources and building views compliant with policies and whom will access the information, our solution requires some minimum repository information: authentication, clinical contexts and integrity keys.

The searching mechanism differs on both approaches as well: with *CMIID* it relies completely on the researcher knowledge to build the query, then the system will retrieve all matching records and variables that may be relevant - as much as metadata possible; *SAIL* only exposes the necessary information to the type of consumer is intended to. Additional variables demand, requires going through the process of validation and schema rebuild. Overall, this indicates if sources change in terms of structure (new/renamed/removed attributes), codification or new information, our solution can automatically adapt to it - except if is a structural change and integrity keys and contexts need to be changed - this needs manual intervention as well.

Making information available using views ensures a better control on the permissions and disclosure of data, however, it constitutes a significant blocker to promote flexible and dynamic research, by having access to more variables and contexts. Supporting multiple terminologies to formulate queries along permissive schema and a search linkage based on clinical context, we may be able to unlock new ways to get more comprehensive information.

## 4.6 Summary

In this chapter we have introduced and explained the results we obtained with the *CMIID* framework on distinct evaluation areas: i) performance; ii) user usability using a Focus Group and a SUS questionnaire) and; iii) solution positioning with *SAIL*.

Results show that for each new context (with two siblings) the overall processing time increases, approximately 8 seconds. This represents an increasing cumulative cost, over new additions, whereas one third of the time is spent on UMLS requests. The participants in the studies evaluated the framework with a median score of 72.5, meaning the scheme is classified as

"GOOD". An significant improvement was identified in order to overcome the lack of definition and accuracy on UMLS terms: add support for additional metadata in the query scheme. The *CMIID* is a very flexible and intuitive solution, allowing the researchers to formulate advanced context-based queries on a cluster of repositories, being agnostic to its characteristics. This acts as an important novel solution when comparing to SAIL, which enforces views to access data and several manual processes.

## Chapter 5

# Conclusions and Further Research

This chapter summarizes the main conclusions of the research work that has been done. The main contributions of this thesis are highlighted, as well as its limitations. Future research directions addressing the main limitations of the work are proposed.

### 5.1 Thesis overview

Undoubtedly researchers go through difficult situations to access clinical data, to understand it and to develop new solutions. The volume of electronic records is increasing at a significant pace. Each year the number of new transitions in the digital form is overwhelming, giving space to several concerns: security; harmonization; data location; coding; format and quality; etc. Thus, research has been focused on techniques to handle these issues and to promote a safe usage of clinical data based on the current *status quo* of healthcare.

When trying to search for information, researchers face four main issues:

- Security and legislation demands - the DSH emerged to diminish this obstacle offering a way-in to access the data;
- Data location - data can be spread in different locations;
- Data coding - depending on the source and characteristics, clinical information may be coded with different standards complying distinct requirements;
- Search harmonization - data from different sources provide users with a view of all meaningful results.

From Data Safe Havens to standard ontologies and ontology mappings, to blocking methods and to virtualization techniques (e.g. RDF), researchers have strived to identify solutions to promote new developments. However, no approach has yet been labeled as optimal in each one of the areas and as so the search process for clinical information is also suffering from these issues.

We have developed a framework that allows researchers to build advanced context-based clinical queries, supporting multiple vocabularies to define the events and whose queries are executed

against a cluster of registered repositories. Such framework contributes with a new, clean and domain-friendly search scheme with the capability of fetching results from distinct types of CDR, without using virtualization techniques.

We have thus elected the search for clinical records in distinct sources, as the main problem to be addressed in this thesis. The application of Graph Theory, Information Retrieval and Boolean Algebra laws was proposed and implemented as a solution to this problem.

The main research question that guided this thesis was "Using a hybrid thesaurus coding scheme embracing a multi-terminology approach, are we able to supply a search solution that harmonizes queries across multiple distinct clinical data sources?"

To answer this main research question we had to solve two main problems. The first one is how to build a query scheme capable of handling advanced knowledge representation and the second one is how to implement a search mechanism resulting from the solution of the first problem. Each of these two problems raised new research questions, that were answered throughout this thesis.

For the first problem: to understand what should be the specifications of a new query scheme and how it could be suitable for further developments, we have conducted several interviews with researchers and physicians from a Portuguese research center. This method, involved in first stage, the study of the healthcare environment where the research center was inserted into and also the understanding of the professional experience of the interviewed (described in Chapter 3). Consequently, an evaluation of the most adequate features was conducted based on the literature review and the outcomes from the field study previously mentioned. The analysis showed that the most important topics were the following: 1) standard and adaptive query language; 2) definition of conditional clauses; 3) options to refine results and 4) up-to-date scheme able to use the latest updates from the vocabularies.

To solve the first problem, a second and crucial phase was carried out, i.e., to develop a query scheme methodology that involves the construction of a solution capable of answering the needs before mentioned, plus providing a solid ground for different healthcare applicabilities. This second phase (described in Section 3.3) required first of all, the construction of a skeleton and then the definition of how the syntax properties would be. To be able to support the most-used and standard vocabularies nowadays and to ensure automatic updates and reliability on the system, we decided to use the UMLS services and semantic groups to shape the syntax. Current categorization and definition by the UMLS for these groups are broaden and detailed enough to be used with this solution, but depending on the nature of the data more specificity may be required (e.g. using semantic types rather than groups).

