



# Data Quality Issues in Electronic Health Records for Large-Scale Databases

By

ABDUL KADER SAIOD

Submitted in fulfilment of the requirements for the degree  
of **DOCTOR of PHILOSOPHY** to be awarded at the Nelson  
Mandela University

December 2019

Promoter: Prof Darelle van Greunen

# PUBLICATIONS

1. *Book Details (First)*: Handbook of Large-Scale Distributed Computing in Smart Healthcare. Published: 2017-08-08. Editors: Samee U. Khan, Albert Y. Zomaya and Assad Abbas.  
ISSN 2520-8632 ISSN 2364-9496 (electronic), Scalable Computing and Communications; ISBN 978-3-319-58279-5 IS BN 978-3-319-58280-1 (eBook), DOI: 10.1007/978-3-319-58280-1

*1.1. Chapter Details*: Part II - Data Quality and Large-Scale Machine Learning Models for Smart Healthcare.

*Chapter Title*: Electronic Health Records: Benefits and Challenges for Data Quality.

*Authors*: A.K. Saiod, D. van Greunen and A. Veldsman.

*Publisher*: © Springer International Publishing AG 2017;

S.U. Khan et al. (eds.), Handbook of Large-Scale Distributed Computing in Smart Healthcare, Scalable Computing and Communications, DOI: 10.1007/978-3-319-58280-1\_6

2. *Book Details (Second)*: Big Data Recommender Systems, Volume 1: Algorithms, Architectures, Big Data, Security and Trust. Editors: Osman Khalid, Samee U. Khan, Albert Y. Zomaya.

*Volume 1*: Published 18<sup>th</sup> July 2019, Hardback 337 pages, Product Code: PBPC035A, ISBN: 978-1-78561-975-5;

*Volume 2*: Published 18<sup>th</sup> July 2019, Hardback 487 pages, Product Code: PBPC035B, ISBN: 978-1-78561-977-9;

Also available as a 2-vol set; 978-1785619793;

*2.1. Chapter Details*: Chapter Five

*Chapter Title*: Novel Hybrid Approaches for Big Data Recommendations.

*Authors*: A.K. Saiod and D. van Greunen.

*Publisher*: The Institution of Engineering and Technology (IET)

UK, Head office, T: +44 (0)1438 313 311, E: [postmaster@theiet.org](mailto:postmaster@theiet.org)

3. *Conference Details (Third)*: ICICIS-2019: Proceedings of 4th International Conference on the Internet, Cyber Security and Information Systems 2019.

*3.1. Conference Paper Details:*

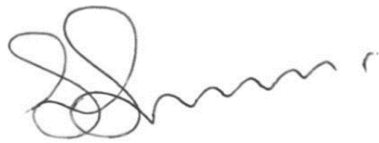
*Topic*: Cloud Integration for eHealth Data.

*Authors*: Abdul Kader Saïod and Darelle Van Greunen

*Publisher*: Kalpa Publications in Computing, Volume 12, 2019, Pages 300-309.

## DECLARATION

I, *ABDUL KADER SAIOD*, student number: s215127943 Nelson Mandela University, Port Elizabeth, South Africa, hereby declare that the thesis titled “*Data Quality Issues in Electronic Health Records for Large-Scale Databases*” for Doctor of Philosophy is my own work and that it has not previously been submitted for assessment or completion of any postgraduate qualification to another University or for another qualification. All information cited from published or unpublished works has been acknowledged.

A handwritten signature in black ink, appearing to read 'ABDUL KADER SAIOD', with a wavy, scribbled end.

ABDUL KADER SAIOD

## DEDICATION

This thesis is dedicated to my beloved parents, **Mrs. HASINA BEGUM** and **Md. NURUL HAQUE**. Thank you for all the sacrifices you made when we were growing up. Thank you very much as you always expected the best from me and putting pressure by making me your role model.

**Sisters** and **Brothers**, thank you for all your help, support, patience and encouragement.

## ACKNOWLEDGEMENT

All praise is due to **ALMIGHTY ALLAH**, the bestowal of knowledge and wisdom, for making it possible for me to complete this thesis.

I want to thank many individuals for their contribution to the completion of this thesis.

Dear **Prof Darelle Van Greunen**, I would not be where I am today if it were not you. Your guidance, mentorship and academic support over the years mean a lot to me. I am extremely grateful to you. I cannot even begin to explain how much your guidance and advice meant to me.

Special thanks to **Mrs. Visha Coopasamy**, who has shown me the way to my promoter.

To my wife, **Svitlana Saiod**, thank you for being my supporter and best friend. I would not wish for any other lifelong companion. To my lovely daughter, **Arina Saiod**, thank you for your patience over the years of the write-up. It was my biggest inability when you have asked me that you would also like to wake up at 02:00 am to see and communicate with me and were closing the door when I was preparing for the university lab over the weekends. I have missed how you have grown up, now we should have more time to 'go places' as a family, now that this is over.

**Family** and **Friends**, I want to thank everyone for your time, patience and encouragement. You are all greatly cherished and appreciated.

May Almighty **ALLAH** bless you!

## PREFACE (NOTE ON WRITING STYLE)

This thesis used the **United Kingdom (UK)** spelling. Words that could be spelled with either ‘s’ or ‘z’ are spelt with ‘s’. This is the case with a word such as ‘organisation’, ‘authorisation’ and ‘standardisation’. However, in spelling the names of organisations, the original spelling as used by the organisation concerned has been preserved. This is the case in the spelling of organisations, such as “the data quality standardization”, where the ‘z’ in standardisation is preserved.

This thesis used the Harvard Referencing Style. **In-text referencing:** This refers to citing references within the body or text of my work. For the in-text reference, I have indicated the following information:

- a)** Author’s surname;
- b)** Year of publication;
- c)** Page number (if available)

In addition, I have adopted the following writing style:

### Bulleted versus numbered lists

- In lists where I have explicitly stated the number of items in the list, I have used numbered lists.
- A list without a specific number of items is provided by using bullets.

### Writing of numbers as numerals versus words

- Numbers up to and including ten are written in words.
- Where a number is in a sentence, it is expressed as a numeral. For example, the number 55 is written as is.
- Where a sentence begins with a two-word number, it is written in words.

## ABSTRACT

Data Quality (DQ) in Electronic Health Records (EHRs) is one of the core functions that play a decisive role to improve the healthcare service quality. The DQ issues in EHRs are a noticeable trend to improve the introduction of an adaptive framework for interoperability and standards in Large-Scale Databases (LSDB) management systems. Therefore, large data communications are challenging in the traditional approaches to satisfy the needs of the consumers, as data is often not capture directly into the Database Management Systems (DBMS) in a seasonably enough fashion to enable their subsequent uses. In addition, large data plays a vital role in containing plenty of treasures for all the fields in the DBMS.

EHRs technology provides portfolio management systems that allow HealthCare Organisations (HCOs) to deliver a higher quality of care to their patients than that which is possible with paper-based records. EHRs are in high demand for HCOs to run their daily services as increasing numbers of huge datasets occur every day. Efficient EHR systems reduce the data redundancy as well as the system application failure and increase the possibility to draw all necessary reports.

However, one of the main challenges in developing efficient EHR systems is the inherent difficulty to coherently manage data from diverse heterogeneous sources. It is practically challenging to integrate diverse data into a global schema, which satisfies the need of users. The efficient management of EHR systems using an existing DBMS present challenges because of incompatibility and sometimes inconsistency of data structures. As a result, no common methodological approach is currently in existence to effectively solve every data integration problem.

The challenges of the DQ issue raised the need to find an efficient way to integrate large EHRs from diverse heterogeneous sources. To handle and align a large dataset efficiently, the hybrid algorithm method with the logical



combination of Fuzzy-Ontology along with a large-scale EHRs analysis platform has shown the results in term of improved accuracy.

This study investigated and addressed the raised DQ issues to interventions to overcome these barriers and challenges, including the provision of EHRs as they pertain to DQ and has combined features to search, extract, filter, clean and integrate data to ensure that users can coherently create new consistent data sets.

The study researched the design of a hybrid method based on Fuzzy-Ontology with performed mathematical simulations based on the Markov Chain Probability Model. The similarity measurement based on dynamic Hungarian algorithm was followed by the Design Science Research (DSR) methodology, which will increase the quality of service over HCOs in adaptive frameworks.

**Keywords:** Electronic Health Records (EHRs), EHRs data quality, Large-Scale EHRs, eHealth, EHRs standard, EHRs interoperability, Mobile health, Cloud Health, EHRs integration, Machine learning EHRs.

# TABLE OF CONTENTS

TITLE PAGE	i
PUBLICATIONS	ii
DECLARATION	iv
DEDICATION	v
ACKNOWLEDGEMENT	vi
PREFACE (NOTE ON WRITING STYLE)	vii
ABSTRACT	viii
TABLE OF CONTENT	x
LIST OF FIGURES	xxiv
LIST OF TABLES	xxix
LIST OF ACRONYMS	xxx
<b>CHAPTER ONE: Introduction</b>	<b>1</b>
1.1 Introduction	2
1.2 Background of the study	5
1.3 Research problem	6
1.3.1 Research statement	8
1.3.2 Research aim and objective	8
1.3.3 Research questions	9
1.3.4 Research scope and rationale	9

1.4 Research process and design	10
1.5 Research approach	12
1.6 Research outline	18
1.7 Research contribution	20
1.8 Ethical considerations	20
1.9 Summary	21
<b>CHAPTER TWO: Research design and methodology</b>	<b>23</b>
2.1 Introduction	24
2.1.1 Chapter background - research design and methodology	25
2.2 Aim and objectives of the survey and experiment	26
2.2.1 Research paradigms and philosophical assumptions in the interpretive framework	27
2.2.2 Overview of research paradigms and their philosophical assumptions	28
2.2.3 Research paradigm and philosophical assumptions applicable to this study	29
2.3 Research strategies	34
2.3.1 General overview of Design Science Research strategy	34
2.3.1.1 Design Science Research products	35
2.3.1.2 Design Science Research methodologies	39
2.3.1.2.1 Design Science Research objectives	43

2.3.1.2.2 Design Science Research artefacts	43
2.3.1.2.3 Design Science Research guidelines	44
2.3.1.2.4 Design Science Research process	45
2.3.1.2.5 Artefacts evaluation in Design Science Research	49
2.3.2 Application of the Design Science Research strategy as used in this study	51
2.3.2.1 Main cycle of the Design Science Research process	52
2.3.2.2 Research sub-cycles to develop the artefact	54
2.3.2.2.1 Sub-Cycle One	54
2.3.2.2.2 Sub-Cycle Two	55
2.3.2.2.3 Sub-Cycle Three	55
2.3.2.2.4 Sub-Cycle Four	56
2.4 Data collection methods	57
2.5 Data analysis	59
2.6 Summary	60

## **CHAPTER THREE: The EHRs landscape in Large Scale Databases**

**62**

3.1 Introduction	63
3.1.1 The Electronic Health Records landscape in Large Scale Databases	63
3.2 The Electronic Health Records landscape	65

3.2.1 The Electronic Health Records application area	65
3.2.2 Different Electronic Health Records Systems	67
3.2.3 Different Electronic Health Records Networks	69
3.2.3.1 Classification of Electronic Health Records systems	71
3.2.3.2 Mobile Electronic Health Records systems	71
3.2.3.3 Cloud Electronic Health Records systems	72
3.2.4 Existing Electronic Health Records terminology	75
3.3 Electronic Health	79
3.3.1 The Electronic Health Records	80
3.3.2 Electronic Health Records in Large Scale Databases	82
3.3.3 Concept of the Electronic Health Data Level	84
3.3.4 The Electronic Health Records data exchange	87
3.3.5 EHRs Data Quality and benefits	88
3.4 Electronic Health Records data structure	89
3.5 Electronic Health Records data synchronisation	91
3.6 Electronic Health Records data collection	95
3.6.1 Improving the Electronic Health Records data collection process	97
3.7 The barriers and threats of Electronic Health Records	100
3.8 Summary	103
3.9 Conclusion	104

<b>CHAPTER 4: The EHRs Data Quality Standards Landscape</b>	<b>106</b>
4.1 Introduction	107
4.1.1 The Electronic Health Records Data Quality Standards landscape	109
4.1.2 The Electronic Health Records Large-Scale Data Structure	109
4.2 Electronic Health Records Data Quality Standards	111
4.2.1 What are the Electronic Health Records Data Quality Standards?	112
4.2.2 Levels of the Electronic Health Records Data Quality Standards	116
4.2.3 The cost of low-level Electronic Health Records Data Quality Standards in the health domain	121
4.2.4 Benefits of high-level standards of Electronic Health Records in the healthcare domain	123
4.2.5 Implementation of Electronic Health Records Data Standards	125
4.2.6 The Electronic Health Records Data integration challenges	126
4.3 Why Data Quality is important for Electronic Health Records?	127
4.3.1 Why care about Data Quality in Electronic Health Records?	128
4.3.2 How to obtain Quality Data in Electronic Health Records?	129
4.3.2.1 Prevention of bad data in Electronic Health Records	129
4.3.2.2 Detection of bad data in Electronic Health Records	130
4.3.2.3 Repairing the bad data in Electronic Health Records	130
4.3.3 Allocating resources for the prevention, detection, and repair in Electronic Health Records	131

4.3.4 The Electronic Health Records process improvement	132
4.3.5 Training medical staff	132
4.4 The Electronic Health Records Data Quality control	132
4.5 The Electronic Health Records Data integration model	134
4.6 Summary	135
4.7 Conclusion	135
<b>CHAPTER FIVE: The EHRs critical issues and Data Quality challenges</b>	<b>137</b>
5.1 Introduction	138
5.1.1 The Electronic Health Records critical issues and Data Quality challenges	139
5.2 The critical issues in Electronic Health Records	139
5.2.1 What are the critical issues in the Electronic Health Records?	141
5.2.2 How the critical issues impact on Electronic Health Records?	142
5.2.3 Different critical issues in the Electronic Health Records	142
5.2.3.1 Data redundancy in the Electronic Health Records	144
5.2.3.2 System application failure in the Electronic Health Records	149
5.3 The Data Quality challenges in Electronic Health Records	150

5.3.1 Why the Data Quality is a challenge in Electronic Health Records	152
5.3.2 How does the Data Quality impact into the Electronic Health Records	153
5.3.3 Why Data Quality became a challenge in Electronic Health Records	154
5.3.4 Different Data Quality challenges in Electronic Health Records	156
5.3.4.1 Incompatible data structure in Electronic Health Records	157
5.3.4.2 Inconsistent data structure in Electronic Health Records	158
5.4 Important barriers and constraints in Electronic Health Records	162
5.5 The most meaningful association among heterogeneous data sources in Electronic Health Records	165
5.6 The integrity constraints in Electronic Health Records	166
5.7 The uncertainties in Electronic Health Record Systems integration	167
5.8 Data materialisation in the Electronic Health Record Systems integration	169
5.9 The query answering in Electronic Health Records	172
5.10 Different approaches to handle large-scale data	173
5.11 Summary	173
5.12 Conclusion	174



<b>CHAPTER SIX: Overview of approaches and recommendations for EHRs integration methods</b>	<b>176</b>
6.1 Introduction	177
6.1.1 Overview of approaches and recommendations for EHRs integration methods	179
6.2 EHRs Data integration method	179
6.3 Methods to detect and reduce data inconsistency	182
6.3.1 Rough Set theory	183
6.3.1.1 Rough Set basic philosophy	186
6.3.1.2 Indiscernibility	187
6.3.1.3 Rough Sets in data analysis	188
6.3.2 Logic analysis of inconsistency data	189
6.3.3 Functional dependencies corresponding relational variables	193
6.3.4 Fuzzy multi-attribute theory	197
6.3.5 Decision-making steps	199
6.3.6 The similarity measurement methods to detect and reduced data redundancy	202
6.3.7 The hybrid data integration method	206
6.4 Summary of the key lessons from the review of approaches and recommendations for EHRs integration methods	207
6.5 Conclusion	209

<b>CHAPTER SEVEN: Integration method for EHRs to address Data Quality issues</b>	<b>210</b>
7.1 Introduction	211
7.1.1 Overview of EHRs integration methods for LSDB	211
7.2 Background of the hybrid method based Fuzzy-Ontology for EHRs integration	213
7.3 Development phases of the hybrid method based on Fuzzy-Ontology for EHRs integration	214
7.4 Research phases	216
7.4.1 Phase 1 – Assessment (Determining interoperability objectives and priorities)	216
7.4.1.1 Identify the goals	217
7.4.1.2 Identify the purposes	218
7.4.1.3 Determining data and the quality needs	221
7.4.1.3.1 Identify the desired workflow	222
7.4.1.3.2 Expectation from the EHRs integration	223
7.4.2 Phase 2 – Strategic planning (Designing the framework to the modelling process)	226
7.4.2.1 Data profiling (Understanding data sources and associated quality)	228
7.4.2.2 Defining the gap between what data is available and the quality versus what the business logic	232
7.4.3 Phase 3 – Analysis (Pattern over time evolve the data integration architecture)	233

7.4.3.1 Data security and visibility	236
7.4.3.2 Performing a data quality evaluation	239
7.4.3.3 Revising the expectations and determining the selected data solution	242
7.4.4 Phase 4 – Implementation (Applying the framework to the modelling process)	244
7.4.4.1 Modelling the data stores necessary	245
7.4.4.2 Data extraction	246
7.4.4.3 Data transformation	247
7.4.4.4 Data loading	248
7.4.5 Phase 5 – Data validation (Extract the underlying DQ in EHRs for LSDB)	249
7.4.5.1 Identify the base DQ standard specified in the EHRs integration for LSDB	250
7.4.5.2 Classify the base DQ standards in EHRs	252
7.4.5.3 The resulting set of DQ in EHRs	253
7.5 Summary	254

## **CHAPTER EIGHT: The applicability of the EHRs integration method for the DQ issues**

8.1 Introduction	257
8.1.1 The applicability of the EHRs integration method for the Data Quality issues	260

8.2 EHRs integration using a hybrid method based on Fuzzy-Ontology	262
8.2.1 Background and overview of HM based on Fuzzy-Ontology	262
8.2.2 Developing the methodological approach for the Crisp Ontology	263
8.2.3 Developing the hybrid methodology for Fuzzy-Ontology	264
8.2.4 Extracting the EHRs key business functions for the proposed HIDM methodology	266
8.3 Identify the specification for the purpose of HIDM for EHRs	269
8.3.1 Phase One: The Purpose of the ontology scope logic	270
8.3.2 Phase Two: Identify Vagueness and Impreciseness to address the Fuzziness	270
8.3.3 Phase Three: Identify vagueness and impreciseness of the related data for Fuzziness	272
8.3.4 Phase Four: Survey applying the appropriate subsist ontology	273
8.3.5 Phase Five: Survey applying the appropriate subsist Fuzzy-Ontology	274
8.3.6 Phase Six: Fuzzy-Ontology components modification	274
8.3.7 Phase Seven: Fuzzy-Ontology component identification	276
8.3.8 Phase Eight: Crisp ontology component identification	276
8.3.9 Phase Nine: Formalisation of the model design	277
8.3.10 Phase Ten: Asseveration and notes	278
8.4 Real-Life Project: Hypertension diagnosis using HIDM based on Fuzzy-Ontology	278