Using the UMLS we were also able to develop several tests to understand if we could define in a conceptual level, different clinical events with conditional clauses. Additionally, if using the services, we could interpret the relationships accordingly and most important, the core of the query. We have applied Graph Theory on top of the scheme to allow an accurate query validation (redundancy of codes for the same clinical context, codes existence, etc.) and an advanced knowledge representation to the search mechanism, ensuring an agnostic definition and implementation. Results show that the use of UMLS, is extremely valuable to act as the core component of the



query scheme, either supporting the *CMIID* query syntax, validating the codes and relationships between them, automatic updates and the additional knowledge inherent in its services (allows further developments).

For the second problem, we have proposed a framework that automates the execution of queries in the CDRs. Two important phases are hereby presented: 1) framework outline (detailed in Section 3.4) and 2) query execution and search harmonization technique (detailed in Section 3.5).

To solve the first part, we developed a framework that uses the knowledge representation resulting from the applicability of Graph Theory, to build queries adjusted to the registered repositories. Having in mind the information collected from the interviews and the analysis of two real-life repositories - (*US FDA API* and *MIMIC-III*) - a set of characteristics needed from the repositories was defined, to allow effective query translations. The analysis showed that, the most important topics were the following: 1) the repository type (e.g. web service or database); 2) the authentication credentials; 3) the clinical contexts available (which UMLS semantic groups exist and which resources hold information coded per each context), and 4) the integrity keys to enable records linkage intra-CDR.

Regarding the second phase, we have developed a search harmonization technique that is able to translate the initial query into several queries depending on each CDR characteristics such as the clinical context there available. We were able to conclude that queries are correctly being translated and executed in the target repositories. The proposed algorithm presented good results, in both translation and search mechanisms. Nonetheless, we have identified some impactful concerns: 1) the deterioration of the system performance with the increase of query complexity (UMLS has a significant impact); and 2) the need for more metadata in the repositories and queries, to enable results linkage when there are no shared contexts between CDRs or when they share similar schema characteristics (with common subject ID attributes that allow referencing).

We have used two repositories that have certified and validated clinical information and that are widely used by researchers in their studies, to simulate a healthcare environment in order to test this framework. Using them proved to be very useful to identify bottlenecks in our application, assess the system performance and study the framework implementation and evolution. Additionally, we have conducted a Focus Group and a System Usability Scale questionnaire with groups of researchers and physicians (from different domains) from the Portuguese health technology and research center already mentioned, in order to evaluate the framework. The participants highlighted the simplicity, openness and limberness of the query scheme, as the advantage of using this solution in research environments. The *CMIID* framework was rated with a score of 72.5 (SUS).

## 5.2 Further Research

For the thesis experiments we have only used the *US FDA API* and *MIMIC-III* repositories but it would be beneficial if we could also have used healthcare repositories that present different characteristics and data records.

We have based our framework on the UMLS features, however it has several limitations that are crucial such as performance. Knowing this, an exhaustive analysis should take place to outline the improvements on the framework side to handle them more effectively - e.g. parallelizing contexts processing (in order to avoid iterative processing and high costs) and implementing caching layers (Markatos, 2001).

Another open issue, is the lack of support for the *NOT* operator in the query syntax. We did not consider it and it is used often during clinical research.

For the aim of this thesis we have used the UMLS system as source of truth and therefore the usage of a code in the query takes its definition as the only one possible. However, the coding process may have been contingent to additional attributes that UMLS is not considering. This metadata should be made available on the repository setup so that on the query definition, and for a each code, we can explicitly indicate more detailed information. Understanding the characteristics of this metadata and how it could be used in the query scheme, are open questions.

As further improvements are concerned, we identify two areas of investigation: 1) the system ability to identify and tag data quality issues and 2) introducing additional layers of safeguards to ensure anonymity. Introducing in the system the possibility to identify and tag data quality issues, would help researchers to understand that for a specific code or set of codes, the records available in the repositories are not coded correctly. This could be possible, having the additional metadata mentioned previously plus the researcher manual intervention to tag the records affected by the query (via the system). Further lookups on that codes would alert for existing tags.

Concerning the latter remark, our solution does not have layers of safeguards to ensure anonymity, since it is focused on the DSH paradigm. Nonetheless, and considering the integration in other systems and data sources via the plug-in characteristic, it is important we ensure this - a proper solution must be investigated in order to not cause degradation of the system performance. Named-Entity Recognition (NER) routines could be used for this task.

Another aspect, and a complex one, is the linkage of search results when there are shared contexts between repositories. The characteristics we first outlined have proven not to be sufficient and more metadata should be added to the registration process to enhance the linkage - for example sources share a common context but the vocabulary in one of them is not available in the UMLS.

The progress done so far and the open questions led to the integration of the *CMIID* into a Brazilian National Tuberculosis Network, containing treatment follow-up information, monthly diagnostic and control examinations, medication regimen, hospitalization and daily medication intake, since 2000 (see Figure 5.1).

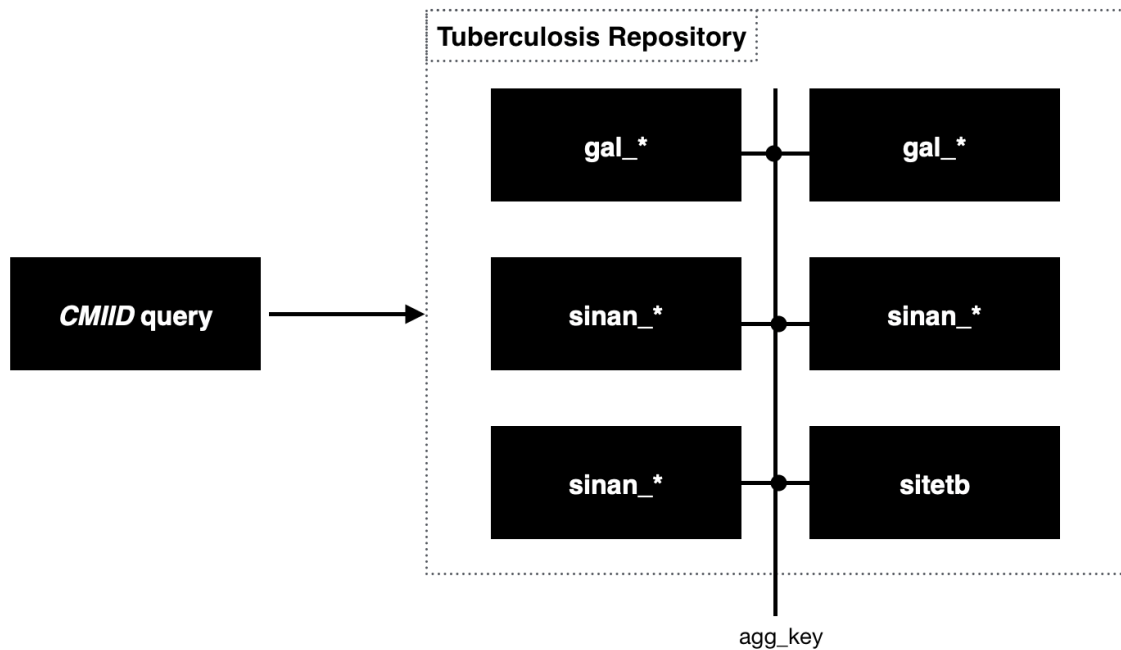


Figure 5.1: Integration of Brazilian National Tuberculosis Network with *CMIID*. Repository has 3 different sources and use an aggregated key to identify individuals. Repositories **gal\_\***, **sinan\_\*** and **sitetb\_\*** contain laboratory, disease and health cases and treatment records, respectively.

The sources available in this repository contain laboratory information (**gal\_\***), notification and investigation of disease and health cases listed on the national compulsory notification list (**sinan\_\***), and records from the Tuberculosis Special Treatment Information System (**sitetb\_\***). All sources disclosure individuals personal information and medical history, therefore the linkage between sources is possible using an aggregated key of 3 fields: patient name, date of birth and mother's name.

To integrate this with *CMIID*, changes are required in the "PATIENT\_KEY" attribute set in the configuration file (see Listing 3) as well in the search engine to consume adequately the aggregated fields. With this project we seek to develop a better understanding of known limitations in a different environment, promoting a solution-driven approach.



## Appendix A

Table A.1: Additional clinical queries used on the evaluation of the *CMHD* framework.

Additional Queries
=snomedctus_441818008;
=icd9cm_003.0,icd9cm_003.1##icd9cm_9910,cpt_99291##rxnorm_866513,rxnorm_0452; icd9cm_003.0*icd9cm_96.6;cpt_99291**icd9cm_43.7;
=icd9cm_003.0; icd9cm_003.0**cpt_99291;
=icd9cm_787.01,icd9cm_787.02,icd9cm_787.0,icd9cm_787.91,icd9cm_009.3,icd9cm_338,icd9cm_789.0, icd9cm_783.0,icd9cm_263.0,icd9cm_263.1,icd9cm_262.0,icd9cm_263.8,icd9cm_263.9, icd9cm_285.8,icd9cm_285.9##cpt_94003;
=icd9cm_003.0,icd9cm_003.1##icd9cm_991.0,cpt_99291##rxnorm_866513,rxnorm_0452,rxnorm_866513, rxnorm_214907,rxnorm_197902;
=icd9cm_003.0,icd9cm_003.1##icd9cm_991.0,icd9cm_991.2,icd9cm_991.3,icd9cm_991.1, icd9cm_991.4,icd9cm_991.8,icd9cm_991.1,icd9cm_991.6,icd9cm_991,cpt_99291;
=snomedctus_2492009,snomedctus_276608005,snomedctus_129845004,snomedctus_422587007, icd9cm_776.5,icd9cm_787.02##snomedctus_266700009,cpt_94003;
=icd9cm_003.0,icd9cm_003.1##icd9cm_991.0,icd9cm_991.2,icd9cm_991.3,icd9cm_991.1,icd9cm_991.4, icd9cm_991.8,icd9cm_991.1,icd9cm_991.6,icd9cm_991;
=snomedctus_409971007##snomedctus_39579001,icd10cm_T78.2B,icd10cm_T78.2C## snomedctus_468846009,rxnorm_1661387,rxnorm_1661398;
=icd9cm_493,snomedctus_195967001##cpt_99211,cpt_99212,cpt_99213,cpt_99214,cpt_99215,hcpcs_CO;
=icd9cm_003.0,icd9cm_003.1##=icd9cm_003.0,icd9cm_003.1;
=icd9cm_787.01##cpt_94003;icd9cm_787.01*icd9cm_787.02,icd9cm_787.0, icd9cm_787.91,icd9cm_009.3,icd9cm_338,icd9cm_789.0,icd9cm_783.0,icd9cm_263.0,icd9cm_263.1, icd9cm_262.0,icd9cm_263.8,icd9cm_263.9,icd9cm_285.8,icd9cm_285.9;