8.4.1 Phase One: HIDA contrivance and excellence	282
8.4.2 Phase Two: Determine and ascertain the necessity for Fuzziness in the hypertension diagnosis	284
8.4.3 Phase Three: Identify the hypertension vagueness and impreciseness related data for Fuzziness	286
8.4.4 Phase Four: Reapplying the appropriate subsisting HIDM resources	287
8.4.5 Phase Five: Survey reapplying the appropriate subsisting Fuzzy-Ontology component resources	288
8.4.6 Phase Six: Appropriate the subsisting of Fuzzy-Ontology components	288
8.4.7 Phase Seven: Identify hypertension Fuzzy-Ontology components	288
8.4.8 Phase Eight: Identify hypertension crisp ontology components	292
8.4.9 Phase Nine: Formalisation of the model design	295
8.4.10 Phase Ten: Hypertension diagnosis result affirmation	296
8.4.11 Phase Eleven: Hypertension asseveration and notes	298
8.5 Mathematical simulation for the hypertension diagnosis based on the Markov Chain Probability Model	298
8.6 The perfect matching	305
8.6.1 The perfect matching analysis	311
8.7 The overall experiment result analysis	313
8.8 Summary	315

<b>CHAPTER NINE: Study contribution</b>	<b>317</b>
9.1 Introduction	318
9.1.1 The study contribution	318
9.2 Contribution to scientific knowledge	322
9.2.1 Contribution to address the DQ issues in EHRs for LSDB	322
9.2.2 Contribution to research aim, objectives and research questions	323
9.2.3 Achievement of the study objectives	324
9.2.4 Contribution to the understanding the DQ standards	327
9.2.5 Contribution to Design Science Research	328
9.3 Contribution to select the appropriate Data Integration Method to address the DQ in EHRs	330
9.3.1 The summary of investigations and findings	332
9.3.2 The limitations of the study	334
9.3.3 The key models contribution	335
9.4 A proposed model	338
9.5 The practical implications	339
9.6 Recommendations based on the results of the study	339
9.6.1 The Data Quality roles on quality care services for EHRs adaptation and interoperability	339
9.6.2 Implementing EHR Systems without considering DQ	339
9.7 The study reflection	340

9.7.1 Personal reflection	340
9.7.2 The methodological reflection	341
9.7.3 Scientific reflection	342
9.8 Recommendations for future research	343
9.9 Summary	345
9.10 The final contribution of the study	346

**REFERENCES** **347**

**APPENDIX A:** Fuzzy hypertension diagnosis using MATLAB

**APPENDIX B:** Mathematical simulation for hypertension diagnosis based on  
Markov Chain Probability Model

## LIST OF FIGURES

Figure 1.1: The patterns through a series of hypotheses “bottom-up approach”	13
Figure 1.2: The patterns through a series of hypotheses “top-down approach”	14
Figure 1.3: Design Science Research phases used in this thesis	16
Figure 1.4: Layout of the thesis chapters	17
Figure 2.1: Methodology architecture	29
Figure 2.2: Research paradigms and their philosophical assumptions	33
Figure 2.3: Design Science Research methodology process model	40
Figure 2.4: Research phases for this study	46
Figure 2.5: Design Science Research phases used in this thesis	50
Figure 3.1: The position of Chapter Three in the Design Science Research process used in this study	64
Figure 3.2: Large-scale cross-platform EHRs system architecture overview	70
Figure 3.3: Mobile eHealth Record Systems (MERS) architecture	72
Figure 3.4: Cloud Electronic Health Record systems	73
Figure 3.5: An overview of eHealth architecture	79
Figure 3.6: EHRs systems framework	80
Figure 3.7: The electronic health record systems in large-scale DBMS	83
Figure 3.8: EHRs information exchange systems architecture	87



Figure 3.9: EHRs database structure	90
Figure 3.10: Two-way data synchronisation workflow	92
Figure 3.11: A healthcare risk manager's organisation data synchronisation architecture	94
Figure 3.12: The data flow control system in large-scale DBMS	98
Figure 3.13: Outcome of Chapter Three	104
Figure 4.1: The position of Chapter Four in the design science research process used in this study	108
Figure 4.2 Large-scale advanced analytic platform architecture	110
Figure 4.3: High-level data quality process architecture framework	116
Figure 4.4: The EHRs Integration Model	134
Figure 4.5: The combined outcome of Chapter Three and Chapter Four	136
Figure 5.1: The position of Chapter Five in the Design Science Research process used in this study	140
Figure 5.2: Data Quality challenges and benefits	151
Figure 5.3: The EHRs Data Quality Framework architecture	153
Figure 5.4: Data Materialisation in Large-Scale EHRs System	170
Figure 5.5: The combined outcome of Chapter Three, Chapter Four and Chapter Five	174
Figure 6.1: The position of Chapter Six in the Design Science Research process used in this study	178
Figure 6.2: Workflow and architecture of the similarity detection service	202

Figure 6.3: The combined outcome of Chapter Three, Chapter Four, Chapter Five and Chapter Six	209
Figure 7.1: The position of Chapter Seven in the design science research process used in this study	212
Figure 7.2: The phases of the proposed hybrid method based Fuzzy-Ontology	215
Figure 7.3: Expectation and conceptual approach of the EHRs Integration	225
Figure 7.4: Strategic planning of the EHRs integration of the framework	227
Figure 7.5 Data profiling modelling process in EHRs	229
Figure 7.6: The EHRs gap analysis process diagram	233
Figure 7.7: The DQ process model in EHRs	234
Figure 7.8: The EHRs data security and visibility	237
Figure 7.9: Dimensions between DQ and data quality assessment	240
Figure 7.10: Data quality functions and characteristics	253
Figure 7.11: The combination outcome of Chapter Three, Chapter Four, Chapter Five, Chapter Six and Chapter Seven	255
Figure 8.1: The position of Chapter Eight in the Design Science Research process used in this study	259
Figure 8.2: Inputs inspiring to conceive the HIDM	266
Figure 8.3: The complete HIDM structure based on Fuzzy-Ontology for EHRs integration systems adopted according to	267
Figure 8.4: The structure of the proposed HM for EHR systems based on Fuzzy-Ontology	269
Figure 8.5: Fuzzy data types for hypertension diagnosis	292

Figure 8.6: The overall visualised the structure of the Fuzzy Hypertension specific ontology	296
Figure 8.7 – The Markov Chain Probability link structure of hypertension progression risk model	298
Figure 8.8: Graphical representation of the Markov Chain Probability link structure of hypertension progression risk when “BMI to BP = 0.35 and BMI to HR = 0.30”	303
Figure 8.9: A different matrix probability simulation according to “BMI to BP” transmission in hypertension diagnosis	305
Figure 8.10: Bipartite graph of the Hungarian algorithm	306
Figure 8.11: Matrix of edge weights	307
Figure 8.12: The combination outcome of Chapter Three, Chapter Four, Chapter Five, Chapter Six, Chapter Seven and Chapter eight	316
Figure 9.1: The position of Chapter Nine in the Design Science Research process used in this study	319
Figure 9.2: Design Science Research cycles adopted according to A.R. Hevner <i>et al.</i> (2013)	329
Figure 9.3 Proposed hybrid method Fuzzy-Ontology	331
Figure 9.4: Inputs inspiring to conceive the HIDM	335
Figure 9.5: The complete HIDM structure based on Fuzzy-Ontology for EHRs integration systems adopted according to	336
Figure 9.6: The overall visualised the structure of the Fuzzy Hypertension specific ontology	336
Figure 9.7: A different Matrix probability simulation according to “BMI to BP” transmission in hypertension diagnosis	337

Figure 9.8: The Perfect Matching result of the Hungarian algorithm	338
Figure 9.9: Structure of the proposed system – Hybrid Integration Development Methodology for EHRs based on Fuzzy-Ontology	338
Figure 9.10: The combination outcome of Chapter Three, Chapter Four, Chapter Five, Chapter Six, Chapter Seven, Chapter Eight and Chapter Nine	345

## LIST OF TABLES

Table 2.1: Design Science Research (DSR) guidelines	44
Table 4.1: Comparison between Systematic and Random	122
Table 8.1 According to various BMI category range around adult South African male's (18-60 years) distribution blood pressure (systolic and diastolic)	280
Table 8.2: According to various BMI category range around adult South African female's (18-60 years) distribution blood pressure (systolic and diastolic)	281
Table 8.3: Adult South African male's (18-60 years) distribution blood pressure "systolic and diastolic" where BMI is the risk factor	281
Table 8.4: Adult South African female's (18-60 years) distribution blood pressure "systolic and diastolic" where BMI is the risk factor	282
Table 8.5: Determination of fuzzy data type of the blood pressure level and its breakdown that ensures the fuzzy description logic	289
Table 8.6: Fuzzy data types determined and identified in the hypertension fuzzy description logic	289
Table 8.7: Determination of appropriate fuzzy concepts in the hypertension specific diagnosis	290
Table 8.8: The appropriate crisp ontology logics in the hypertension diagnosis	293
Table 8.9: The description of the hypertension-related dataset	294
Table 8.10: The hypertension Fuzzy dataset corresponding to feature with the numerical presentation	294

Table 8.11: A different matrix probability simulation according to “BMI to BP”  
transmission in hypertension diagnosis

## LIST OF ACRONYMS

AAMI	Association for the Advancement of Medical Instrumentation
ANSI	American National Standards Institute
ASC	Accredited Standards Committee
ASTM	American Society for Testing and Materials
BMI	Body Mass Index
BP	Blood pressure
CFDs	Conditional Functional Dependencies
COs	Commercial Organisations
CPOE	Computerised Physician Order Entry Systems
CPT	Current Procedural Terminology
CRM	Customer Relationship Management
DBMS	Database Management Systems
DHS	Diverse Heterogeneous Sources
DICOM	Digital Imaging and Communications in Medicine
DI	Data Integration
DIM	Data Integration Method
DQ	Data Quality
DQA	Data Quality Assessment
DSS	Decision Support System
DQSS	Data Quality Standards

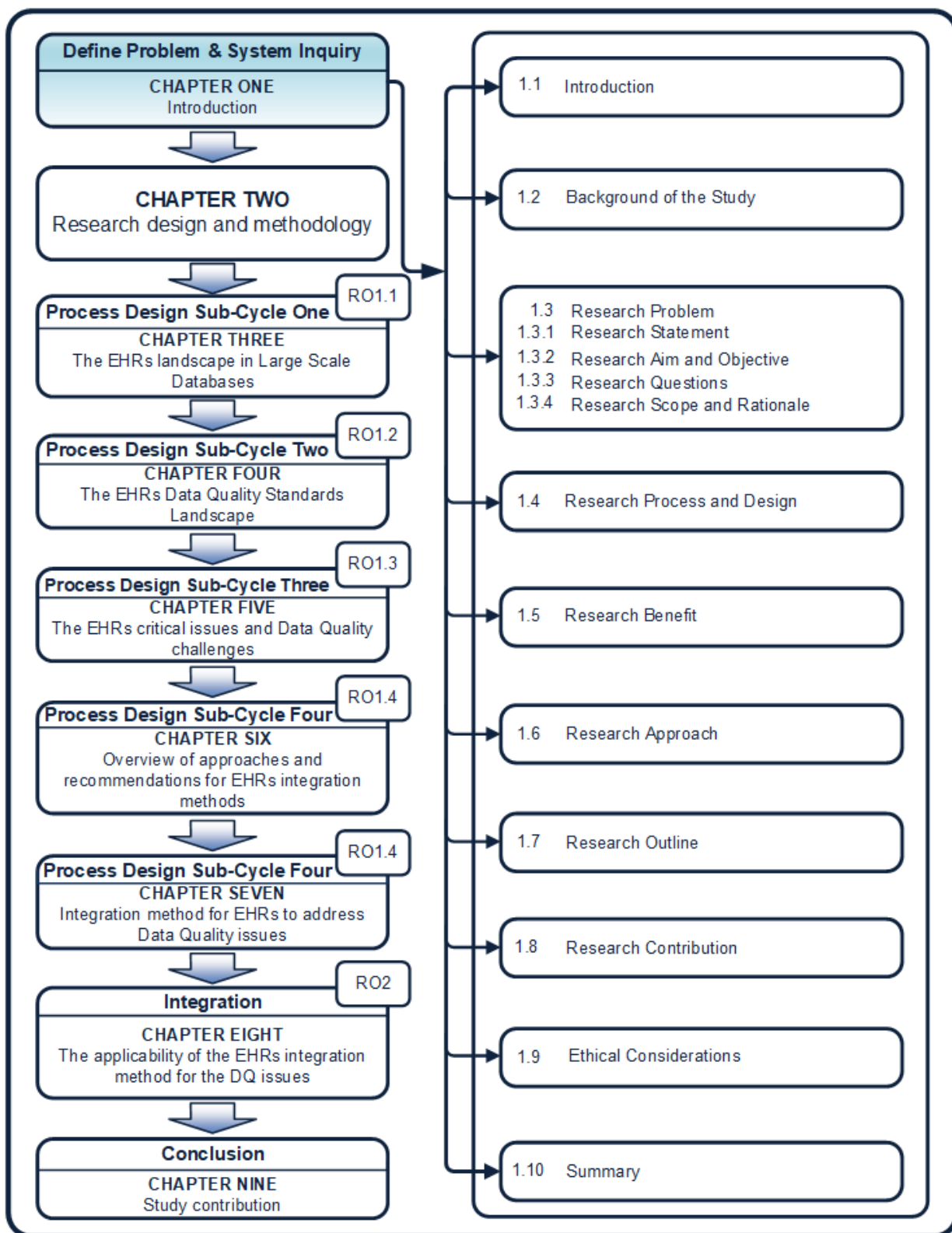
DR	Data Redundancy
DSR	Design Science Research
EDA	Exploratory Data Analysis
eHealth	Electronic Health
EHRs	Electronic Health Records
EHDL	Electronic Health Data Level
EMRs	Electronic Medical Records
ERP	Enterprise Resource Planning
FDCRVs	Functional Dependencies Corresponding Relational Variables
FOGA	Fuzzy-Ontology Generation Framework
PCP	Potentially Common Parts
HAI	Hybrid Approach Integration
HAs	Hybrid Approaches
HCOs	Healthcare Organisations
HCRM	Healthcare Risk Managers
HCSs	Healthcare Systems
HIDM	Hybrid Integration Development Methodology
HL7	Health Level Seven
HIPAA	Health Insurance Portability and Accountability Act
HISs	Health Information Systems
HIT	Health Information Technology



HR	Heart Rate
ICD	International Statistical Classification of Diseases and Related Health Problems
ICT	Information and Communication Technologies
IM	Integration Method
IMS	Information Management System
INR	Item Non-response
IS	Information Systems
ISDO	International Standard Development Organisations
IT	Information Technology
LSDB	Large Scale Databases
MADM	Multi-Attribute Decision-Making
MAGDM	Multi-Attribute Group Decision-Making
NAHIT	National Alliance for Health Information Technology
NCPDP	National Council for Prescription Drug Programs
NDC	National Drug Catalog
NIST	National Institute of Standards and Technology
OM	Operations Management
OR	Organisational Rules
PHP	Hypertext Preprocessor (Programming Language)
PIPS	Province Information Processing Standard
QAT	Quality Assessment Tools

QA	Question Answering
QIs	Quality Indicators
SDOs	Standards Development Organisations
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution
UNR	Unit Non-response

## CHAPTER ONE: Introduction



Outline of the Chapter One

# CHAPTER 1

## 1.1 Introduction

Electronic Health Records (EHRs) refer to implemented structured digital manifestations of real-time, patient-centred health records. EHRs are considered as one of healthcare’s innovative heuristic items and are widely adopted over Healthcare Organisations (HCOs) and becoming an important mechanism to perform healthcare daily services. But, improving Data Quality (DQ) to achieve benefits through EHR systems is neither low-cost nor easy. EHRs are in high demand for HCOs to run their daily services as increasing numbers of huge data occur every day.

However, one of the main challenges in EHRs is the inherent difficulty to coherently manage incompatible and sometimes inconsistent data structures from diverse heterogeneous sources. Secure EHR systems provide information available instantly and accurately to the authorised users; therefore, the user can coherently create new consistent data sets (Illhoi *et al.* 2012; Weng *et al.* 2013).

In general, the EHRs have the problem of combining health data that resides at different sources and providing an accurate, comprehensive up-to-date patient history (Terence 2015; Ibrahim 2016). Improvements in the DQ have brought about efficiency, scalability and safety in the implementation of large-scale healthcare Database Management Systems (DBMS) (Jens *et al.* 2012). Health data can, therefore, be composed and managed by the authorised user and consulted by authorised providers from across multiple HCOs nations or global wide and can be shared across them. The EHRs include an enormous range of patient data sets, including patient details, history, references, medication, immunisation, allergies and radiology reports including images, laboratory data, laboratory test reports, admission information, discharge details and personal statistics such as Body Mass

Index (BMI), blood pressure and sugar levels. These datasets are electronically stored in the database as narrative (free text) or encrypted data.

The EHR databases are structured to store health information accurately and securely over time. It can reduce data replication risk as all access points are retrieving data from the main data server, as well as lots of paperwork. The principle of data replication is to share information across multiple resources. The replication reduces fault tolerance, increase high accessibility and reliability. Many distributed database systems use replication to avoid single access point failure and high traffic. It can be possible to dynamically improve load-spreading and load-balancing performance by providing replication (Yeturu *et al.* 2016).

Replication supports restoring replicated databases to the same server and database from which the backup was created (Andrew *et al.* 2015). Backup is one of the important processes of database server routine maintenance plans that copies and archives data to an external device. In this way, backup data can be used to restore the original information after any data loss event.

Now a days, electronic data is searchable even from heterogeneous sources and it is possible to combine them into a single data set. EHRs are even more effective when analysing the longterm patient medical history (Gombert *et al.* 2015). Due to EHRs data being tractable and easy to identify patients preventive visits or screening information, the overall progress can be monitored, effectively than the paper-based record in HCOs. EHRs improve patient care, increase patient participation, improve care coordination, improve diagnostics and patient outcomes, practice efficiencies for cost savings and allow more case studies for research purposes. Despite the many advantages and functionalities of EHR systems, a considerable number of disadvantages are still associated with this technology (Brent 2014).

One of the key concerns is the quality of the data, which includes inconsistency, privacy protection and record synchronisation, lack of standardised terminology, system architecture indexing and deficient standardised terminologies. The productivity may drop temporally with the associated EHRs adaptation as

workflows have changed. Several long-standing consequences are emerging from the critical issue of EHRs adaptation (Andrea *et al.* 2005). Therefore, it is of the utmost importance to advise healthcare organisations to choose the correct EHR systems and to provide a proper setup to establish the complete system to become successful users of EHR systems (Nir *et al.* 2011).

Healthcare organisations using tangible augmented EHR systems in their facilities can make better decisions based on the comprehensive information available to them. Improving healthcare distribution systems are becoming the most consequential technology for medical innovation of all times. EHR systems exhibit promising potential, which has played the crucial role in HCOs to ensure the provision of excellent patient care service, quality management, accurate information, perfect diagnosis, patient safety information, disease management and investigation as advance innovation deftness (Sumit 2014).

Most of the existing EHRs integration methods such as peer-to-peer, data warehouses, middleware, data grid, data mining, semantics and ontology actually establish semantic connections between heterogeneous data sources. Although these methods offer advantages in some aspects, but they do not provide a coherent mechanism to solve every data integration problem (Christina *et al.* 2014; Yuchu *et al.* 2016; Beata 2017). In addition, none of them pays strong attention to data inconsistency, which has been a long-standing DQ challenge in health database environments (Malcolm *et al.* 2012; John 2013; Matthias *et al.* 2013). This implies that the integrated data can show inconsistency because different HCOs have several standards and different major systems, which have emerged as critical issues and practical challenges (Jiawei *et al.* 2011; Yusuf *et al.* 2012; Ralph *et al.* 2013).