## Appendix B

# Comprehensive Medical Information Identifier (CMIID)

Nowadays governments and healthcare entities are more aware of the benefits of using patient data from different domains and locations to promote research advancements. However there are strong barriers such as ethical, financial, security and data search (several paradigms have emerged to overcome these issues - e.g. Data Safe Havens and SAIL Databank).

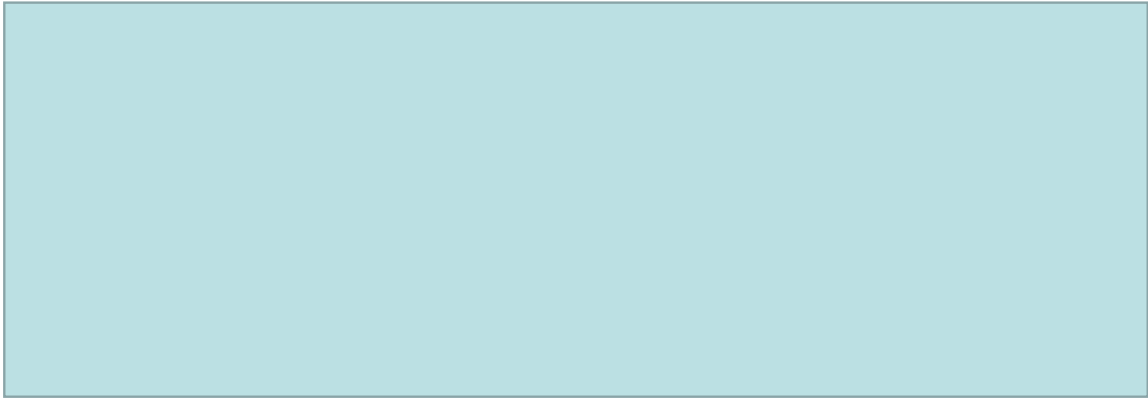
In terms of data search interoperability, researchers struggle with different source codifications, different query systems, the need to learn new search mechanisms requiring prior knowledge about the sources.

This understanding led us to develop a hybrid coding scheme to formulate clinical questions and consequently use it to enhance the harmonization of queries in the context of clinical data lakes (via a framework we developed) - capable of supporting clinical questions using a definition that all experts understand and are familiar to.

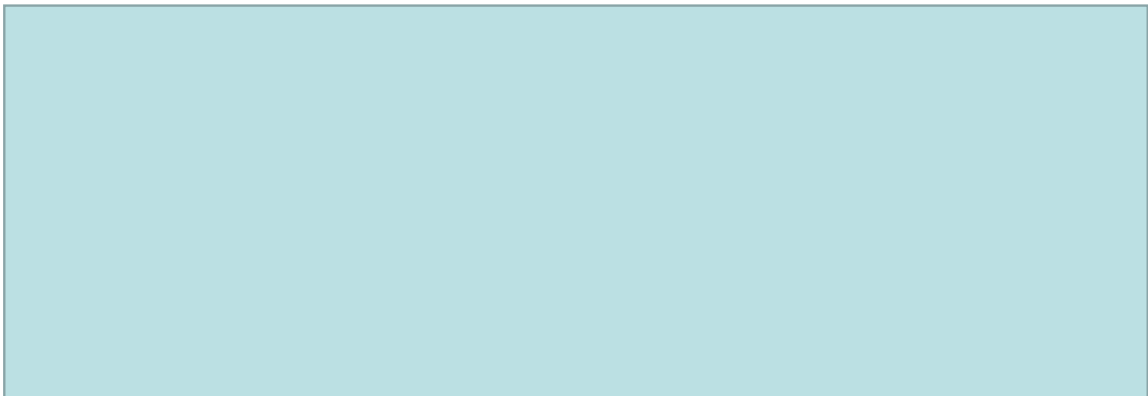
This follows the idea that researchers can formulate queries without knowing exactly the sources codification and just build universal queries based on standard codes. Some codes from the query may have a match on the repositories while others not. Such proposal allows the query to last in time and embrace several codes if for some reason, sources add a new codification.

# Clinical questions

Researcher question

A large, empty light blue rectangular box, likely intended for a researcher to write a question.

CMIID translation (a possible one)

A large, empty light blue rectangular box, likely intended for a CMIID translation of the researcher's question.

# CMIID translation

**Note:** This system uses *UMLS* active release. Codes prefixes are the Root Source Abbreviations (RSABs) in the Metathesaurus.