This suggests that integrated adoptive EHRs can show inconsistencies because the data structure and standard from the various HCOs are different. Therefore, the principal consideration of this study is DQ issues in EHRs for Large Scale Databases (LSDB). Data quality introduces smart interfaces and perfect data mapping in present EHR systems as well as mobile and cloud computing.

## 1.2 Background of the study

In our daily lives, data inconsistency may cause uncertain incidents, if unstructured data is composed in the data collection process (Leonardo *et al.* 2015). For example, a web healthcare domain page base is largely composed of free text data. Another example is medical data collections process methods that are paper-based and/or from archived information.

Information may be collected by the data retrieval system and index them by non-text data so that the user can access and find data using special keywords to obtain accurate data sets (Moon *et al.* 2016).

Using the non-text data to index large text may lead the data structure design in EHR systems to the other efficient way for accessing and searching information as a vast amount of non-text data is available in EHRs (Wang *et al.* 2016).

The four latest methods to detect and reduce data inconsistency are:

- a)** Rough set theory (Ewa 2013);
- b)** Logic analysis of the inconsistent data method (Cheng 2014);
- c)** Corresponding relational variables of functional dependencies (Bernhard 2013);
- d)** Fuzzy multi-attribute theory (Abdolhadi *et al.* 2012);

The fuzzy multi-attribute method has the ideal performance of inconsistent data and it can obtain the highest average level of correct information compared to other solutions (Evangelos *et al.* 2013). A method for reducing data inconsistency has to be combined with a method for data integration to coherently solve the data inconsistency and the data integration problems simultaneously. The domain ontology may effectively combine data from diverse heterogeneous sources for data integration. The existing ontology data integration methods are, however, not sufficient to implement fuzzy-ontology (Hai *et al.* 2013).

The benefits of EHRs are numerous when compared to the physician's time and finances, the health benefits for patients and the impact on the environment. The sparse health data may have multi-dimensions and it is practically challenging to investigate and analyse for different reasons. These include the heterogeneous features of the system, encompassing quantitative data as well as the categorical information. This results in the random systematic error with affects the DQ badly and reduces it. Most data integration methods are sufficiently robust to random systematic errors for large datasets of input and process. This is commonly identical to bring them on the same scale when using the pre-processing principal component analysis and data simplification algorithm (Peter *et al.* 2012).

DQ issues in EHRs might include a patient incorrect unique identification number. Other examples include a misplaced name, incorrect gender, incorrect date of birth, numeric diagnosis code is written in text, or wrongly saved radiology image, incorrect inserting the standard code, such as the National Drug Catalog (NDC) for drugs and derailing bulk analysis (for example ICD10 code: International Classification of Diseases Tenth Revision or CPT code: Current Procedural Terminology). Data quality refers to the concepts with immensely large-scale multi-dimensional in DBMS, which include not only data search, validation, extract and verification, but also the appropriateness of use to take even further beyond the traditional concerns with the accuracy of data. The EHR system's design, data structure, aggregation algorithm, simplification methodology and reporting mechanisms highly reflect on the DQ.

### **1.3 Research problem**

Quality data, appropriate for use, comprise characteristics that include completeness, uniqueness, consistency, accuracy, validity, correctness and accurate timelines. DQ has emerged as a crucial issue in many health application domains (Hirak *et al.* 2016; Umut *et al.* 2017). The data flow process has several factors that influence the quality of information obtained from such data at a later stage. The objectives of EHRs become even more important in the case of merging



systems of different similar health organisations (Francky *et al.* 2013). The uncertainties are the other important integration aspect in EHRs that should be minimised to improve the DQ (Umberto *et al.* 2015). The most important barriers and constraints that hinder of the successful EHRs DQ should be identified to improve the implementation of Health Information Systems (HIS) and EHRs to achieve the maximum benefit in healthcare services.

Generally, two major barriers and challenges are in the way of the successful EHRs implementation, namely: the human barriers (for example, professional and belief) and the financial barriers (for example, available money and funding opportunity). The human factors become even more important as the benefits are only expected after the implementation and use of EHR systems (Mohamed 2013).

It improves the potential of EHRs as well as accuracy, accessibility, productivity, efficiency and to reduce the costs of healthcare including medical errors. The consolidation of information from diverse sources to provide a unified view of an organisation's data assets is technically challenging (Risto *et al.* 2011). This difficulty involves the way to practically combine data from disparate, incompatible, inconsistent and typically heterogeneous sources (Vladimir *et al.* 2015). The other difficult objective in EHRs is that data has a structure, which is usually complex and cannot be treated as a simple string of bytes (Roy 2012). Often data inconsistency occurs because the data structure may depend on other structures; therefore, in a distributed system, this kind of data management is very difficult (Alejandro *et al.* 2017).

Another important aspect of a data integration system is whether the system is able to materialise data which are retrieved from diverse sources through mappings (Matthias *et al.* 2013). The query answering in the context of data exchange is the final important issue for DQ (Christoph *et al.* 2014).

### 1.3.1 Research statement

***The eradication of data quality issues in electronic health records will benefit the integration of electronic health records and systems in large scale databases.***

### 1.3.2 Research aim and objectives

The overarching aim of the study is to tackle the indigent EHR's data quality to provide a single, centralised and homogeneous interface for users to efficiently integrate data from diverse heterogeneous sources. The DQ can be analysed from multiple dimensions. A dimension is a DQ measurable property that represents some aspect of the data accuracy and consistency that can be used to guide the process of understanding the quality (Nuno *et al.* 2015). DQ may rise along from collecting raw data into EHR information systems.

One of the key challenges of healthcare services is extinguishing medication errors in the medication process, where those badly affect the quality of patient care and these errors could also lead to death. The Implementing of EHR systems can result in improved patient safety by reducing medical errors in hospitals. To achieve the aim of this study, the following research objectives have been considered:

- a)** To analyse the impact of EHRs formal concept analysis adaptation on the research productivity;
- b)** To examine research productivity using DQ conceptual clustering;
- c)** To examine research productivity using DQ generation;
- d)** To examine research productivity using traditional systems;
- e)** To design a model on EHRs adoption for the increase of research productivity in the Grid-File for multi-attribute search and semantic representation conversion;

### 1.3.3 Research questions

The main research problem to be addressed in this study is stated as follows:

*What measures can be introduced to eradicate data quality issues in electronic health records that will benefit the integration of electronic health records and systems in large scale databases?*

To effectively answer this single research question, the following open sub-research questions are considered:

- 1) What are the most meaningful associations among heterogeneous data sources that can be explored to improve EHRs data quality?
- 2) What kinds of integrity constraints are specified in the global schema of data mapping that can be explored to improve EHRs data quality?
- 3) What are the uncertainties in the data integration that when minimised, would result in an improved EHRs data quality?

### 1.3.4 Research scope and rationale

The reason for conducting this study is to research, investigate and address the data quality issues in Electronic Health Records for Large Scale Databases. The main contribution of this study is the improvement of a novel framework for an effective method for electronic health records to achieve its maximum benefits and reduce the Data Quality challenges across healthcare organisations. A consensus method has been applied to solve the matching conflicts in EHRs integration.

In practice, the applicability and interoperability of the EHRs implementation to address the DQ issue for LSDB, the study constructed as follows:

- a) Developed and implemented a Hybrid Integration Development Methodology (HIDM) based on Fuzzy-Ontology for hypertension diagnosis;
- b) Performed a mathematical simulation to measure hypertension risk probability based on the Markov Chain Probability Model;
- c) Provided a similarity measurement based on the Hungarian Algorithm matching tool by combining Potentially Common Parts (PCP) and consensus techniques;

The EHRs consist of the following essential steps to achieve the goal: the formal concept analysis, the conceptual clustering, the ontology generation, the Grid-File for multi-attribute search and semantic representation conversion. The growth of the Internet and ICT technologies had a large impact on modern healthcare service. A fundamental need is to design novel EHRs services that not only improve people's health and well-being but also extend beyond the individual towards the sustainability of our society. Introducing EHR systems in healthcare service can offer several benefits to HCOs and society. Under the proposed framework, a database was developed and its result has shown the method to be effective regarding accurate performances.

This method presents the outline of the main theoretical properties considered to tackle the theoretical and thereafter, the practical problems for both the qualitative and quantitative methodologies of the research. The systems have no limits though and can be modified to benefit its scalability.

## 1.4 Research process and design

This study was conducted within a quantitative paradigm and the target data was collected from several HCOs and medical aids. **Firstly**, the entire departments were included in this study. **Secondly**, it was experimental, based on the use of traditional methods without EHRs software/tools. For **the first phase**, it was

important to see the impact and necessity of the EHRs adoption on the research productivity and for **the second phase**, using the EHRs to improve the research productivity. Experiments, testing, modelling, simulation, as well as theorem proving, are more about the qualitative and positivist aspect (Creswell *et al.* 2011; Vaishnavi *et al.* 2013).

On the one hand, a data integration method was extracted from the existing literature for the EHRs adoption and validated within the theoretical framework proposed by Bland *et al.* (2005). On the other hand, for the EHRs impact, the mathematical model of a heuristic methodology based on the Markov Chain Probability Model and a combination of the perfect matching for the similarity measurement based on the Hungarian Algorithm for distributed concepts, unanimity techniques for inconsistency and conflict resolution performance were used to improve data quality. The probability of hypertension behaviour and the diagnosis were measured using the Markov chain algorithm and graph theory. A consensus was created between perfect matching and similarity measurement which resolved to unite data inconsistency, mismatches and the conflict ontology entity regarding diverse data sources.

Matching is a process of finding alignment between sets of correspondences with a semantic verification output of the matching process. This merging is a process of creating a new set of possibly overlapping data. However, the aim of this study is to determine the best illustrate object to find a semantically fundamental equivalent motive in EHR systems to address DQ issues in LSDB. This provides a strong theoretical and practical framework to work with heterogeneous, complex, conflicting and automatic consensus methods for EHRs. This means that each and every EHRs amalgamated data associated with a distinguishable adumbration characteristic. Therefore, conflict may happen on EHRs integration, if a diverse amalgamated data associated with the same apprehension in the diverse EHR systems.

The similarity measurement often discovered that those conflicting entities are approximately identical among EHRs entities. Especially, with a chronic health

circumstance, the EHRs statistic predicts individual development and effectuation, since EHRs adoption better meets the needs of the growing modern community. The main objective of this study is namely, designing EHRs strategy to address the DQ issue; this incorporates Design Science Research (DSR) or creation and design, as DSR focuses on the production of artefacts. DSR is a problem-solving strategy aimed at building and evaluating artefacts to address real-life situations (Hevner *et al.* 2010).

This has been done to ease data access, extract information, search mechanisms, synchronises and establish semantic connections, filter data and provide different levels of security, provide data inconsistency solutions, resolve equivalently matching or conflicting information in multiple entities, resolve queries and achieve data compression and automatic data integration simultaneously.

## 1.5 Research approach

This research reported in this study is based on the **Design Science Research (DSR)** paradigm. The DSR is a problem-solving paradigm aimed at creating innovative artefacts to address and better understand a given problem.

The DSR methodology by Vaishnavi *et al.* (2015) and Kuechler *et al.* (2008) was used to guide the research process. This methodology consists of the five following iterative phases:

- 1) Defining the problem;
- 2) Systematic Inquiry;
- 3) Process design;
- 4) Integration;
- 5) Conclusion;

In addition to these phases, the methodology also involves a process known as circumscription. During the process design or integration phase, circumscription allows a researcher to fill the potential knowledge gaps by iterating back to the

define problem phases. The process of circumscription also enables the knowledge contribution that is only possible through the process of artefact construction. During the phases of the Vaishnavi *et al.* (2015) and Kuechler *et al.* (2008) DSR methodology, the following cognitive reasoning is typically employed by a design science researcher:

**a) Abduction:** This involves the use of interfaces to explain observations. Abduction is used in the suggestion phases of the DSR process to propose a solution to define the problem as it is a form of logical inference. Figure 1.1 (researcher source) describes the patterns through a series of hypotheses “bottom-up approach”, as follows:

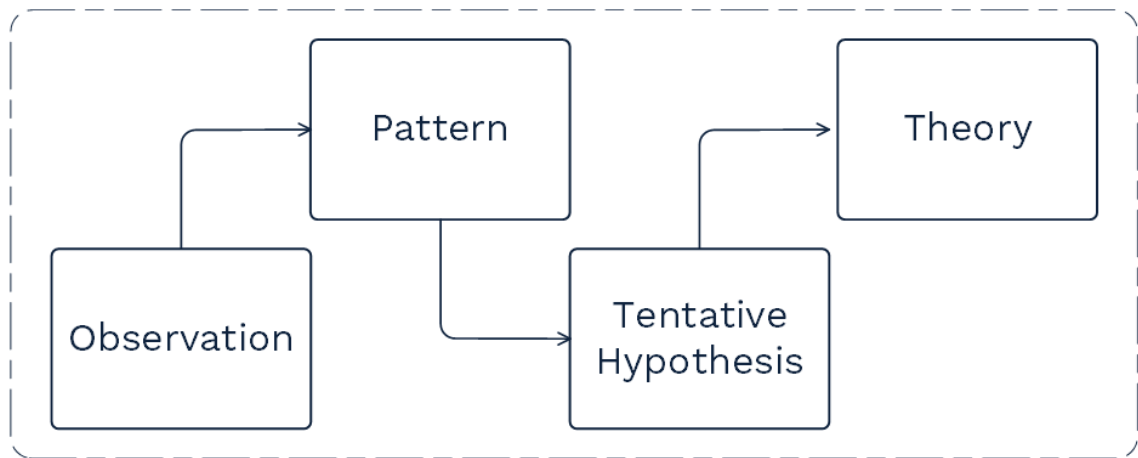


Figure 1.1: The patterns through a series of hypotheses “bottom-up approach” (researcher source)

This starts with an observation and then seeks to find the simplest and most likely explanation. The inductive approach, also known in inductive reasoning, starts with the observations and theories are proposed towards the end of the research process as a result of observations (Goddard *et al.* 2004).

Inductive research “involves the search for a pattern from observation and the development of explanations – theories – for those patterns through a series of hypotheses” (Bernard 2011). Inductive reasoning works the other way, moving from specific observations to broader generalisations and

theories. Informally, we sometimes call this a "bottom-up" approach (please note that it is "bottom up" and not "bottoms-up")

**b) Deduction:** This is the derivation of a specific hypothesis or theory from general theories. A deductive approach is concerned with “developing a hypothesis (or hypotheses) based on existing theory and then designing a research strategy to test the hypothesis” (Wilson 2010). Figure 1.2 (researcher source) describes the patterns through a series of hypotheses “top-down approach”, as follows:

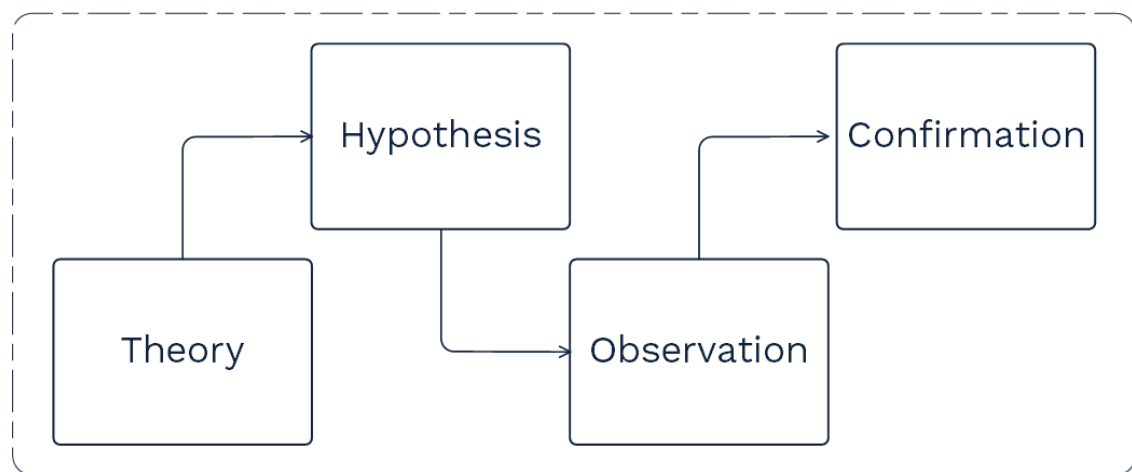


Figure 1.2: The patterns through a series of hypotheses “top-down approach” (researcher source)

Deductive reasoning is used during the process design and integration phases of the DSR process. Deductive reasoning works from the more general to the more specific. Sometimes this is informally called a "top-down" approach.

**c) Reflection and abstraction:** Reflection, in broad terms, refers to contemplating and learning from, experiences in the past. Together, these mental activities offer the potential to generate generic knowledge out of design practice (Shirley *et al.* 2013). This involves contemplation of the research process to determine the lessons learned from the research. Reflection and abstraction are used in the concluding phases of the DSR process and facilitate a contribution to the body of knowledge. As illustrated in figure 1.1, the research



reported in this thesis entails one main DSR cycle consisting of the following phases:

1. ***Defining the problem (Main DSR cycle):*** This phase defines the problem of the Data Quality issues associated with EHRs for LSDB.
2. ***Systematic inquiry (Main DSR cycle):*** This involves a systematic inquiry for the process design of a generic integration method that could use the integration of an appropriate set of integration.
3. ***Process design (Main DSR cycle):*** This phase involves the process design of addressing the Data Quality issues in EHRs for LSDB. This process design phases consist of four DSR sub-cycles, each with a defined problem, systematic inquiry and process design sub-phase:
  - a) ***Sub-cycle One:*** This sub-cycle involves the review of the EHRs landscape.
  - b) ***Sub-cycle Two:*** This sub-cycle involves the review of the EHRs used in the context of the healthcare domain.
  - c) ***Sub-cycle Three:*** This sub-cycle involves the critical issues and challenges associated with EHRs integration as well as adaptation and interoperability.
  - d) ***Sub-cycle Four:*** This sub-cycle involves the actual process design which is proposed.
4. ***Integration (Main DSR cycle):*** This phase involves a demonstration of the applicability of the proposed data integration method to address Data Quality issues in EHRs for LSDB.
5. ***Conclusion (Main DSR cycle):*** This phase involves reflections on the process design of addressing the Data Quality issues in EHRs to determine the study contributions to the body of knowledge.

The outcome of each sub-phase provides input for subsequent sub-phases. Following the fourth sub-cycle, the DSR processes iterate back the process design phase of the main DSR cycle. The integration and conclusion phases of the main DSR cycle are then conducted here.

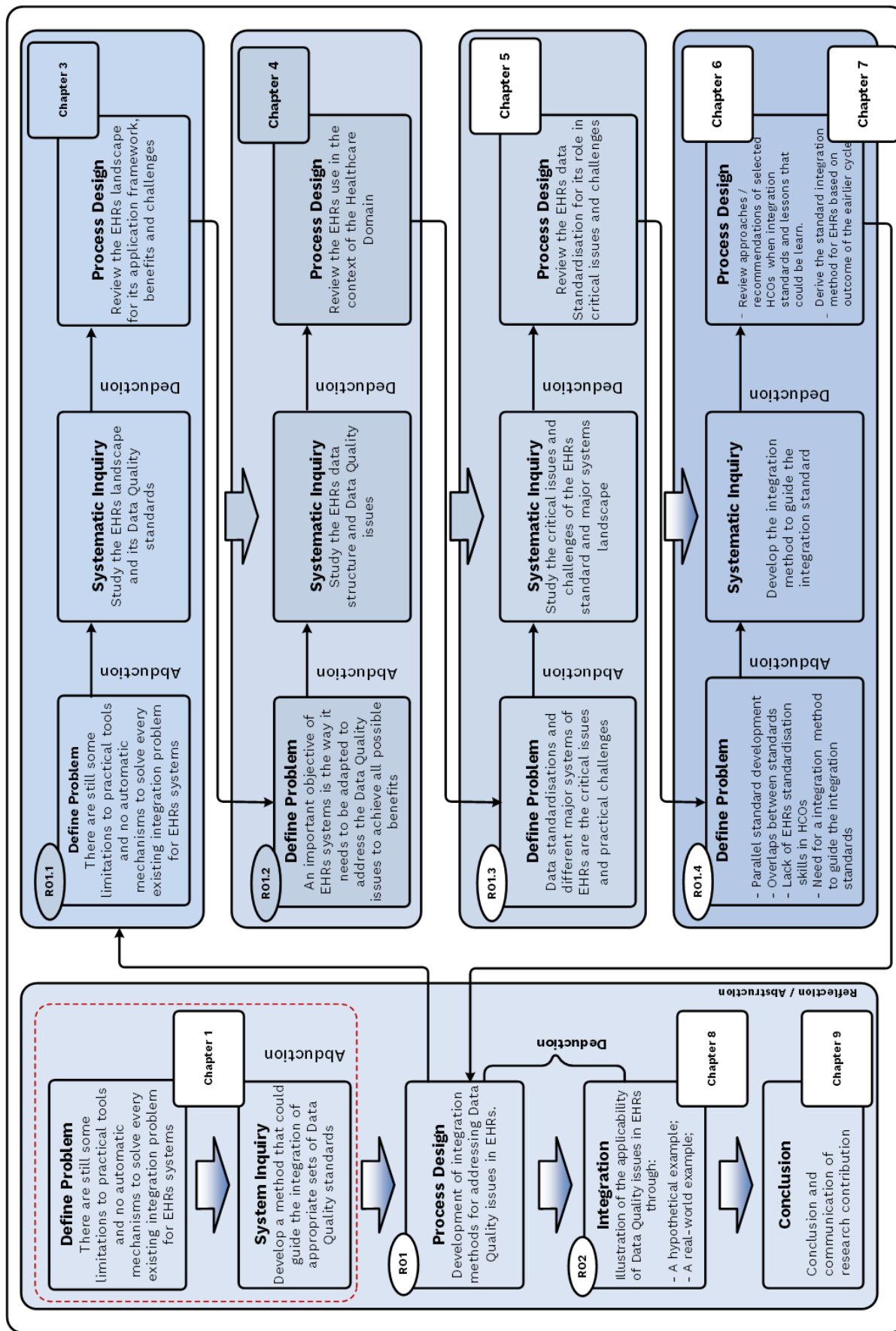


Figure 1.3: Design Science Research (DSR) phases used in this thesis

A detailed discussion of the research process is provided in Chapter two. Figure 1.2 describes the layout of the thesis chapters, as follows:

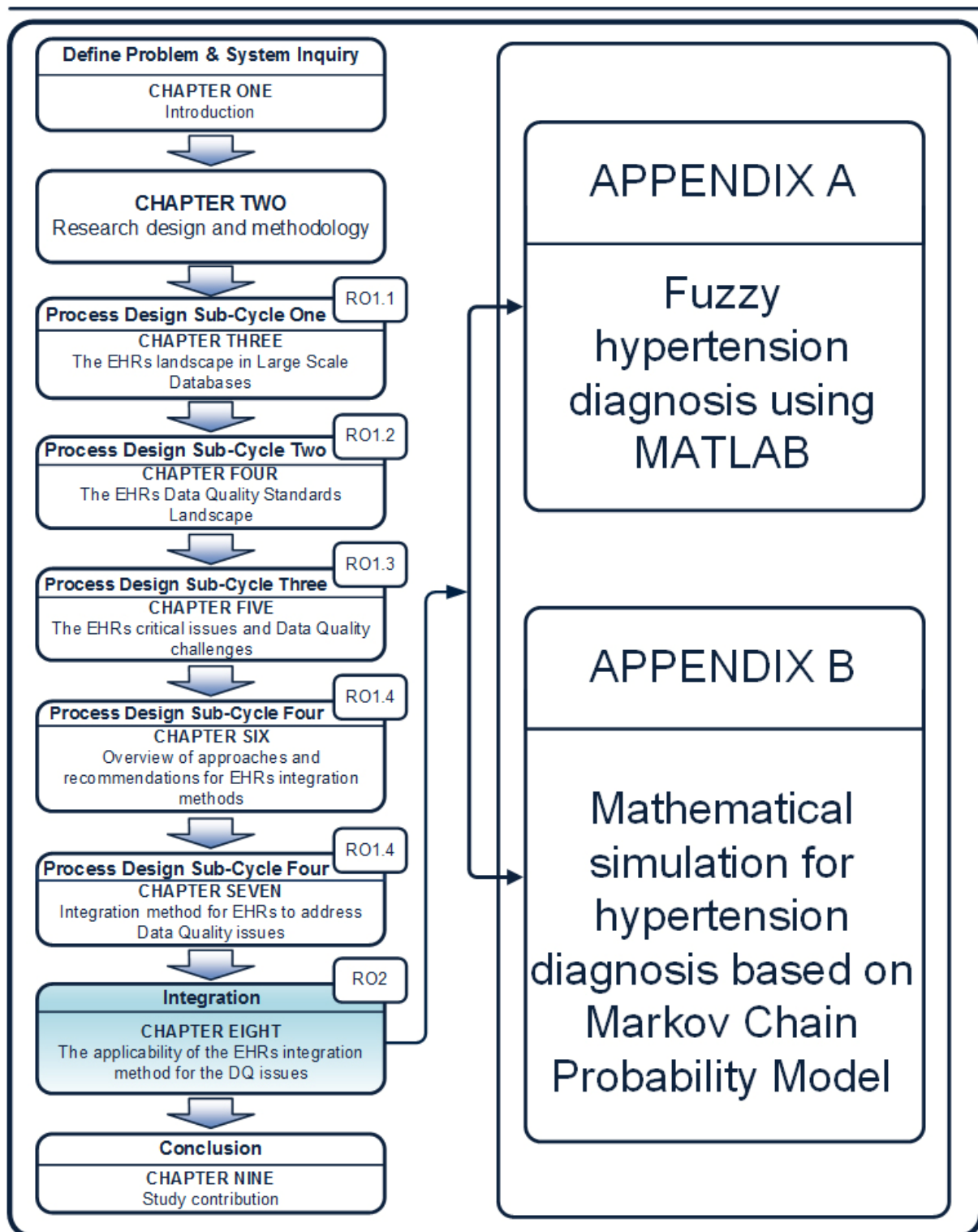


Figure 1.4: Layout of the thesis chapters

## 1.6 Research Outline

### ***CHAPTER ONE: Introduction - define problem and system inquiry***

Chapter One present the motivation for the thesis subject, research problem, aim, goal and objectives, the reason for it being necessary to work out new methods for EHRs to resolve the DQ issues for LSDB, a description of the solutions employed and an overview of EHR systems and ethical considerations.

### ***CHAPTER TWO: Research design and methodology***

Chapter Two offers the research paradigms and the philosophical assumptions, research strategies and data collection methods.

### ***CHAPTER THREE: The EHRs landscape in Large Scale Database***

The state of the art and the related works aimed at processing the distributed data in various forms, eHealth interoperability, standardisation, eHealth standardisation initiatives and services, are discussed in Chapter Three. The solutions having a fundamental influence on this thesis were briefly described.

### ***CHAPTER FOUR: The EHRs Data Quality standards landscape***

The fundamentals of EHR systems and challenges are given, based on the existing solutions. A summary of the main results presented in the literature for EHRs research is presented.

### ***CHAPTER FIVE: The EHRs critical issues and Data Quality challenges***

Chapter Five discusses the concepts and the assumptions of the developed and implemented methodology for automatic integrating from heterogeneous resources into a virtual repository. Furthermore, criticism of such systems is presented with respect to their effective capabilities in providing solutions for the DQ issues in EHR systems.

### ***CHAPTER SIX: Overview of approaches and recommendations for EHRs integration methods***

Assumptions about DQ issues of the testing process, as well as the results of testing, are presented in Chapter Six. Several algorithms for efficient consistent query answering have been presented to enable the system to deal with the expressive forms of integrity constraints. A series of experimental results are presented here.

### ***CHAPTER SEVEN: Integration method for EHRs to address Data Quality issues***

Chapter Seven shares the research contribution to the scientific body of knowledge, the solution to address the DQ issues, lack of interoperability, implementing eHealth world-class standards and Design Science Research (DSR).

### ***CHAPTER EIGHT: The applicability of the EHRs integration method for the DQ issues***

The detailed description of the development and implementation of the automatic integration mechanism is presented, which exploits EHRs on adaptation framework in Large Scale Databases (LSDB) to provide a robust and expressive solution to address the DQ issue. It contains architectures of the healthcare domain. Additionally, assumptions concerning a virtual network platform are explained in detail. Prototype activities are depicted by demonstrative examples based on introducing schemata. The chapter contains a number of listings, which have shown complex transformations of integration views and a presentation of the way that the view generation mechanism works. This chapter presents a Hybrid Integration Development Methodology (HIDM) based on the Fuzzy-Ontology development with a real-life project for hypertension diagnosis, performs a mathematical simulation based on Markov Chain Probability Method and a similarity measurement using the Hungarian method.

### ***CHAPTER NINE: Study contribution***

The conclusions and future works that can be conducted for the summary of the research findings, reflection and recommendation to further research of automatic EHR mechanism prototype development. The thesis text contains appendices describing DQ in EHRs for LSDB.

## 1.7 Research contribution

The contribution of the study presented in this thesis to the body of knowledge can be summarised as:

- a)* Addressing the problem fragmentation and the DQ issues in EHRs which could be used for Healthcare Organisations (HCOs) to deliver a higher quality of care to their patient than that which is possible with paper-based records;
- b)* Contributing to the number of DSR efforts that focuses on the method form of artefacts;
- c)* Contributing to the understanding of DQ issues in EHRs by the existing incompatibility and inconsistency data structure from heterogeneous sources;
- d)* Contributing to published documents on DQ issues in EHRs from academic and research institution;
- e)* Exploitability of outputs (for example, applicability to community development, improved products, processes, services, the region and/or the continent);
- f)* Expected effects of research results;

## 1.8 Ethical considerations

The systems have been designed in such a way that it is very unlikely for data and confidentiality to have been exposed. Confidential data has been securely stored on the external secured portable device – accessible only by the credentials. Researchable information has been stored for future consultation purposes for a period of 15 years. Irrelevant materials have been deleted or shredded upon completion of the research. During the recruitment process information has been being published in two books as a chapter detail provided on page ii.

For purposes of improvement in past researches and comparing the past and present (AS-IS model) to determined and contributed to model, certain

information might find in databanks that may have privacy legislation attached- which would require seeking access permission to those databanks. This particular research focused on setting up an African System with most of the African regional blocks with key stakeholders of this programme who have been involved in the setting up the ASAS project focused on Africa's requirements. Other stakeholders have been the project-owners and development partners.

Since this is a continental-tailored programme, employees, students and others will benefit. Employees will receive the knowledge and remuneration upon implementation and students will receive the research grounds to further their research; countries will benefit by using the services offered by the project. The research fulfills the criteria for informed consent. This research has been considered as competitive to other existing –similar systems established in developed regions of the world. Hence its success may be limited to various conditions that may include resistance to the idea, financial constraints by developed-linked countries and sabotage, for example.

## 1.9 Summary

This chapter has been presented as a research agenda to the study, to investigate and address the DQ issues in EHRs for LSDB. The main contribution of this study is the improvement of a novel framework for an effective method for electronic health records to achieve its maximum benefits and reduce the data quality challenges in healthcare organisations. A consensus method has been applied to solve the matching conflicts in EHRs integration. The main research question is: **What measures can be introduced to eradicate data quality issues in electronic health records that will benefit the integration of electronic health records and systems in large scale databases?** In practice, a Hybrid Integration Development Methodology (HIDM) was developed, based on Fuzzy-Ontology with a real-life challenge hypertension diagnosis ontology, a mathematical simulation was performed based on the Markov Chain Probability Method and a dynamic Hungarian algorithm matching tool has been implemented

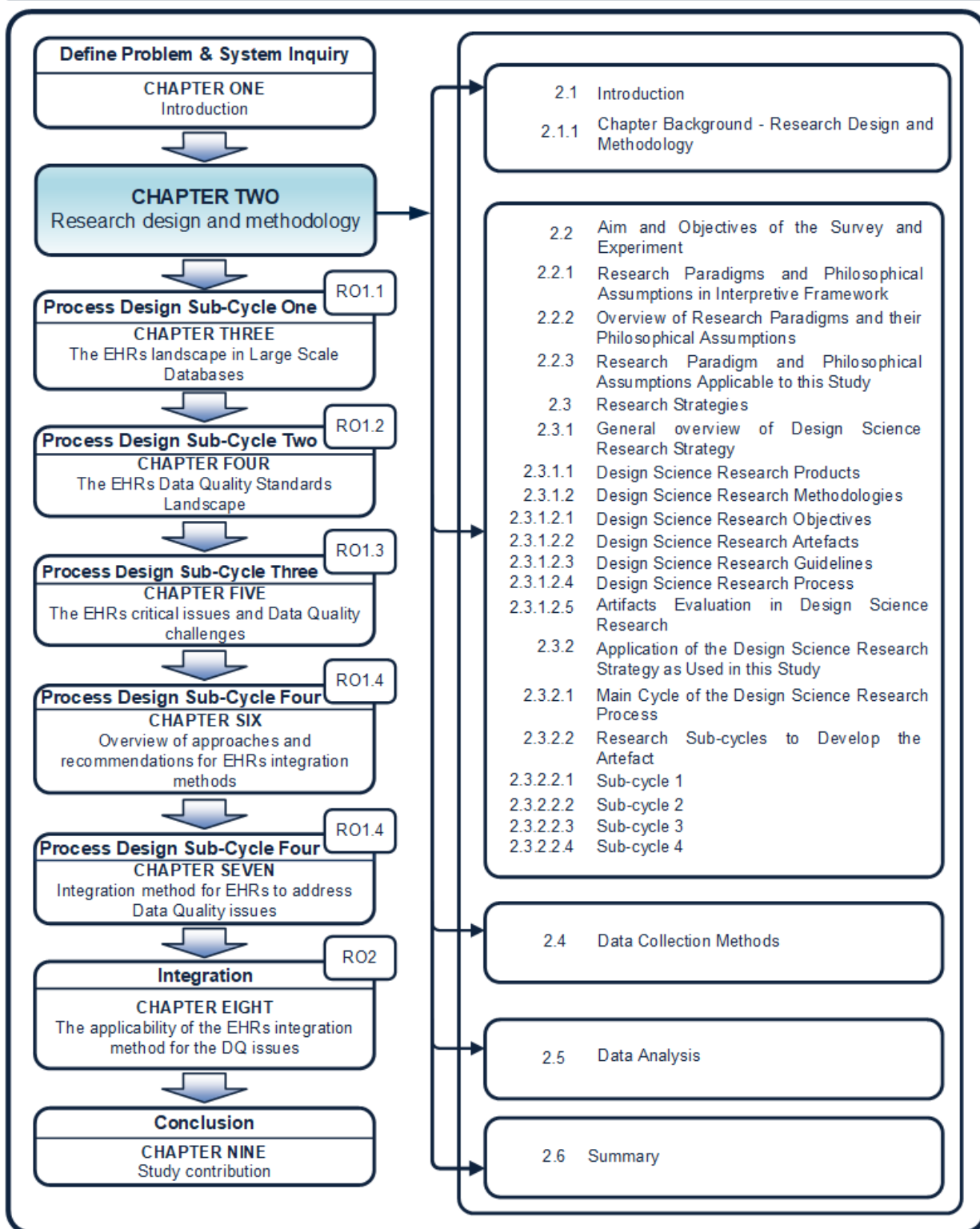
by combining Potentially Common Parts (PCP) and consensus techniques. The EHRs consist of the following essential steps to achieve the goal:

- 1) The formal concept analysis;
- 2) The conceptual clustering, the ontology generation;
- 3) The Grid-File for multi-attribute search;
- 4) Semantic representation conversion;

The growth of the Internet and ICT technologies had a large impact on modern healthcare service. A fundamental need is to design novel EHRs services that improve people's health and well-being, but also extend beyond the individual towards the sustainability of our society. Introducing EHR systems in healthcare service can offer several benefits to HCOs and society. Under the proposed framework, a database has been developed and its result has shown the method to be effective with regard to accuracy performance. This method presents the outline of the main theoretical properties considered to tackle the theoretical and thereafter, the practical problems for both the qualitative and quantitative methodologies of the research. The system does not have any limit and can be modified to benefit scalability.



## CHAPTER TWO: Research design and methodology



Outline of the Chapter Two

***“Theories are nets to catch what we call the ‘world’ to rationalise, to explain, and to master it. We endeavour to make the mesh ever finer and finer.”***  
– Sir Karl Popper

---

## CHAPTER TWO

### 2.1 Introduction

After having presented the research introduction, problem statement and the objectives, the purpose of this chapter is to share the philosophical assumptions underpinning this research, as well as to introduce the research strategy and the empirical techniques conducted in this study. Walliman *et al.* (2011) defined research as “an activity that involves finding out, in a more or less systematic way, things you did not know”. The chapter defines the scope and limitations of the research design and situates the research amongst existing research traditions in information systems. According to Ackoff *et al.* (1961), the term “Research” refers to a careful investigation or inquiry especially through searching for new facts in any branch of knowledge. In other words, research is an art of a scientific voyage of discovery investigation.

“All progress is born of inquiry. Doubt is often better than overconfidence, for it leads to inquiry and inquiry leads to the invention” is a famous Hudson Maxim in the context of which the significance of research can well be understood. This inquisitiveness is the mother of all knowledge and the method, which man employs for obtaining the knowledge of whatever the unknown, can be termed as research (Kothari 2004). Collected data were subjected to descriptive statistical and inferential statistical analysis, measurement modelling and structural equation modelling to provide the objectives formulated by this study.

The second part of the chapter presents the general research paradigms and philosophies and the way that this study can best be undertaken. This chapter does not describe the way in which the model ties up with the design of the survey or with that of the experiment, because these links are related to the findings and it is not appropriate to reveal such findings at this stage. In section 2.3, an overview of the general research strategies is presented, where section 2.4 discusses the

methods of data collection. Section 2.5 shares the method of data analysis. The chapter concludes with a summary in section 2.6.

## 2.1.1 The chapter background - Research design and methodology

The methodology is defined as a generic combination of methods commonly used as a whole – in soft systems methodology, strategic options development and analysis or survey methodology covers the design and analysis of questionnaires (Mingers 2003). Guba *et al.* (1998) stated that “the methodology is the way a researcher discovers whatever they believe can be known”, where Welment *et al.* (2003) regard, “it as the plan by which research participants are integrated and how the information is collected from them”. The philosophical assumptions underlying this research come from the interpretive tradition.