## Explanation



<In-detail explanation (part 1)>



<In-detail explanation (part 2)>

**Note:** <More explanation>



## System Usability Scale

**Instructions:** For each of the following statements, mark one box that best describes your reactions to the system.

		Strongly Disagree				Strongly Agree
1.	I think that I would like to use this system frequently.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2.	I found this system unnecessarily complex.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3.	I thought this system was easy to use.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4.	I think that I would need assistance to be able to use this system.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5.	I found the various functions in this system were well integrated.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6.	I thought there was too much inconsistency in this system.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7.	I would imagine that most people would learn to use this system very quickly.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8.	I found this system very cumbersome/awkward to use.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9.	I felt very confident using this system.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10.	I needed to learn a lot of things before I could get going with this system.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Please provide any comments about this system:



# Bibliography

- Abolhassani, N., Tung, T., Gomadam, K., and Ramaswamy, L. (2016a). Knowledge graph-based query rewriting in a relational data harmonization framework. In *2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC)*, pages 433–438. IEEE.
- Abolhassani, N., Tung, T., Gomadam, K., and Ramaswamy, L. (2016b). Knowledge graph-based query rewriting in a relational data harmonization framework. In *Collaboration and Internet Computing (CIC), 2016 IEEE 2nd International Conference on*, pages 433–438. IEEE, IEEE.
- Alakrawi, Z. M. (2016). Clinical terminology and clinical classification systems: A critique using ahima’s data quality management model. *Perspectives in Health Information Management*, 45(Summer).
- Almeida, J. P. (2016). A disruptive big data approach to leverage the efficiency in management and clinical decision support in a hospital. *Porto Biomedical Journal*, 1(1):40–42.
- Andargoli, A. E., Scheepers, H., Rajendran, D., and Sohal, A. (2017). Health information systems evaluation frameworks: A systematic review. *International Journal of Medical Informatics*, 97:195–209.
- Angles, R. (2012). A comparison of current graph database models. In *2012 IEEE 28th International Conference on Data Engineering Workshops*, pages 171–177. IEEE.
- Avillach, P., Coloma, P. M., Gini, R., Schuemie, M., Mougin, F., Dufour, J.-C., Mazzaglia, G., Giaquinto, C., Fornari, C., Herings, R., et al. (2012). Harmonization process for the identification of medical events in eight european healthcare databases: the experience from the eu-adr project. *Journal of the American Medical Informatics Association*, 20(1):184–192.
- Balamurugan, R. and Zubar, H. A. (2018). The application of graph theory in healthcare sector.
- Baliga, R. R. (2012). *250 Cases in Clinical Medicine E-Book*. Elsevier Health Sciences.
- Bangor, A., Kortum, P., and Miller, J. (2009). Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123.
- Bangor, A., Kortum, P. T., and Miller, J. T. (2008). An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction*, 24(6):574–594.

- Bansler, J. P. and Havn, E. (2010). Pilot implementation of health information systems: Issues and challenges. *International journal of medical informatics*, 79(9):637–648.
- Baxter, R., Christen, P., and Churches, T. (2003). A comparison of fast blocking methods for record linkage. In *ACM SIGKDD*, volume 3, pages 25–27. Citeseer, Citeseer.
- Berg, M. (2001). Implementing information systems in health care organizations: myths and challenges. *International journal of medical informatics*, 64(2-3):143–156.
- Berge, C. (2001). *The theory of graphs*. Courier Corporation.
- Betawadkar-Norwood, A., Pirahesh, H., and Simmen, D. E. (2013). Efficient processing of queries in federated database systems. US Patent 8,538,985.
- Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl\_1):D267–D270.
- Braithwaite, J., Hibbert, P., Blakely, B., Plumb, J., Hannaford, N., Long, J. C., and Marks, D. (2017). Health system frameworks and performance indicators in eight countries: A comparative international analysis. *SAGE Open Medicine*, 5:2050312116686516.
- Brandt, M. M., Rath, A., Devereau, A., and Aym  , S. (2011). Mapping orphanet terminology to umls. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 194–203. Springer.
- Brooke, J. et al. (1996). Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.
- Burgun, A. and Bodenreider, O. (2001). Mapping the umls semantic network into general ontologies. In *Proceedings of the AMIA Symposium*, page 81. American Medical Informatics Association.
- Burton, P. R., Murtagh, M. J., Boyd, A., Williams, J. B., Dove, E. S., Wallace, S. E., Tass    , A.-M., Little, J., Chisholm, R. L., and Gaye, A. (2015). Data safe havens in health research and healthcare. *Bioinformatics*, 31(20):3241–3248.
- Campbell, J. R., Brear, H., Scichilone, R., White, S., Giannangelo, K., Carlsen, B., Solbrig, H. R., and Fung, K. W. (2013). Semantic interoperation and electronic health records: context sensitive mapping from snomed ct to icd-10. In *MedInfo*, pages 603–607.
- Cardillo, E. (2015). Mapping between international medical terminologies.
- Chute, C. G., Cohn, S. P., and Campbell, J. R. (1998). A framework for comprehensive health terminology systems in the united states. *Journal of the American Medical Informatics Association*, 5(6):503–510.
- Ciccarelli, S. M. (1999). System and method for query translation/semantic translation using generalized query language. US Patent 6,009,422.

- Cimino, J. J. (1998). Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of information in medicine*, 37(4-5):394.
- Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387.
- Cragun, B. J., Fish, D. R., Rath, C. T., and Wall, D. A. (2007). Federated query management. US Patent 7,243,093.
- Cruz-Correia, R. J. (2010). Implementation, monitoring and utilization of an integrated hospital information system - lessons from a case study. *MedInfo*, 2010:238–241.
- Cruz-Correia, R. J., Vieira-Marques, P. M., Ferreira, A. M., Almeida, F. C., Wyatt, J. C., and Costa-Pereira, A. M. (2007). Reviewing the integration of patient data: how systems are evolving in practice to meet patient needs. *BMC medical informatics and decision making*, 7(1):14.
- Doiron, D., Burton, P., Marcon, Y., Gaye, A., Wolffenbuttel, B. H., Perola, M., Stolk, R. P., Foco, L., Minelli, C., and Waldenberger, M. (2013). Data harmonization and federated analysis of population-based studies: the bioshare project. *Emerging themes in epidemiology*, 10(1):12.
- Dotson, P. (2013). Cpt® codes: What are they, why are they necessary, and how are they developed? *Advances in Wound Care*, 2(10):583–587.
- Dusetzina, S. B., Tyree, S., Meyer, A., Meyer, A., Green, L., and Carpenter, W. (2014). *Linking data for health services research: a framework and instructional guide*. Agency for Healthcare Research and Quality (US), Rockville (MD).
- DuVall, S. L., Fraser, A. M., Rowe, K., Thomas, A., and Mineau, G. P. (2012). Evaluation of record linkage between a large healthcare provider and the utah population database. *Journal of the American Medical Informatics Association*, 19(e1):e54–e59.
- Ehrlinger, L. and Wöß, W. (2016). Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48:1–4.
- Falconer, S. M., Noy, N. F., and Storey, M.-A. D. (2007). Ontology mapping-a user survey. In *OM*. Citeseer.
- Fang, H. (2015). Managing data lakes in big data era: What’s a data lake and why has it become popular in data management ecosystem. In *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, pages 820–824. IEEE.
- Ferrante, A. and Boyd, J. (2010). Data linkage software evaluation: a first report (part i). *Public Health Research Network Centre for Data Linkage, Curtin University: Perth, Western Australia*, page 45.

- Ford, D. V., Jones, K. H., Verplancke, J.-P., Lyons, R. A., John, G., Brown, G., Brooks, C. J., Thompson, S., Bodger, O., and Couch, T. (2009). The sail databank: building a national architecture for e-health research and evaluation. *BMC health services research*, 9(1):157.
- Fox, J., Meir, R., and Schreiber, Z. (2013). Method and system for federated querying of data sources. US Patent 8,412,746.
- Garla, V. N. and Brandt, C. (2012). Semantic similarity in the biomedical domain: an evaluation across knowledge sources. *BMC bioinformatics*, 13(1):261.
- Gross, J. L. and Yellen, J. (2004). *Handbook of graph theory*. CRC press.
- Harron, K., Wade, A., Gilbert, R., Muller-Pebody, B., and Goldstein, H. (2014). Evaluating bias due to data linkage error in electronic healthcare records. *BMC medical research methodology*, 14(1):36.
- He, S., Hurdle, J. F., Botkin, J. R., and Narus, S. P. (2010). Integrating a federated healthcare data query platform with electronic irb information systems. In *AMIA Annual Symposium Proceedings*, volume 2010, page 291. American Medical Informatics Association, American Medical Informatics Association.
- Hripcsak, G., Duke, J. D., Shah, N. H., Reich, C. G., Huser, V., Schuemie, M. J., Suchard, M. A., Park, R. W., Wong, I. C. K., Rijnbeek, P. R., et al. (2015). Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. *Studies in health technology and informatics*, 216:574.
- Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.
- Khan, A., Uddin, S., and Srinivasan, U. (2016). Adapting graph theory and social network measures on healthcare data: a new framework to understand chronic disease progression. In *Proceedings of the Australasian Computer Science Week Multiconference*, page 66. ACM.
- Knoppers, B. M. and Chadwick, R. (2015). The ethics weathervane. *BMC medical ethics*, 16(1):58.
- Ko, E.-J., Lee, H.-J., and Lee, J.-W. (2006). Ontology-based context-aware service engine for u-healthcare. In *Advanced Communication Technology, 2006. ICACT 2006. The 8th International Conference*, volume 1, pages 632–637. IEEE, IEEE.
- Krishnankutty, B., Bellary, S., Kumar, B. N., and Moodahadu, L. S. (2012). Data management in clinical research: an overview. *Indian journal of pharmacology*, 44(2):168.
- Kumar, P. and Clark, M. L. (2012). *Kumar and Clark’s Clinical Medicine E-Book*. Elsevier health sciences.