This methodology chapter systematically discovers, describe and motivates the research design implemented scientifically in this study to reach its aim and objectives. An outline of the systematic methodological approach that has followed to find appropriate answers to the research questions is discussed. **Firstly**, this is to provide a plan or mastermind for the research. **Secondly**, this should enable to anticipate the appropriate research design, to ensure the validity of the final results. Nevertheless, it is important that different views are analysed after, the methodology has been discussed. **Finally**, it is important to consider a theoretical framework for the research design.

This study has positioned a hybrid research method as the natural complement to traditional qualitative and quantitative research. It has been provided with a framework for designing and conducting hybrid research methods. This chapter covers the instrument design, target population, sample, data collection, data analysis, development and research analysis. This has been briefly described as the tent of pragmatism and the fundamental principle of the hybrid method and how to apply them. Compared to the mono-method, a key feature of the hybrid

methodological pluralism or eclecticism is that it is frequently superior to the research result.

## 2.2 Aim and objectives of the survey and experiment

The research methodology for the present study explains the research objectives and a suitable methodology to achieve those objectives. The objectives of this survey and experiment were to tackle the indigent DQ issues in EHRs for LSDB to provide a single, centralised and homogeneous interface for users to efficiently integrate data from diverse heterogeneous sources. The DQ can be analysed from multiple dimensions. A dimension is a DQ measurable property that represents some aspect of the data accuracy and consistency that can be used to guide the process of understanding the quality (Nuno *et al.* 2015). However, data accuracy and consistency should not be seen as isolated but need to be placed in the context of the research productivity. The assessment of the impact of data accuracy and consistency on the research productivity for EHRs was conducted in this present study, using hybrid research paradigms in Design Science Research (DSR) methodology.

This involved an exhaustive study of the investigation and addresses of the DQ issues associated with Electronic Health Records (EHRs) in Large Scale Databases (LSDB). The risk dimensions were explored and then of these dimensions were compared across the diverse health data from heterogeneous sources including defining and identifying characteristics of the quality data.

Secondly, the most meaningful among heterogeneous data, the integrity constraining in the global schema and uncertainties in the health data integration that is present in HCOs data domain, was associated. This was done by detecting the similarity measurement through a health data survey and comparing these across data inconsistency and uncertainties.

Thirdly, the moderating effect of EHRs characteristics factors was studied through research analysis. In addition, the study also assessed the impact of EHRs in HCOs on the success and the three performances construct of success, namely:

- a)* Finance;
- b)* Implement;
- c)* Quality separately;

This was followed by model validation through four case studies involving health data analysis of the projects. The research methodology has to be robust to minimise errors in data collection and analysis. Owing to this, various methodologies namely survey, interviews (telephonic, structured and unstructured) and case studies were chosen for data collection. This chapter describes the pilot of the study, participants of the study, instrumentation for the research, data collection and data analysis procedures of the entire study.

### **2.2.1 Research paradigms and philosophical assumptions in the interpretive framework**

“Methodology is the philosophical framework within which the research is conducted or the foundation upon which the research is based” (Brown 2006). O’Leary (2004) describes methodology as the framework which is associated with a particular set of paradigmatic assumptions that we have used to conduct our research. The research design is the arrangement of conceptual structure within which data is collected and analysed in such a way that aims to combine relevance to the research purpose with economy in procedure. The research design conducted constitutes the blueprint for the collection, measurement and analysis of the data. This section provided a brief description of different research paradigms and their philosophical assumptions, as well as the research paradigm and philosophy applicable to this thesis.

## 2.2.2 Overview of research paradigms and their philosophical assumptions

The research paradigm is defined as the first principle and/or ultimate handle by a set of basic beliefs (metaphysics) and agreements shared between scientists, about how problems should be understood and addressed (Kuhn 1962). Terre *et al.* (1999) state that “the research paradigm is a summary of a whole range of the all-encompassing system of interrelated practice, assumptions, research techniques, established results, methodologies and thinking that define the nature of enquiry along three dimensions”, as follows:

- a)** Ontology;
- b)** Epistemology;
- c)** Methodology;

According to Guba *et al.* (1994), “it represents a worldview that defines, for its holder, the nature of the world, the individual’s place in it and the range of possible relationships to that the world and its parts, as, for examples, cosmologies and theologies do”. Guba *et al.* (1994) made a significant contribution in articulating four differing worldviews of research, as follows:

- a)** Positivist;
- b)** Post-positivist;
- c)** Critical;
- d)** Constructivist-based on their ontological, epistemological and methodological assumptions;

Heron and Reason (1997) argue for a fifth worldview - a **participatory paradigm**. **Community-based research is situated within this paradigm and also embraces the ideology and methodology of co-operative inquiry** (Heron 1997 & 1996; Reason 1994 and 1988). In essence, research has been described as a systematic investigation (Burns 1997) or inquiry whereby data is collected, analysed and interpreted in some way in an effort to "understand, describe, predict or control

an educational or psychological phenomenon or to empower individuals in such contexts (Mertens 2005). Kuhn (1962) defines the research paradigm to be sufficiently unprecedented to attract an enduring group away from competing modes of scientific activity and sufficiently open-ended to leave all sorts of problems for the redefined group of practitioners to resolve. These are the so-called quantitative and qualitative research paradigms. In other words, the research paradigm of necessity has a preoccupation with theory and particularly with the philosophical trinity of theory.

### 2.2.3 Research paradigm and philosophical assumptions applicable to this study

Design Science Research (DSR) is a set of analytical techniques and perspectives for performing research in Information Systems (IS). Qualitative and quantitative approaches are rooted in philosophical traditions with different epistemological and ontological assumptions. Figure 2.1 describes the research methodology architecture applied in this study, as follows:

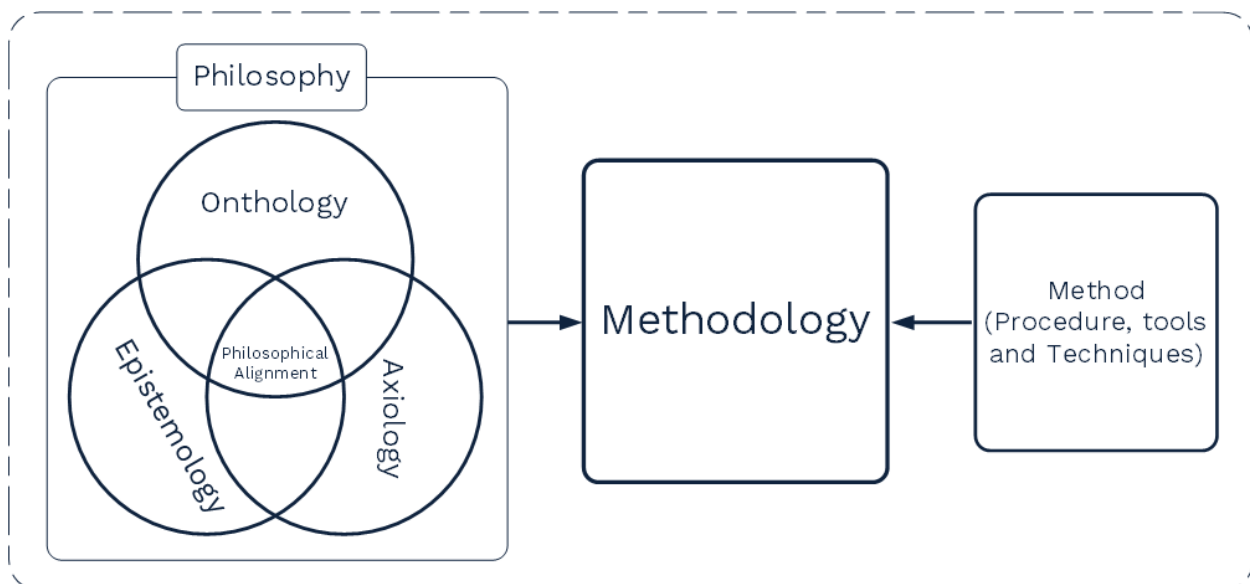


Figure 2.1: Methodology architecture (adapted according to Hevner *et al.* 2013)

Research methodology is a generic combination of methods commonly used as a whole – as in soft systems methodology, strategic options development and analysis or survey methodology covering the design and analysis of questionnaires (Mingers 2003). The quantitative research paradigm typically has an essentialist ontology, empiricist epistemology and either an Aristotelian or applied axiology. The qualitative research paradigm characteristically has an anti-foundationalist ontology, a realist or idealist epistemology and an applied or Aristotelian axiology. The realist epistemology in this paradigm gives rise to the constructivist research tradition and the idealist epistemology results in the subjectivist tradition of inquiry.

**Quantitative research:** Quantitative research is defined by Bryman *et al.* (2015) as “entailing the collection of numerical data and exhibiting the view of the relationship between theory and research as deductive, a predilection for natural science approach and as having an objectivist conception of social reality”. A type of research in which the researcher decides what to study asks specific, narrow questions, collects quantifiable data from participants, analyses these numbers using statistics and conducts the inquiry in an unbiased, objective way.

**Qualitative research:** Monette *et al.* (2005) credit qualitative methods with the acknowledgement of abstraction and generalisation. As defined by Polonsky *et al.* (2005), “the categorise vision, images, forms and structures in various media, as well as spoken and printed word and recorded sound into qualitative data collection methods”. A type of research in which the researcher relies on the views of participants asks broad, general questions, collects data consisting largely of words (or text) from participants, describes and analyses these words for themes and conducts the inquiry in a subjective, biased way. The fourth defining characteristics of a research paradigm axiology put in issues “values of being, about what human states are to be valued simply because of what they are” (Heron *et al.* 1997). The four philosophical assumptions are:

**a) Ontological assumptions:** The form and nature of reality and what can be known about it is called ontology (Guba *et al.* 1994) or the philosophy of the



world view or “Weltanschauung” of reality (Heron *et al.* 1997; Hitchins 1992) and it concerns the philosophy of existence and the assumptions and beliefs that we hold about the nature of being and existence. In contrast to orthodox research that utilises quantitative methods in its claim to be value-free (but which is more accurately described as valuing objectivity) and many qualitative approaches that value subjectivity, community-based research endorses a subjective-objective stance.

Subjective-objective ontology means that there is "underneath our literate abstraction, a deeply participatory relation to things and to the earth, felt reciprocity" (Abram 1996). As Heron *et al.* (1997) explain, this encounter is transactional and interactive. "To touch, see or hear something or someone does not tell us either about our self all on its own or about a being out there all on its own. It tells us about a being in a state of interrelation and co-presence with us. Our subjectivity feels the participation of what is there and is illuminated by it". So, community-based research is interested in investigating people's understandings and meanings as they experience them in the world.

**b) Epistemology assumptions:** Epistemology refers to the nature of the relationship between the knower and the can be known as well as the philosophy of knowledge and justification (Audi 2000). Epistemology is the theory of knowledge and the assumptions and beliefs that one has about the nature of knowledge. Guba *et al.* (1994) claim that orthodox science, because of its belief in a “real” world that can be known, requires the knower to adopt a posture of objective detachment in order "to discover how things really are".

There is a presumption that the knower and the known are separate and independent entities that do not influence one another. There is a search for the “truth”, for the facts in objective and quantifiable terms which hold empirical data in the highest esteem. In contrast, community-based research rests on an extended epistemology that endorses the primacy of practical knowledge. In community-based research, the knower participates in the known and that evidence is generated in at least four interdependent ways,

namely experiential, presentational, propositional and practical (Heron *et al.* 1997; Heron, 1996).

**c) *Axiology assumptions:*** Ontology and epistemology deal with truth, however, axiology is about values and ethics (Mingers 2003). Axiology is a branch of philosophy that studies judgments about the value (Saunders *et al.* 2012). Specifically, axiology is engaged with an assessment of the role of the researcher's own value on all stages of the research process (Li 2016). Axiology primarily refers to the "aims" of the research. This branch of the research philosophy attempts to clarify if one is trying to explain or predict the world or only seeking to understand it (Lee 2008).

The participatory paradigm addresses this axiological question in terms of human flourishing. Human flourishing is viewed as a "process of social participation in which there is a mutually enabling balance, within and between people, of autonomy, co-operation and hierarchy. It is conceived as interdependent with the flourishing of the planet ecosystem" (Heron 1996). Human flourishing is valued as intrinsically worthwhile and participatory decision-making and is seen as a means to an end", which enables people to be involved in the making of decisions, in every social context, which affect their flourishing in any way" (Heron 1996).

***Methodological assumptions:*** Research methods are the tools, techniques or processes, which we use in the research. These might be, for example, surveys, interviews, photo and voice or participant observation. Methods and how they are used are shaped by methodology. As illustrated the research methodology architecture in figure 2.1, the philosophical assumption and method influence the research methodology. One methodology that is particularly well suited to community-based research is the co-operative inquiry (Heron, 1996; Reason, 1994). The co-operative inquiry is a participatory action methodology that does research with people not on to or about them. This methodology engages people in a transformative process of change by cycling through several iterations of action

and reflection. Figure 2.2 demonstrates the details of research paradigms and their philosophical assumptions, as follows:

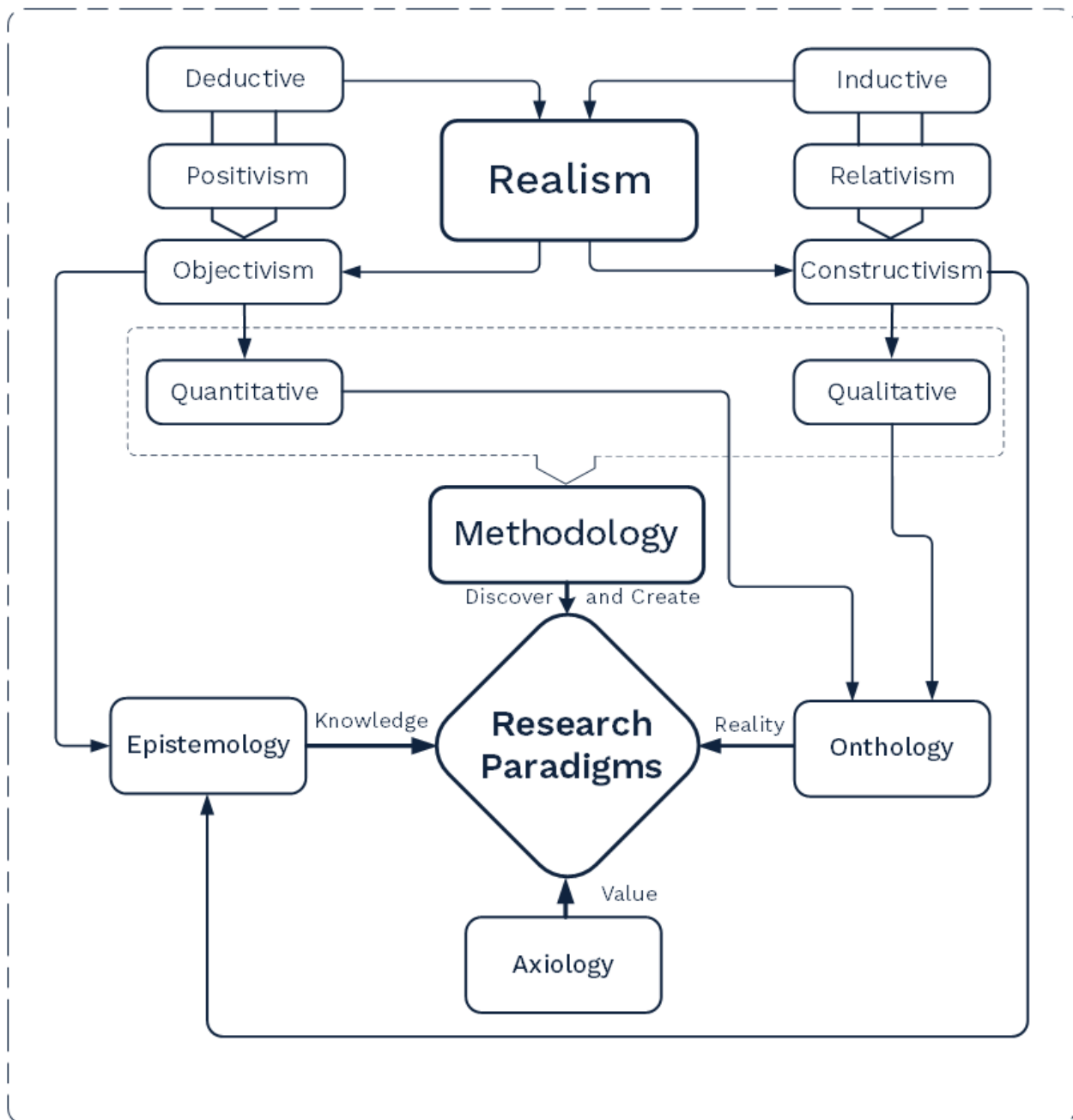


Figure 2.2: Research paradigms and their philosophical assumptions  
(researcher source)

Co-operative inquiry consists of a series of logical steps including identifying the issues/questions to be researched, developing an explicit model/framework for

practice, putting the model into practice and recording what happens, reflecting on the experience and making sense out of the whole venture (Reason 1988). Therefore, evidence about what constitutes “best practice” is generated by people examining their practices in practice and reflecting on these practices.

Paradigms are the models or frameworks derived from a worldview or belief system about the nature of knowledge and existence. Paradigms are shared by a scientific community and guide how a community of researchers acts with regard to the inquiry. Methodology defines how we gain knowledge about the world or "an articulated, theoretically informed approach to the production of data" (Ellen 1984).

## **2.3 Research strategies**

A research strategy is an overall plan for conducting a research study. A research strategy guides a researcher in planning, executing and monitoring the study. Whereas the research strategy provides useful support on a high level, it needs to be complemented with research methods that can guide the research work on a more detailed level. Research methods tell the researcher the way in which to collect and analyse data, for example, through interviews, questionnaires or statistical methods (Paul *et al.* 2014).

Section 2.3.1 provides a general overview of the DSR strategy, whereas section 2.3.2 gives a detail description of how DSR strategy was applied in this thesis.

### **2.3.1 General overview of the Design Science Research strategy**

The different types of artefacts that could emanate from DSR are presented in section 2.3.1.1, whereas 2.3.1.2 discusses the connective reasoning processes typically employed in DSR. Section 2.3.1.3 describes two commonly used DSR methodologies, with section 2.3.1.4 discussing the different evaluation methods for DSR artefacts.

### 2.3.1.1 Design Science Research products

Research is the production of knowledge to guide practice, with the modification of a given reality occurring as part of the research process itself. Design science products are assessed against the criteria of value or utility (Pertti 2007). Real problems must be properly conceptualised and represented, appropriate techniques for their solution must be constructed and solutions must be implemented and evaluated using appropriate criteria. The typical research product is the prescription discussed above or in terms of Bunge's technological rule: "an instruction to perform a finite number of acts in a given order and with a given aim". Hevner *et al.* (2004) state that much of the work performed by Information Systems (IS) practitioners deals with design. According to (March *et al.* 1995) design science products includes four types, as follows:

- 1) **Constructs:** Constructs provide the language in which problems and solutions are defined and communicated (Hevner *et al.* 2004). As stated by Winter (2008), foundational concepts correspond to constructs, means-end relations match with models and methods and concrete choices agree with instantiations.
  
- 2) **Models:** Models use constructs to represent a real-world situation, the design problem and its solution space (Hevner *et al.* 2004). Applicable ontology and metamodels constitute the foundation for theoretical statements. The design practice theory gives explicit prescriptions on how to design and develop an artefact by means of models and methods. For example, methodology providing models and methods, which clearly constitute a constructive design theory, is a demo (Dietz 2006). The demo is based on an explanatory design theory (called 'Performance in Social Interaction). Demo (Dietz 2006) is a methodology in the area of enterprise engineering that allows the revealing of the essence of an enterprise by means of supporting the construction of conceptual enterprise models. By

the essence is understood that the models completely abstract from all realisation and implementation issues.