- Kuo, J. and Kuo, A. (2017). Integration of health information systems using hl7: A case study. *Studies in health technology and informatics*, 234:188–194.
- Kuo, M.-H., Kushniruk, A., and Borycki, E. (2011). A comparison of national health data interoperability approaches in taiwan, denmark and canada. *International Journal of Electronic Healthcare*.
- Lea, N. C., Nicholls, J., Dobbs, C., Sethi, N., Cunningham, J., Ainsworth, J., Heaven, M., Peacock, T., Peacock, A., and Jones, K. (2016). Data safe havens and trust: Toward a common understanding of trusted research platforms for governing secure and ethical health research. *JMIR Medical Informatics*, 4(2).
- Lewis, J. R. and Sauro, J. (2009). The factor structure of the system usability scale. In *International conference on human centered design*, pages 94–103. Springer.
- Li, Z., Wen, J., Zhang, X., Wu, C., Li, Z., and Liu, L. (2012). Clindata express - a metadata driven clinical research data management system for secondary use of clinical data. In *AMIA Annual Symposium Proceedings*, volume 2012, page 552. American Medical Informatics Association, American Medical Informatics Association.
- Lippeveld, T., Sauerborn, R., and Bodart, C. (2000). *Design and implementation of health information systems*. Citeseer.
- Livne, O. E., Schultz, N. D., and Narus, S. P. (2011). Federated querying architecture with clinical and translational health it application. *Journal of medical systems*, 35(5):1211–1224.
- Lyons, R. A., Jones, K. H., John, G., Brooks, C. J., Verplancke, J.-P., Ford, D. V., Brown, G., and Leake, K. (2009a). The sail databank: linking multiple health and social care datasets. *BMC medical informatics and decision making*, 9(1):3.
- Lyons, R. A., Jones, K. H., John, G., Brooks, C. J., Verplancke, J.-P., Ford, D. V., Brown, G., and Leake, K. (2009b). The sail databank: linking multiple health and social care datasets. *BMC medical informatics and decision making*, 9(1):3.
- Marcelo, A. (2010). Health information systems: a survey of frameworks for developing countries. *IMIA Yearbook*, pages 25–29.
- Markatos, E. P. (2001). On caching search engine query results. *Computer Communications*, 24(2):137–143.
- MarkLogic (2016). Data lakes, data hubs, federation: Which one is best? <http://www.marklogic.com/blog/data-lakes-data-hubs-federation-one-best>. Accessed: 2017-05-31.
- Martins, A. I., Rosa, A. F., Queirós, A., Silva, A., and Rocha, N. P. (2015). European portuguese validation of the system usability scale (sus). *Procedia Computer Science*, 67:293–300.

- McCray, A. T. (1989). The umls semantic network. In *Proceedings. Symposium on Computer Applications in Medical Care*, pages 503–507. American Medical Informatics Association.
- McCray, A. T. (2003). An upper-level ontology for the biomedical domain. *International Journal of Genomics*, 4(1):80–84.
- McCray, A. T., Burgun, A., and Bodenreider, O. (2001). Aggregating umls semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, 84(0 1):216.
- McGinnis, J. M., Olsen, L., Goolsby, W. A., and Grossmann, C. (2011). *Clinical data as the basic staple of health learning: creating and protecting a public good: workshop summary*. National Academies Press.
- Morgan, D. L. (1997). *The focus group guidebook*, volume 1. Sage publications.
- Mulder, N. J., Akinola, R. O., Mazandu, G. K., and Rapanoel, H. (2014). Using biological networks to improve our understanding of infectious diseases. *Computational and structural biotechnology journal*, 11(18):1–10.
- NLM (2017). Active release - 2017aa release. [https://www.nlm.nih.gov/research/umls/knowledge\\_sources/metathesaurus/release/active\\_release.html#](https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/active_release.html#). Accessed:2017-05-05.
- NLM (2018). The umls semantic network. <https://semanticnetwork.nlm.nih.gov/>. Accessed: 2018-10-20.
- OBIBA (2017). Data harmonization with opal. <https://wiki.obiba.org/display/OPALDOC/Data+Harmonization+with+Opa>. Accessed: 2017-05-28.
- Pathak, J., Kiefer, R. C., and Chute, C. G. (2013). Using linked data for mining drug-drug interactions in electronic health records. *Studies in health technology and informatics*, 192:682.
- Pavis, S. and Morris, A. D. (2015). Unleashing the power of administrative health data: the scottish model. *Public Health Res Pract*, 25(4):e2541541.
- Rabiee, F. (2004). Focus-group interview and data analysis. *Proceedings of the nutrition society*, 63(4):655–660.
- Rahm, E. and Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13.
- Randall, S. M., Ferrante, A. M., Boyd, J. H., and Semmens, J. B. (2013). The effect of data cleaning on record linkage quality. *BMC medical informatics and decision making*, 13(1):64.
- Rector, A. L. (1999). Clinical terminology: why is it so hard? *Methods of information in medicine*, 38(4/5):239–252.