**3) *Methods:*** Methods define the processes. They provide guidance on how to solve problems, that is, how to search the solution space (Hevner *et al.* 2004). **Further**, they clarify how an artefact is designed and developed by means of the given models and methods. Only models and methods which are founded in theory, either in the explanatory and/or predictive theory or in the explanatory design theory and which are sufficiently generalised to solve a class of design problems rather than a singular design problem, can be considered as constructive design theories.

**4) *Instantiations:*** Instantiations show that constructs, models or methods can be implemented in a working system (Hevner *et al.* 2004). Concrete design solutions are actually instantiations of those models and methods that have been applied for solution construction.

Over the years, five common paradigms of research paradigms are used in information systems research, which has been defined in the research, namely:

**1) *Assumptions and beliefs of the interpretivist paradigm:*** Interpretivist views have different origins in different disciplines. The interpretivist paradigm developed as a critique of positivism in the social sciences. In general, interpretivists share the following beliefs about the nature of knowledge and reality namely, interpretive approaches rely heavily on naturalistic methods (interviewing, observation and analysis of existing texts). These methods ensure an adequate dialogue between the researchers and those with whom they interact to collaboratively construct a meaningful reality. Generally, meanings are emergent from the research process. Typically, qualitative methods are used.

**2) *Assumptions and beliefs of the positivist paradigm:*** As a philosophy, positivism adheres to the view that only “factual” knowledge gained through

observation (the senses), including measurement, is trustworthy. In positivism studies, the role of the researcher is limited to data collection and interpretation through the objective approach and the research findings are usually observable and quantifiable. Positivist approaches rely heavily on experimental and manipulative methods. These ensure a distance between the subjective biases of the researcher and the objective reality he or she is studying. This generally involves hypothesis generation and testing. Typically, quantitative methods are used.

**3) *Assumptions of the critical or subtle realist paradigm:*** Critical realism is a meta-theory for social sciences. It is concerned with aspects of the philosophy of science, ontology, epistemology and axiology, along with conceptions of what constitutes an explanation, a prediction and what the objectives of social science ought to be. Critical or subtle realist paradigms have emerged recently and in the context of the debate about the validity of interpretive research methods and the need for appropriate criteria for evaluating qualitative research.

Realist approaches tend to rely on a combination of qualitative and quantitative methods. Research is conducted in more natural settings and more situational or contextual data is collected. It incorporates methods to elicit the participants' ways of knowing and seeing (interview, observation, text). Research designs provide opportunities for discovery (emergent knowledge) as opposed to operating by testing a priori hypothesis.

**4) *Assumptions of critical theory paradigms:*** These assume a "Reality" that is apprehendable. This is a reality created and shaped by social, political, cultural, economic, ethnic and gender-based forces, which have been reified or crystallised over time into social structures taken to be natural or real. People, including researchers, function under the assumption that for all practical purposes these structures are real. Critical theorists believe this assumption is inappropriate. Critical theoretical approaches tend to rely on dialogic methods,

methods combining observation and interviewing with approaches that foster conversation and reflection. This reflective dialogic allows the researcher and the participants to question the “natural” state and challenge the mechanisms for order maintenance. This is a way to reclaim conflict and tension.

**5) *Assumptions of feminist paradigms:*** Feminist concerns shape the research questions and interpretation, but researchers are committed to traditional research methods. Feminist empiricists adhere to the standards of current qualitative and quantitative methods. They believe that any method can be feminist. Their ontological and epistemological stance is similar to interpretivists or realists. Feminists use a wide range of research methods, including naturalistic approaches to social inquiry, quantitative and dialogic methods that combine observation and interviewing with approaches that foster conversation, reflection and change with regard to the “natural” and oppressive social order.

**6) *Design Science Research:*** Design science research (DSR) has staked its rightful ground as an important and legitimate IS research paradigm. Design science is an outcome-based information technology research methodology, which offers specific guidelines for evaluation and iteration within research projects.

Most qualitative research emerges from the “interpretivist” paradigm. While one describes the epistemological, ontological and methodological underpinnings of a variety of paradigms, one need not identify with a paradigm when doing qualitative research. Up until the 1960s, the “scientific method” was the predominant approach to social inquiry, with little attention given to qualitative approaches such as participant observation.

In response to this, a number of scholars across disciplines began to argue against the centrality of the scientific method. They argued that quantitative approaches might be appropriate for studying the physical and natural world, these were not



appropriate when the object of study was people. Qualitative approaches were better suited to social inquiry.

To understand the tension between paradigms one must understand that this tension - the either-or approach that emerged in the context of a debate about the capacity and importance of qualitative methods. Byrman and others, most recently Morgan (2007), argue for a more pragmatic approach, “one that is disentangled from the entrapments of this paradigm debate, one that recognises the ties or themes that connect quantitative and qualitative research and one that sees the benefits of blending quantitative and qualitative methods”. The brief details have been discussed below:

### **2.3.1.2 Design Science Research methodologies**

Design science research involves the design of novel or innovative artefacts and the analysis of the use and/or performance of such artefacts to improve and understand the behaviour of aspects of IS (Vaishnavi *et al.* 2015; Kuechler *et al.* 2008). In **Design Science Research**, as opposed to explanatory science research, academic research objectives are of a more pragmatic nature. Research in these disciplines can be seen as a quest for understanding and improving human performance (Aken 2005). Aken (2004) states that “there are two similar yet orthogonal essays groups addressing the subject of DSR”.

The DSR differs in some important aspects from other operations management (OM) research strategies; this essay examines in some depth its challenges and possible solutions (Aken *et al.* 2016). To underscore both the similarities among and distinctions between design-based research and related methods, design-based research is defined as a systematic but flexible methodology aimed to improve educational practices through iterative analysis, design, development and implementation.

This is based on collaboration among researchers and practitioners in real-world settings and leading to contextually-sensitive design principles and theories. Figure 2.3 demonstrates the DSR process model, as follows:

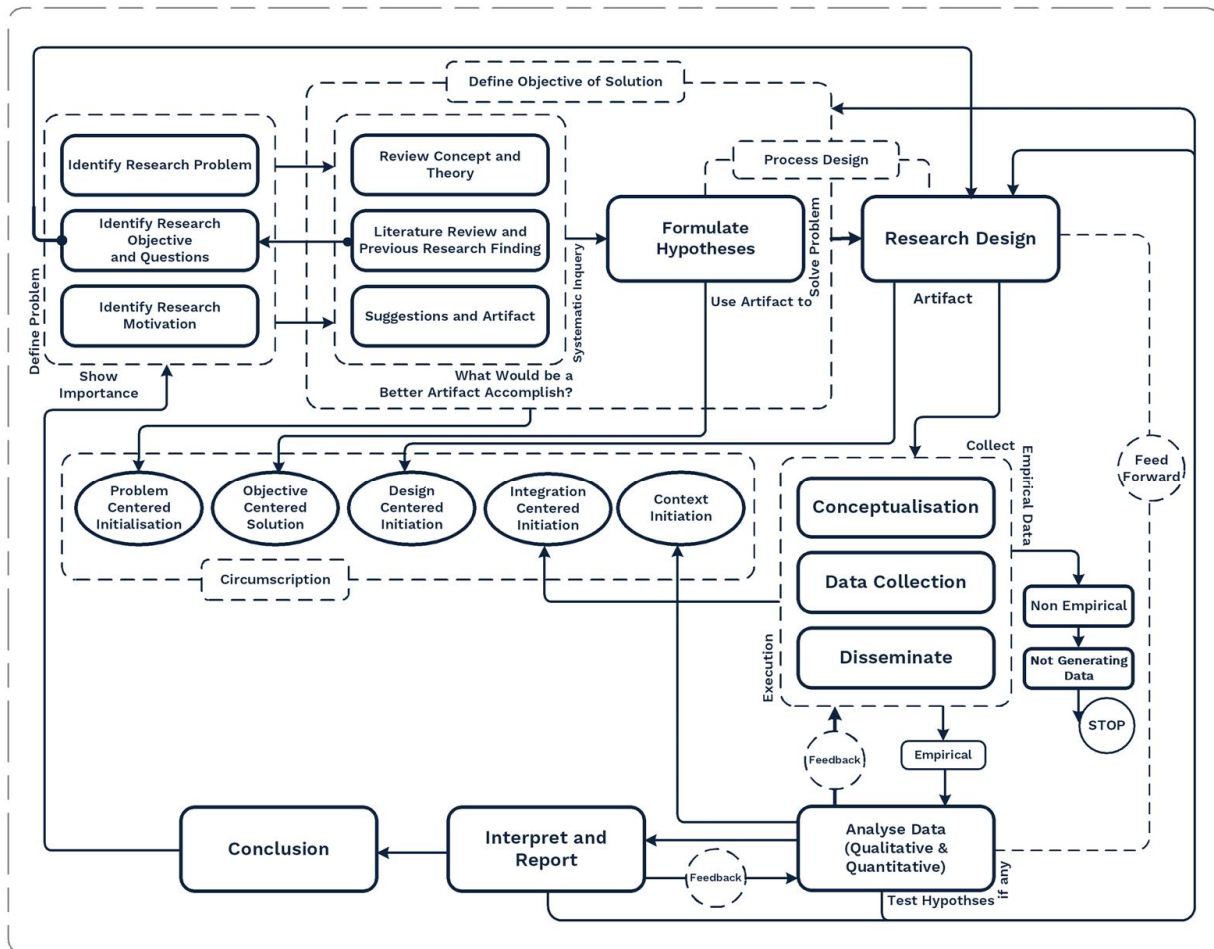


Figure 2.3: Design Science Research methodology process model (researcher source)

The five basic characteristics are, namely:

- a) Technological and empirical;
- b) Founded;
- c) Conversational, reiterative and comprehensive;
- d) Associative;
- e) Applicable;

The DSR has the inventive feature of new solutions for the new problem. True invention is a radical break through a clear departure from the accepted way of thinking and solving. The inventions are rare and inventors are rare still. The invention cannot be advanced as design theory before it was demonstrated in a physical artefact.

In many ways, DSR is intrinsically linked to and its development nourished by, multiple designs and research methodologies. Researchers assume the functions of both designers and researchers, drawing on procedures and methods from both fields, in the form of a hybrid methodology. For example, design-based research requires significant literature review and theory generation, uses formative evaluation as a research method and utilises many data collection and analysis methods widely used in quantitative or qualitative research (Orrill *et al.* 2003; Reigeluth *et al.* 1999).

In other ways, however, the convergence of design research, theory and practice extend beyond current methodologies. For example, participatory action research, a qualitative approach akin to design-based research involves collaboration between researchers and participants, local practices that support systematic theorising and improvement in both theory and practice.

However, local improvements in participatory action research typically derive from participants own research, which is facilitated by researchers rather than interventions designed and progressively refined jointly with researchers (Kemmis *et al.* 2000; Patton 2002; Stringer 1999). Design-based methodologies are especially important considering that EHRs have often been developed using incompatible or contradictory theoretical and epistemological foundations (Hannafin *et al.* 1997). Consequently, gaps are evident between what EHRs are and the way in which they should be used in theory compared to the way they are indeed used in practice.

Alternative approaches are needed to align learning environments with their fundamental assumptions (Hannafin *et al.* 1997; Jonassen *et al.* 1999) and “encourage flexibility as well” (Schwartz *et al.* 1999). Design-based research emphasises closely linked strategies for developing and refining theories rather

than testing intact theories using traditional methodologies (Edelson 2002). In this regard, design-based research does not replace other methodologies, but rather provides an alternative approach that emphasises direct, scalable and concurrent improvements in research, theory and practices.

Design-based research guides theory development improves the instructional design, extends the application of results and identifies new design possibilities (Cobb *et al.* 2003; Edelson 2002; Gustafson 2002; Reigeluth *et al.* 1999). Design-based research can “help create and extend knowledge about developing, enacting and sustaining innovative learning environments” (DBRC 2003). Richey *et al.* (2003) proposed two types of developmental research.

- a) Type One:** Research is context specific, conclusions typically take the form of lessons learned from the development of a specific product and conditions that improve the effectiveness of that product;
- b) Type Two:** Research, in contrast, yields generalisable design procedures or principles;

Likewise *et al.* (2002) identified three types of theories of potential relevance to EHRs, namely:

- 1) Domain theories;
- 2) Design frameworks;
- 3) Design methodologies;

Design methodologies are generic procedures that guide the process, such as the way in which to achieve a design goal and to develop the needed expertise. Both design frameworks and design methodologies are prescriptive in nature. Design-based research has the potential to generate theories that both meet teachers’ needs and support educational reforms (Reigeluth *et al.* 1999).

As with all disciplined inquiry, design-based research implementations need to be both purposeful and systemic. In this regards, design-based research parallels instructional design in many ways. Traditional Interaction Design (ID) activities are applied to address local design needs and requirements a goal shared by design-

based research. To generate practical, credible and contextual design theories, however, a rigorous, disciplined and iterative inquiry is needed.

Design-based research extends past the immediate local goal shared by traditional ID designers to generate pragmatic and generalisable design principles. Therefore, design activities and research activities usually cannot be conducted separately, systematic ID processes can be referred to as design-based research procedures. As described in the following sections, nine principles have been identified, which are central to planning and implementing EHRs design-based research.

### 2.3.1.2.1 Design Science Research objectives

According to Van Aken (2005), “the main goal of design science research is to develop knowledge that the professionals of the discipline in question can use to design solutions for their field problems. This mission can be compared to one of the ‘explanatory sciences’, such as the natural sciences and sociology, which is to develop knowledge to describe, explain and predict”. The DSR artefact is defined by Hevner *et al.* (2004) as, “the main purpose of DSR is achieving knowledge and understanding of a problem domain by building an application of a designed artefact”.

### 2.3.1.2.2 Design Science Research artefacts

Artefact design is a creative engineering process (Marne *et al.* 2017; Peter *et al.* 2015). Artefacts within DSR are perceived to be knowledge containing. This knowledge ranges from the design logic, construction methods and tool to assumptions about the context in which the artefact is intended to function. The creation and evaluation of artefacts thus form an important part in the DSR process which was described by Hevner *et al.* (2004) and supported by March *et al.* (2008) as, “revolving around **build and evaluate**”.

DSR artefacts can broadly include: models, methods, constructs, instantiations and design theories (March *et al.* 1995; March *et al.* 2008, Gregor *et al.* 2013), social innovations, new or previously unknown properties of technical/social or informational resources (March *et al.* 2008), new explanatory theories, new design and developments models and implementation processes or methods (Ellis *et al.* 2010).

### 2.3.1.2.3 Design Science Research guidelines

The design science research paradigm is generally comfortable with the world stage for the researchers as Hevner *et al.* (2004) counts seven guidelines summarised in table 2.1, for a design science research, as follows:

Table 2.1: Design Science Research (DSR) guidelines (Hevner *et al.* 2004)

Guideline	Description
<i>1. Design as an artefact</i>	DSR must produce a viable artefact in the form of a construct, a model, a method or an instantiation.
<i>2. Problem relevance</i>	The objective of DSR is to develop technology-based solutions to important and relevant business problems. DSR provides invention new solution for the new problem as well as to improve new solution for the known problem, explain known solution solutions extended to new problems and routine design to a known solution for the known problem.
<i>3. Design evaluation</i>	The utility, quality and efficacy of a design artefact must be rigorously demonstrated via well-executed evaluation methods.
<i>4. Research contributions</i>	Effective DSR must provide clear and verifiable contributions in the areas of the design artefact, design

	foundations and/or design methodologies. DSR is applying the knowledge contribution framework.
<i>5. Research rigour</i>	<i>Design Science Research</i> relies upon the application of rigorous methods in both the construction and evaluation of the design artefact.
<i>6. Design as a search process</i>	The search for an effective artefact requires utilising available means to reach desired ends while satisfying laws in the problem environment.
<i>7. Communication of research</i>	DSR must be presented effectively both to technology-oriented as well as management-oriented audiences.

#### 2.3.1.2.4 Design Science Research process

The DSR paradigm has been presented and motivated to be appropriate for this study. A research process “is the application of scientific method to the complex task of discovering answers (solutions) to questions (problems)” (Nunamaker *et al.* 1991). According to Leedy *et al.* (2001), “a research process is a systematic process of collecting and analysing information with the objective of increasing the understanding of the phenomenon under investigation to bring the solution”.

The research process combines qualitative and quantitative research and references well-known research methods for conducting design science. Therefore, the process is not a research method on its own, but a formalised combination of existing methods. In addition, it is believed that a defined process is fundamental to create new insights, as only then results from different projects become comparable. As the process is designed for design science, it provides support for other design artefacts as well (Philipp *et al.* 2009). The process also covers considerations and results from research using other theory types such as

explanation and prediction. The DSR process followed by this study is shown in figure 2.4, as follows:

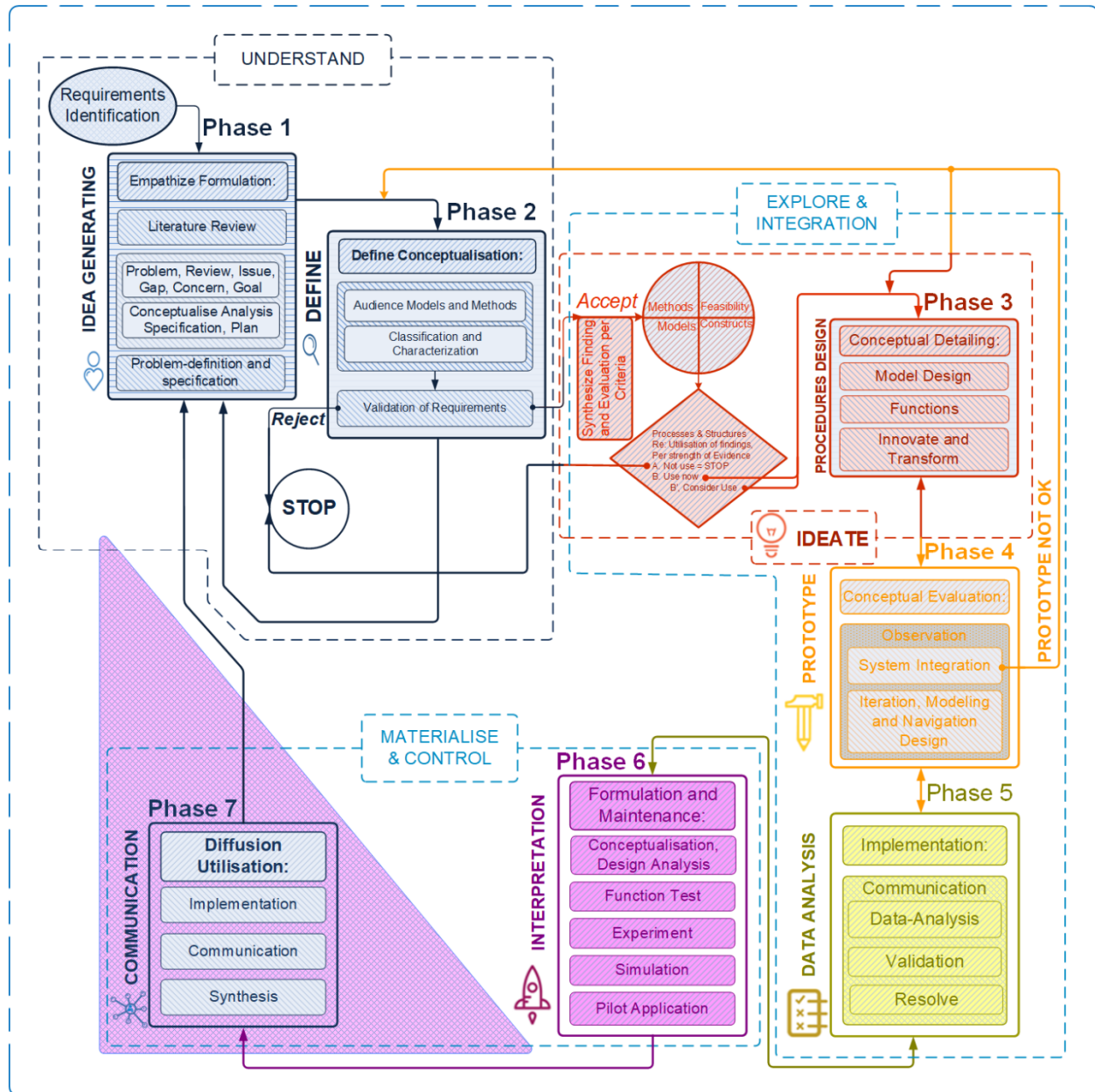


Figure 2.4: Research phases for this study

At the end of the research process, results are summarised and published (Wincup 2017). It combines different research methods used for qualitative and quantitative research. In contrast, seven iterative DSR phases are applied in this study and divided into three sections, which can interact with each other within the research



process and each phase is divided into a few steps. The arrows indicate a transition from one step to another dotted line indicating the research sections as follows:

**1. *Understanding:*** In the first phase of the research process, a problem is identified. It has to be ensured that the problem has practical relevance (Paul *et al.* 2017; Richard *et al.* 2016) or that it might be of relevance once solved. Criteria for problem relevance are reviewed in (Bardach *et al.* 2015).

**1.1. *Phase One (Idea Generating):*** Knowledge or idea exchange is an important function of groups in organisations. Numerous research has demonstrated that idea sharing in groups involves relatively inefficient processes. The participation at the moment of idea generation is an important place to be practising participatory design. This involves the definition of a specific research problem to address and justify the value of the proposed solution. However, “**participation at the moment of decision**” is gaining in interest as well. This conceptual analysis specification has fed into the process of creating an artefact to solve the identified problem.

**1.2. *Phase Two (Defining):*** Defining the specific research problem and justifying the value of a solution. The motivation stems from the plan to create a method that specialises in method engineering. The method factory has provided knowledge about construction, documentation, evaluation and configuration of methods to individual method construction. The defined process is fundamental to create new insights, as only then results from different projects become comparable. As the process is designed for design science, it provides support for other design artefacts as well.

**2. *Exploring and Integrating:*** This involves the procedure design, prototype and data analysis.

**2.1 *Ideating:*** This is the process of inferring the objectives of a solution from both the problem definition and the available knowledge of that which is possible and feasible.

**2.1.1. *Models/Methods:*** This is more about process and structure.

**2.1.2. Phase Three (procedure design):** This is a more formal model design, functions, innovate and transform.

**2.2. Phase Four (prototype):** This involves the conceptual evaluation which includes the observation that refers to system integration, iteration, and modelling and navigation design.

**2.3. Phase Five (data analysis):** This involves implementations which include communication as well as data-analysis, validation and resolving the issue.

**3. Materialising and Controlling:** This section is a combination of interpretation and communication.

**3.1. Phase Six (interpretation):** This involves formulation and maintenance, including the conceptualisation, design analysis, functional test, experiment, simulation and the pilot application

**3.2. Phase Seven (Communication):** This involves diffusion utilisation which includes the implementation, communication and synthesis.

Design activities are central to most applied disciplines. The DSR paradigm is highly relevant to Information System (IS) research because it directly addresses two of the key issues of the discipline namely the central, albeit controversial, role of the Information Technology (IT) artefact in IS research (Weber 1987; Orlikowski *et al.* 2001; Benbasat *et al.* 2003) and the perceived lack of professional relevance of IS research (Benbasat *et al.* 1999; Hirschheim *et al.* 2003). Design science, as conceptualised by Simon (1996), supports a pragmatic research paradigm that calls for the creation of innovative artefacts to solve real-world problems. DSR attempts to focus human creativity into the design and construction of artefacts that have utility in application environments. Design science offers an effective means of addressing the relevancy gap that has plagued academic research, particularly in the management and information systems disciplines.

One issue that must be clearly addressed in Design Science Research is differentiating high-quality professional design or system building from DSR. The difference is in the nature of the problems and solutions. Professional design is

the application of existing knowledge to organisational problems, such as constructing a financial or marketing information system using "best practice" artefacts (constructs, models, methods and instantiations) existing in the knowledge base. On the other hand, DSR addresses important unsolved problems in unique or innovative ways or solve problems in more effective or efficient ways. The key differentiator between professional design and design research is the clear identification of a contribution to the archival knowledge base of foundations and methodologies and the communication of this contribution to the stakeholder communities.

DSR has been interpreted as including two distinctly different classes of research – “design as research” and “researching design”. Design as research encompasses the idea that doing an innovative design that results in clear contributions to the knowledge base constitutes research. Design research projects are often performed in a specific application context, the resulting designs and design research contributions may be clearly influenced by the opportunities and constraints of the application domain.

### **2.3.1.2.5 Artefacts evaluation in Design Science Research**

Design Science Research is motivated by the desire to improve the environment through the introduction of new innovative artefacts and the processes for building these artefacts (Simon 1996). Good DSR often begins by identifying and representing opportunities and problems in an actual application environment. As livari (2007) notes, “It is the rigor of constructing IT artefacts that distinguishes IS as design science from the practice of building IT artefacts”.

During the performance of the design cycle, it is important to maintain a balance between the efforts spent in constructing and evaluating the evolving design artefact. Both activities must be convincingly based on relevance and rigor. Having a strongly grounded argument for the construction of the artefact, as discussed above, is insufficient if the subsequent evaluation is weak.

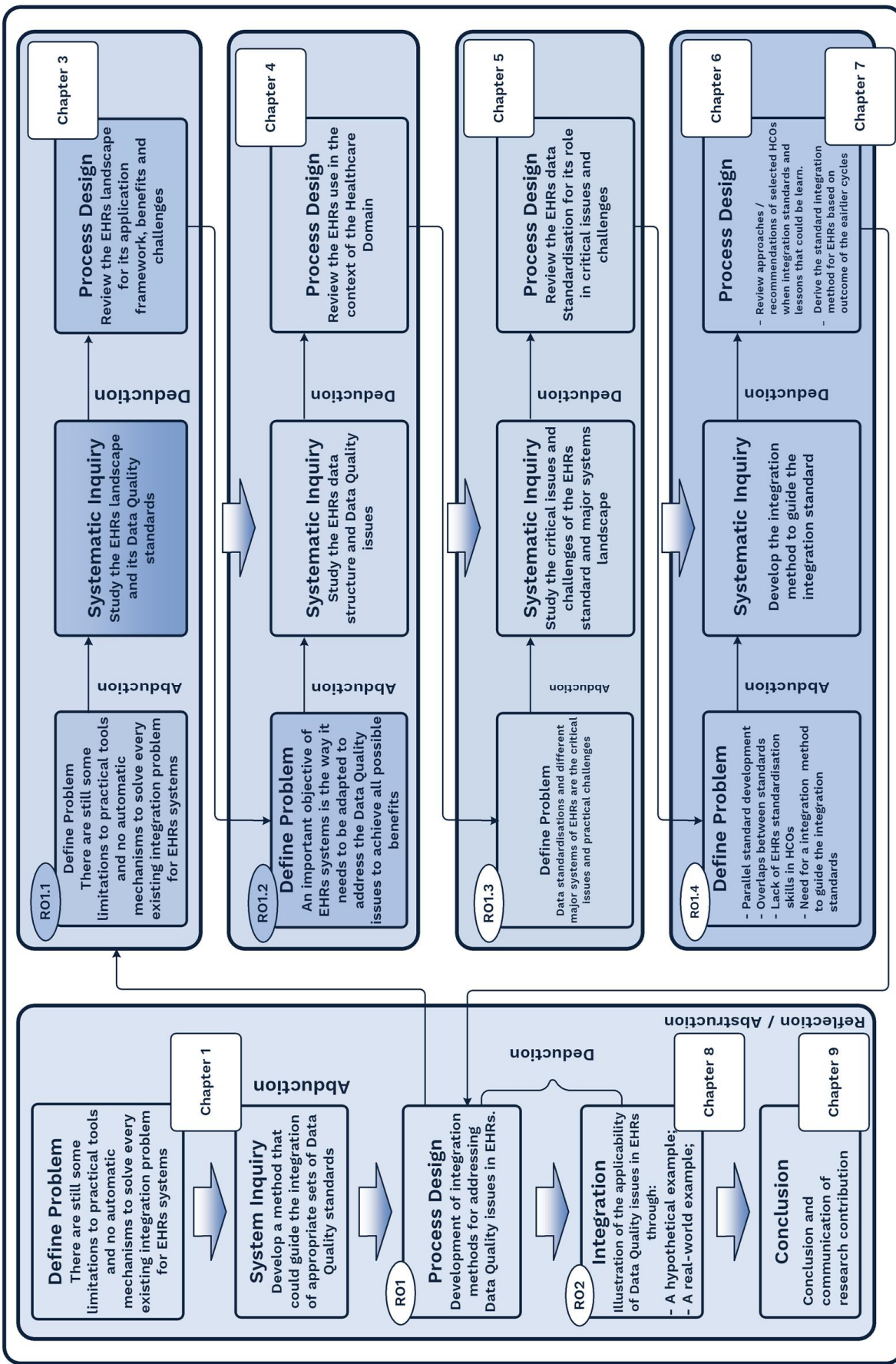


Figure 2.5: Design research phases used in this study

As livari (2007) states in his essay, “The essence of Information Systems as design science lies in the scientific evaluation of artefacts”.

This is in line with Juhani (2007), “I agree that artefacts must be rigorously and thoroughly tested in laboratory and experimental situations before releasing the artefact into field testing along the relevance cycle”. This calls for multiple iterations of the design cycle in DSR before contributions are put out into the relevance cycle and the rigour cycle. Can a research project effectively balance the goals of fundamental scientific understanding with considerations of the usefulness of the resulting artefacts?

### **2.3.2 Application of the Design Science Research strategy as used in this study**

The DSR strategy used in this thesis is the DSR methodology as proposed by Vaishnavi *et al.* 2015 and Aken *et al.* 2016. As discussed in section 2.3.1.2.4, this study DSR methodology involves seven phases, namely:

- 1) Idea generation;
- 2) Defining;
- 3) Process design;
- 4) Prototype;
- 5) Data analysis;
- 6) Interpretation;
- 7) Communication;

The DSR strategy for the research presented in this thesis, as illustrated in figure 2.3.1.2.4, involves one main DSR cycle consisting of seven phases. The development phase involves four sub-cycles, each consisting of the awareness of the problem, the suggestion of a solution and development phases. The form of the artefact developed is a method. In essence, the method employed in developing this essay is in itself a DSR approach (Gregor *et al.* 2013).

A DSR project can explore fundamentally new approaches to a certain issue without producing well-defined generic designs. A DSR article also can present a methodological innovation, for example, approaches for field testing generic designs in volatile environments (Aken *et al.* 2016). Personal learning is limited by the scope of one's personal experiences, yet experiential learning also can be the basis of research design, namely systematic and methodical experiential social learning.

This body of evidence is to be compiled through field testing a number of instantiations of the design within the intended application domain. In most cases, this involves rigorous case-studies using methods such as controlled observations, triangulation, “thick” descriptions, careful cross-case analyses and member checks. In DSR, field testing is key, often beginning with alpha testing (testing by the designers themselves) followed by beta testing (testing by third-party stakeholders). It is not a specific method with fixed rules rather, it is a strategy that can be operationalised in various ways (Aken *et al.* 2016). DSR is a research strategy. The differences from the more common explanatory research strategy lie at the level of strategy.

In principle, no differences exist at the tactical level of methods for data gathering and data analysis. DSR does not need specific methods at this tactical level (Aken *et al.* 2016). Research opportunities are less obvious and these situations rarely require research methods to solve the given problem (Gregor *et al.* 2013). The remainder of this section provides a detailed discussion of the DSR methodology followed in this thesis.

### **2.3.2.1 Main cycle of the Design Science Research process**

This section provides an overview of the main cycle of the DSR process for the research presented in this thesis and consists of seven phases. According to Vaishnavi *et al.* (2015), “DSR is sometimes called “Improvement Research” and this

designation emphasises the problem solving or performance improving nature of the activity”. It enables the realisation of the primary goal of the study, which is:

*The applicability of the Hybrid Integration Development Methodology (HIDM) to address the Data Quality (DQ) issues in Electronic Health Records (EHRs) for Large Scale Database (LSDB)*

The main DSR cycle consists of the following five phases Figure 2.3.1.2:

1. **Defining the problem:** Defining the problem of interesting research may come from multiple sources including new developments in the organisation or in a reference discipline. This phase of the DSR process was initiated by an initial literature study of EHRs, interoperability, adaptation and, EHRs standardisation.
2. **Systematic inquiry:** Suggestions for a problem solution are abductively drawn from the existing knowledge/theory base for the problem area (Peirce, 1931). These suggestions may, however, be inadequate for the problem or suffer from significant knowledge gaps (which make the problem a research problem). A systematic inquiry for the development of a generic method to guide the selection of an appropriate set of standards was proposed.
3. **Process design:** This phase involves the development of the standard integration method for EHRs for DQ issues, which triggered the four sub-cycles (discussed in section 2.3.2.2). It also involves iterations through the four sub-cycles of DSR with three phases in each sub-cycle, namely:
  - a) Defining the problem;
  - b) Systematic inquiry;
  - c) Integration;

The need for further literature studies on the different aspects of EHRs standardisation was deduced at the end of each DSR sub-cycle.

4. **Integration:** The integration phase involves testing the effectiveness of a given system or process, whereas field testing in DSR also has a crucial function in optimising and generalising a design and the demonstration of the applicability. The integration was done through the descriptive integration method (see section 2.3.2.2.4).
5. **Conclusion:** This phase involved reflections on investigating and addressing the EHRs DQ issues to abstract the study contributions to the body of knowledge. The contributions were communicated through two scientific publications (two book chapters). A summary of the research contributions is provided in Chapter Nine.

## 2.3.2.2 Research sub-cycles to develop the artefact

The research sub-cycles to develop the artefact involve the creation of innovative artefacts to solve real problems, namely constructs, models, methods and instantiations. It appreciates the levels of artefact abstractions that may be DSR contributions to include design theory.

### 2.3.2.2.1 Sub-Cycle One

Sub-cycle one enables the achievement of a broad understanding of DQ issues in EHRs for LSDB, its benefits and challenges.

Sub-cycle one consists of the following three phases:

1. **Defining the problem:** To develop a generic method, the DSR contents can be similar, but the problem definition and research objectives have specified the goals required of the artefact for development. The relevance of the research problem has started clearly that, it was important to obtain a brief understanding of EHRs.
2. **Systematic inquiry:** The need for the literature study on the EHRs was abducted to obtain a brief understanding of DQ issues in EHRs.



- 3. *Process design:*** This phase involved a broad analysis and investigation of the literature on EHRs, its application platform, and the benefits and challenges preventing its adaptation framework. The literature study on EHRs is presented in Chapter Three.

### 2.3.2.2.2 Sub-Cycle Two

This sub-cycle enables the achievement of RO1.2 (see figure 2.5), which is to determine what integration entails in the healthcare domain.

Sub-Cycle Two consists of the following three phases:

- 1. *Defining the problem:*** The literature review on EHRs revealed that the important objective of EHR systems is the way it needs to be adapted to address the DQ issues to achieve all possible benefits.
- 2. *Systematic inquiry:*** The need for the literature study on the EHRs was abducted to obtain a brief understanding of the meaning of DQ standards in EHRs integration.
- 3. *Process design:*** An in-depth analysis of EHRs integration literature was done. This includes the different levels of integration, which could be achieved, the essential method of integration and the benefit and challenges to the DQ issues in EHRs integration.

### 2.3.2.2.3 Sub-Cycle Three

This sub-cycle enables the achievement of RO1.3 (see figure 2.5), which is to study the EHR standards landscape and the role of standardisation in enabling integration in the healthcare domain.

Sub-Cycle Three consists of the following three phases:

- 1. Defining the problem:** The literature review on EHRs revealed that standardisation is one of the keys to address the critical issues and practical challenges.
- 2. Systematic inquiry:** To understand the role of standardisation in enabling integration, the need for an explanation of the EHRs data quality standards landscape was abducted.
- 3. Process design:** This phase involved an in-depth literature study on DQ issues in EHRs, the HCOs involved in EHRs data quality structure initiatives, the benefits of DQ and its challenges and an overview of EHRs data quality issues in Africa. Discussions were also held with a data expert to understand the risk associated with the combination of several DQ standards to address the DQ issues in EHRs. The literature study on EHRs data structure and its role in critical issues and challenges are presented in Chapter Five.

#### 2.3.2.2.4 Sub-Cycle Four

This sub-cycle enables the achievement of RO1.4 (see figure 2.5), which is deriving a generic data integration method to address DQ issues in EHRs.

Sub-Cycle Four consists of the following three phases:

- 1. Defining the problem:** The study of EHRs integration revealed many of the challenges associated with DQ issues. These included parallel integration, the overlap between the quality standard and lack of integration guidelines to accompany many of the published standards and the need for the integration structured method to address the DQ issues in EHRs.
- 2. Systematic inquiry:** This phase involved a proposal for the structure of a generic integration method to address the DQ issues in EHRs integration.

**3. *Process design:*** This phase represented the actual integration method to address the DQ issues in EHRs. It began with an explanation of the initiatives and approaches taken by the integration of countries and organisations and that which could be learned from the processes they followed (presented in Chapter Six). Based on the analyses of literature and discussions with the expert, the steps in EHRs integration were derived by drawing on the outcomes and lessons from earlier sub-cycles (presented in Chapter Six).

## 2.4 Data collection methods

Data collection is the process of gathering and measuring information on targeted variables in an established systematic fashion, which then enables one to answer relevant questions and evaluate outcomes. As mentioned in the research cycles discussed in section 2.3.2, several data collection methods were employed during the research presented in this study. According to John Dudovskiy (2016), the data collection method can be divided into two categories:

**1. *Secondary data collection methods:*** Secondary data collection methods use the type of data that have already been published in books, newspapers, magazines, journals, online portals etc. An abundance of data is available in these sources about research areas in business studies, regardless of the nature of the research area. Therefore, the application of an appropriate set of criteria to select secondary data to be used in the study plays an important role in terms of increasing the levels of research validity and reliability.

These criteria include, but are not limited to the date of publication, the credential of the author, the reliability of the source, the quality of discussions, the depth of analyses, the extent of the contribution of the text to the development of the research area. Secondary data collection is discussed in greater depth in the chapter on the literature review.

**2. Primary data collection methods:** Primary data collection methods can be divided into two groups, as follows:

**a) Quantitative data collection methods:** Quantitative data collection methods are based on mathematical calculations in various formats. Quantitative data collection methods and analysis include questionnaires with closed-ended questions, methods of correlation and regression, mean, mode and median and others.

Correlation can be explained as a single number which describes the extent of the relationship between two variables. The relationship between these two variables is described through a single value, which is the coefficient.