- Rector, A. L., Qamar, R., and Marley, T. (2009). Binding ontologies and coding systems to electronic health records and messages. *Applied Ontology*, 4(1):51–69.
- Regenstrief Institute, Inc (2017). The international standard for identifying health measurements, observations, and documents. <https://loinc.org>. Accessed: 2017-07-05.
- Reich, C., Ryan, P. B., Stang, P. E., and Rocca, M. (2012). Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. *Journal of biomedical informatics*, 45(4):689–696.
- Riaz, F. and Ali, K. M. (2011). Applications of graph theory in computer science. In *2011 Third International Conference on Computational Intelligence, Communication Systems and Networks*, pages 142–145. IEEE.
- Ribeiro, L., da Silva Cunha, J. P., and Correia, R. J. C. (2010). Information systems heterogeneity and interoperability inside hospitals-a survey. In *Healthinf*, pages 337–343.
- Robertson, D., Giunchiglia, F., Pavis, S., Turra, E., Bella, G., Elliot, E., Morris, A., Atkinson, M., McAllister, G., Manataki, A., et al. (2016). Healthcare data safe havens: towards a logical architecture and experiment automation. *The Journal of Engineering*, 2016(11):431–440.
- Rodrigues, J. M., Robinson, D. J., Della Mea, V., Campbell, J. R., Rector, A. L., Schulz, S., Brear, H., Üstün, B., Spackman, K. A., Chute, C. G., et al. (2015). Semantic alignment between icd-11 and snomed ct. In *MedInfo*, pages 790–794.
- Rodrigues, J. M., Schulz, S., Rector, A. L., Spackman, K. A., Millar, J., Campbell, J. R., Üstün, B., Chute, C. G., Solbrig, H. R., and Della Mea, V. (2014). Icd-11 and snomed ct common ontology: circulatory system. In *MIE*, pages 1043–1047.
- Rodrigues, J. M., Schulz, S., Rector, A. L., Spackman, K. A., Üstün, B., Chute, C. G., Della Mea, V., Millar, J., and Persson, K. B. (2013). Sharing ontology between icd 11 and snomed ct will enable seamless re-use and semantic interoperability. In *MedInfo*, pages 343–346.
- Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B., and Mark, R. G. (2011a). Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952.
- Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B., and Mark, R. G. (2011b). Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952.
- Saitwal, H., Qing, D., Jones, S., Bernstam, E. V., Chute, C. G., and Johnson, T. R. (2012). Cross-terminology mapping challenges: a demonstration using medication terminological systems. *Journal of biomedical informatics*, 45(4):613–625.

- Sauleau, E. A., Paumier, J.-P., and Buemi, A. (2005). Medical record linkage in health information systems by approximate string matching and clustering. *BMC medical informatics and decision making*, 5(1):32.
- Schmidt, G. and Ströhlein, T. (2012). *Relations and graphs: discrete mathematics for computer scientists*. Springer Science & Business Media.
- Schulz, S., Schober, D., Daniel, C., and Jaulent, M.-C. (2010). Bridging the semantics gap between terminologies, ontologies, and information models. In *Medinfo*, pages 1000–1004.
- Shahmoradi, L. and Habibi-Koolaei, M. (2016). Integration of health information systems to promote health. *Iranian Journal of Public Health*, 45(8):1096–1097.
- Sligo, J., Gauld, R., Roberts, V., and Villa, L. (2017). A literature review for large-scale health information system project planning, implementation and evaluation. *International Journal of Medical Informatics*, 97:86–97.
- Stearns, M. Q., Price, C., Spackman, K. A., and Wang, A. Y. (2001). Snomed clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association, American Medical Informatics Association.
- Stevanovic, D. (2014). *Spectral radius of graphs*. Academic Press.
- Stroetman, V., Kalra, D., Lewalle, P., Rector, A., Rodrigues, J., Stroetman, K., Surjan, G., Ustun, B., Virtanen, M., and Zanstra, P. (2009). Semantic interoperability for better health and safer healthcare [34 pages]. *Information Society and Media*.
- Stuart-Buttle, C., Read, J., Sanderson, H., and Sutton, Y. (1996). A language of health in action: Read codes, classifications and groupings. In *Proceedings of the AMIA Annual Fall Symposium*, page 75. American Medical Informatics Association.
- Tao, C., Pathak, J., Solbrig, H. R., Wei, W.-Q., and Chute, C. G. (2013). Terminology representation guidelines for biomedical ontologies in the semantic web notations. *Journal of biomedical informatics*, 46(1):128–138.
- Teodoroa, D., Paschea, E., Wipflia, R., Gobeilla, J., Choquetb, R., Danielb, C., Rucha, P., and Lovisa, C. (2009). Integration of biomedical data using federated databases. *Swiss Medical Informatics*, 25(67):57–60.
- Vicknair, C., Macias, M., Zhao, Z., Nan, X., Chen, Y., and Wilkins, D. (2010). A comparison of a graph database and a relational database: a data provenance perspective. In *Proceedings of the 48th annual Southeast regional conference*, pages 1–6.
- Vuokko, R., Mäkelä-Bengs, P., and Härkönen, M. (2014). Improving health care service delivery with a national code service. In *MIE*, pages 323–327.

- Waijen, S. A. (1997). Linking large administrative databases: a method for conducting emergency medical services cohort studies using existing data. *Academic emergency medicine*, 4(11):1087–1095.
- Weber, G. M. (2013). Federated queries of clinical data repositories: the sum of the parts does not equal the whole. *Journal of the American Medical Informatics Association*, 20(e1):e155–e161.
- Weber, G. M. (2015). Federated queries of clinical data repositories: Scaling to a national network. *Journal of biomedical informatics*, 55:231–236.
- Witham, M., Frost, H., McMurdo, M., Donnan, P., and McGilchrist, M. (2015a). Construction of a linked health and social care database resource - lessons on process, content and culture [published online ahead of print march 20, 2014 [published online ahead of print march 20, 2014]. *Inform Health Soc. Care*, 40(3):229–239.
- Witham, M. D., Frost, H., McMurdo, M., Donnan, P. T., and McGilchrist, M. (2015b). Construction of a linked health and social care database resource—lessons on process, content and culture. *Informatics for Health and Social Care*, 40(3):229–239.
- Zhang, J., Haider, S., Baran, J., Cros, A., Guberman, J. M., Hsu, J., Liang, Y., Yao, L., and Kasprzyk, A. (2011). Biomart: a data federation framework for large collaborative projects. *Database*, 2011.