**Correlation analysis:** Correlation coefficient 'r' is a number that represents the level of the relationship between two individual variables (Washington *et al.* 2010). For instance, the correlation coefficient can assist in identifying the relationship between consumer age groups and the type of atmosphere in a restaurant they enjoy the most. Similarly, the correlation coefficient can be used to establish the nature of the relationships between consumer genders and the level of their interests.

The coefficient of correlation is expressed by the formula:

$$c = \frac{n \sum ab - \sum a \sum b}{\sqrt{(n \sum a^2 - (\sum a)^2)(n \sum b^2 - (\sum b)^2)}} \quad (2.1)$$

Where,

*a and b* – The correlation coefficient;

*c* – Correlation;

And others the range of value 'c' can take changes from +1 to -1 depending on the type of correlation. Specifically,

(i) The interrelation would be completely positive if 'c' is equal to +1;

(ii) The interrelation would be completely negative if 'c' is equal to -1;

(iii) The two variables relationship would be apprehending to be non-interrelated if 'c' is equal to 0 (zero);

*b) Qualitative:* According to William (2005), “the qualitative data collection methods emerged after it has become known that traditional quantitative data collection methods were unable to express human feelings and emotions”. Qualitative research methods, on the contrary, do not involve numbers or mathematical calculations and are closely associated with words, sounds, feeling, emotions, colours and other components that are non-quantifiable. Qualitative studies aim to ensure a greater level of depth of understanding and qualitative data collection methods include interviews, questionnaires with open-ended questions, focus groups, observation, game or role-playing and case studies etc.

Primary data collection methods have been used to collect data in this study. All data has been collected under Non-Disclosure Agreements (NDA). The details of data collection sources and questioners used, section 8.4 and 9.2.3 provides the overview. This section briefly explains the data collection methods in this study.

## 2.5 Data analysis

This methodology chapter also includes the methods of data analysis to explain in a brief way, the way to analyse the method used to collect the primary data.

Qualitative data analysis differs from the analysis of quantitative data. In qualitative research, focus groups and experiments, among others, are used. Data analysis involve identifying the common patterns within the responses and critically analysing them to achieve research aims and objectives.

Data analysis for quantitative studies, on the other hand, involves critical analysis and interpretation of figures and numbers and attempts to find the rationale behind the emergence of main findings. Comparisons of primary research findings with the findings of the literature review are critically important for both types of studies – qualitative and quantitative.

Data analysis methods in the absence of primary data collection can involve the discussion of common patterns, as well as, controversies within secondary data directly related to the research area.

***Qualitative data analysis:*** Qualitative data analysis can be conducted through the following three steps:

***Step One (Developing and applying codes):*** Coding can be explained as the categorisation of data. A 'code' can be a word or a short phrase that represents a theme or an idea. All codes need to be assigned meaningful titles. A wide range of non-quantifiable components such as events, behaviours, activities and meanings can be coded.

***Step Two (Identifying themes, patterns and relationships):*** Unlike in quantitative methods, in qualitative data analysis, no universally applicable techniques are used that can be applied to generate specific findings. The researcher's analytical and critical thinking skills play a significant role in data analysis in qualitative studies. Therefore, no qualitative study can be repeated to generate the same results.

***Step Three (Summarising the data):*** At this last stage the method findings are linked to the hypotheses or research aim and objectives. When writing the data analysis chapter, the method uses noteworthy quotations from the transcript to highlight major themes within the findings and possible contradictions.

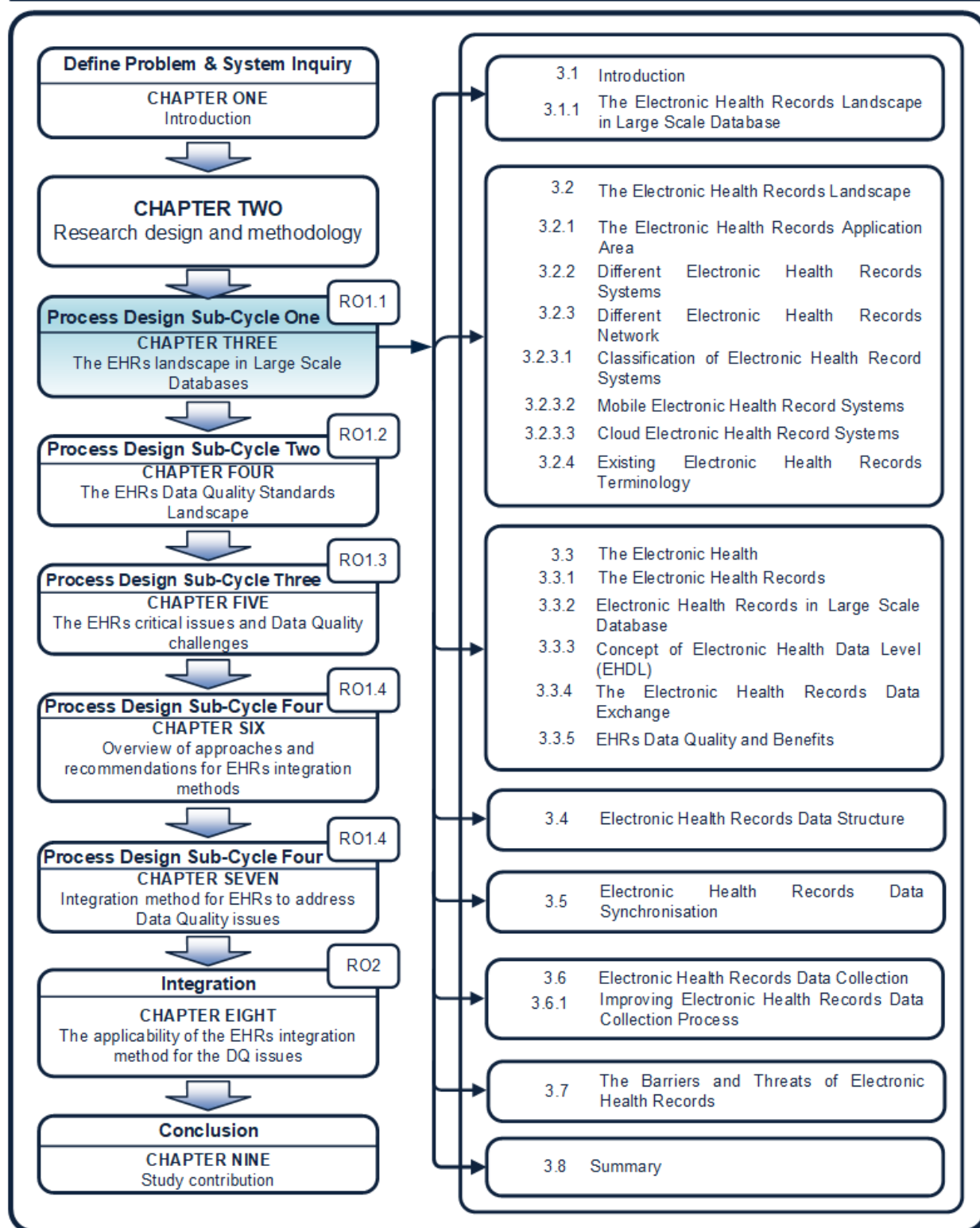
***Quantitative data analysis:*** In quantitative data analysis the outcome expected is, namely, to turn raw numbers into meaningful data through the application of rational and critical thinking. The same figure within the data set can be interpreted in many different ways; therefore, it is important to apply fair and careful judgement.

## 2.6 Summary

Chapter Two present the study design and methodology used in this research as guided by the research objectives. To justify the basis for the research paradigm

followed, the chapter began with a general discussion of the four research paradigms and the philosophical assumptions that blueprinted them. This was followed by a detailed discussion of the way in which the DSR strategy was used to guide the EHRs integration to address the DQ issues. Chapter Three provides detailed discussions of a literature review on DQ issues in EHRs.

## CHAPTER THREE: The EHRs landscape in Large Scale Databases



Outline of the Chapter Three



---

## CHAPTER THREE

### 3.1 Introduction

The purpose of this chapter is to review the existing knowledge from a substantive solution of scholarly papers to obtain a broad understanding of the large-scale EHRs landscape and the challenges of DQ issues. The main purpose of this review is to study and identify relevant knowledge on research productivity. EHRs are defined to use Information Technology (IT) to allow HCOs to deliver a higher quality of care to their patients than that which is possible with the paper-based record. It maps to Sub-Cycle One of the DSR process, as described in section 2.3.2.2.1 and highlighted in Figure 3.1.

To contextualise the discussion, section 3.2 provides an overview of the general concept architecture, different technology and existing EHR systems. This is followed by a detailed discussion and broad overview of eHealth, EHRs, EHRs in LSDB, the concept of EHRs Data Level (EHDL), EHRs data exchange and EHRs data quality and benefits in section 3.3. The EHRs data structures are discussed in section 3.4. Section 3.5 explores the data synchronisation and section 3.6 the EHRs data collection. In section 3.7 the barriers and challenges of DQ in EHRs are described. A summary of the chapter is provided in section 3.8, with section 3.9 concluding the chapter.

#### 3.1.1 The Electronic Health Records landscape in Large Scale Databases

Electronic Health Records (EHRs) refer to implemented structured digital manifestations of real-time, patient-centred health records (Abdel *et al.* 2015). The EHRs landscape in LSDB of the study is presented, followed by an outline of this chapter, and is detailed below:

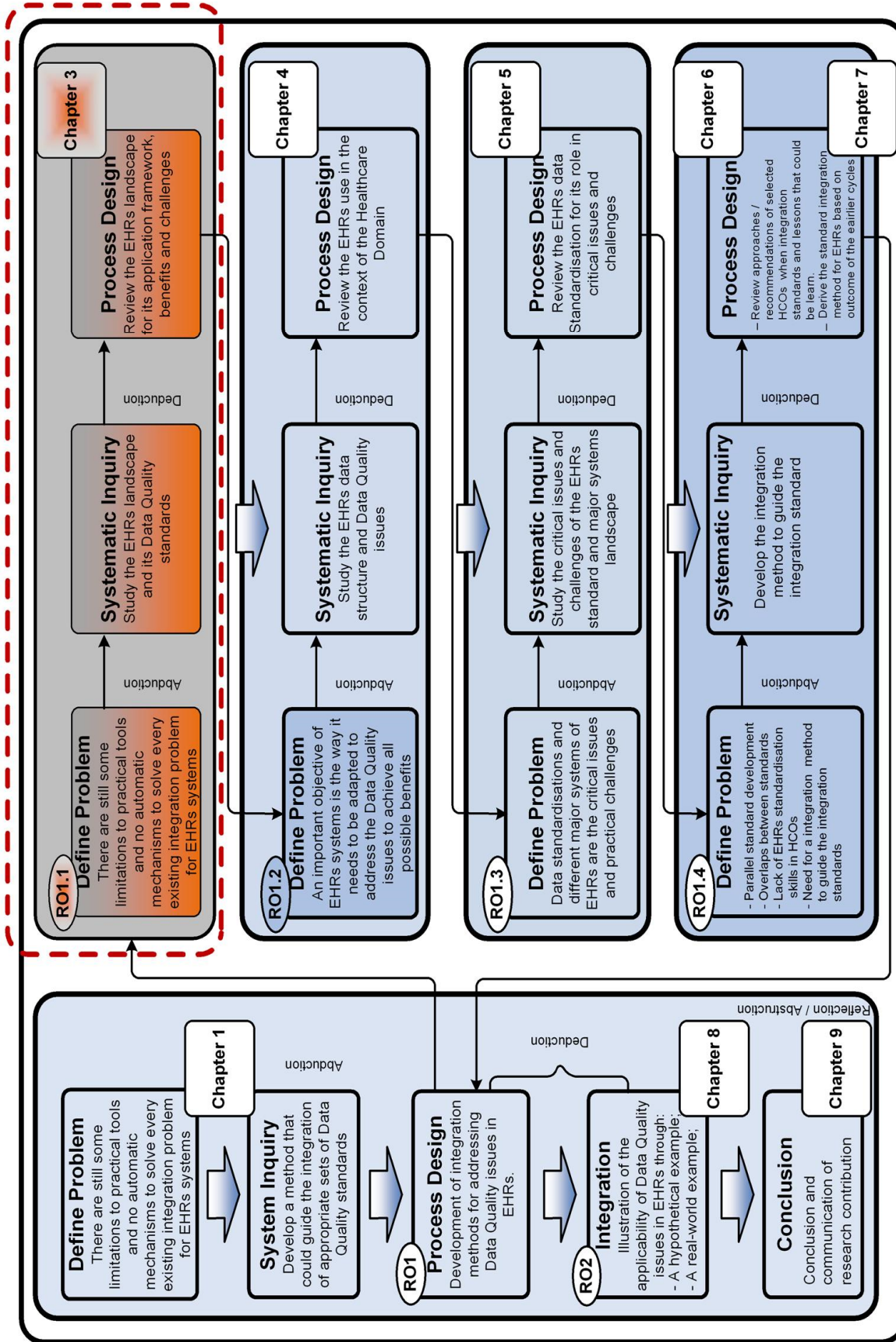


Figure 3.1: The position of Chapter 3 in the design science research process used in this study

The literature reviews are the secondary source which refers to the existing knowledge including the substantive solution of scholarly papers, as well as theoretical and methodological contributions to a particular topic. It means that literature reviews are not the original experimental works and do not report the new findings.

## **3.2 The Electronic Health Records landscape**

The Electronic Health Records (EHRs) refer to the system that provides secure, centralised and private lifetime implemented structured digital manifestations of real-time, patient health records of a person's health and healthcare history. The EHR systems store and share over the HCOs such information as demographics, lab results and reports, medication profiles, diagnostics and immunisations history. The electronic record is available instantly and accurately to authorised users anywhere and anytime in support of delivering a higher quality of care to their patients than that which is possible with paper-based records.

The goal of the EHR is to provide a patient-centric solution to the challenges currently facing the provinces healthcare sector. It does this by providing the means for diverse healthcare systems and data sources to securely share information. As a result, authorised users can have a seamless, efficient way to request information from different sources and input information into designated data sources. For example, through the EHR an authorised lab will be able to enter a lab test result into the lab data repository and the provider who requested the test can retrieve the result from the same repository. In this way, a comprehensive healthcare picture for a patient can be assembled as if it were coming from a single system.

### **3.2.1 The Electronic Health Records application area**

The first EHR systems were known as clinical information systems. In the mid-1960s, Lockheed developed one such product, which has since been handed down

to the vendor Technicon, then to TDS Healthcare and then to Eclipsys, now part of All-scripts (Amatayakul 2007). It influenced later systems because its processing speed and flexibility allowed many users in the system at once (Dick *et al.* 1997). Since the 1980s, more concerted efforts have been made to increase the use of EHR. The EMRs of today first appeared in 1972 from the Regestrief Institute in Indianapolis but were so expensive that they did not spread among physicians. Instead, they were used by government hospitals, according to the University of Scranton in Scranton.

By the 1990s, technology had entered most medical offices and computers were being used to a limited degree for record keeping purposes. But, it was not until the age of the Internet that large-scale change became far more visible. Even in its early stages, the Internet became a vital tool for recording and transferring prescription histories and other medical records. Finally, within the last decade or so, most major medical systems in the developed world could easily communicate with each other when needed.

Today, medical records are increasingly paperless, although some private practices continue to use a combination of paper and computerised records. Patient medical records are more accessible than ever before with data technology becoming increasingly portable and comprehensive. Current refinements in the medical records industry are aimed at the continued specialisation of systems to further streamline workflows, boost productivity and improve doctor-patient interactions

Two major challenges, however, remain when it comes to electronic medical records. The first challenge is, of course, the security. Due to the unique nature of doctor-patient privacy, questions around electronic data and privacy have been shaping both public policy and private software development. HIPAA guidelines, for example, were designed to deal with the security of patient medical records. Challenges in this area remain and both the public and private sectors are focused on strengthening the security of medical records at all access and transmission points.

### 3.2.2 Different Electronic Health Records systems

Electronic Health Records (EHRs) are a digital version of the paper-based health record. Electronic Medical Records (EMRs) are a digital version of the paper chart in the provider's office. The EMRs contains the medical and treatment history of the patients in one practice. EMRs have advantages over paper records, for example, EMRs allow clinicians to:

- a)* Track data and patient history over time;
- b)* Easily identify which patients are due for preventive screenings or check-ups;
- c)* Check how their patients are doing on certain parameters, such as blood pressure readings or vaccinations;
- d)* Monitor and improve the overall quality of care within the practice;

The information in EMRs, however, does not travel easily out of the practice. In fact, the patient's record might even have to be printed out and delivered by mail to specialists and other members of the care team. In that regard, EMRs are not much better than a paper record.

The EHRs do all those things and more. EHRs focus on the total health of the patient going beyond standard clinical data collected in the provider's office and inclusive of a broader view on a patient's care. EHRs are designed to reach out beyond the health organisation that originally collects and compiles the information. They are built to share information with other healthcare providers, such as laboratories and specialists; therefore, so they contain information from all the clinicians involved in the patient's care. The National Alliance for Health Information Technology (NAHIT) stated that EHRs "can be created, managed and consulted by authorised clinicians and staff across more than one healthcare organisation."

The information moves with the patient to the specialist, the hospital, the nursing home, the next state or even across the country. In comparing the differences between record types, HIMSS analytics stated that "The EHR represents the ability

to easily share medical information among stakeholders and to have a patient's information follow him or her through the various modalities of care engaged by that individual". EHRs are designed to be accessed by all people involved in the patients' care including the patients themselves. Indeed, that is an explicit expectation in stage one definition of "meaningful use" of EHRs. And that makes all the difference because when information is shared in a secure way, it becomes more powerful. Healthcare is a team effort and shared information supports that effort. After all, much of the value derived from the healthcare delivery system results from the effective communication of information from one party to another and ultimately the ability of multiple parties to engage in interactive communication of information.

***Benefits of Electronic Health Records, as follows:***

With fully functional EHRs, all members of the team have ready access to the latest information allowing for more coordinated, patient-centered care. With EHRs:

- a) The information gathered by the primary care provider tells the emergency department clinician about the patient's life-threatening allergy so that care can be adjusted appropriately, even if the patient is unconscious.
- b) Patients can log on to their own record and see the trend of the lab results over the last year, which can help motivate them to take the medications and keep up with the lifestyle changes that have improved the numbers.
- c) The lab results run the previous week are already in the record to tell the specialists what they need to know without running duplicate tests.
- d) The clinician's notes from the patient's hospital stay can help inform the discharge instructions and follow-up care and enable the patient to move from one care setting to another more smoothly.

So, yes, the difference between "*electronic medical records*" and "*electronic health records*" are just one word. But in that word, there is a world of difference.

### 3.2.3 Different Electronic Health Records network

In co-operating distributed health information systems and networks, the EHR Systems provided a lifelong patient record advance towards core applications (Bernd 2006). Several researchers have shown that only the arbitrary access to patient health information is the proximal motive of the accurate decision in healthcare during decision-making and the effective communication between patient care team members (Chery *et al.* 2012). The number of hospitals and clinics are increasing every day, as well as increasing health information. Health information has been digitalised and archived in their health record with the universal use of the computer and information technology network. Vast types of wired and wireless network layout exist, consisting of the type of device, including hardware, software, connectivity protocols and communication mode of transmission. It also includes knowledge about the types of networks grouped according to types such as LAN, MAN and WAN.

Such is cloud computing, that refers to fast computation and its capability to store large storage space. Now-a-days cloud computing is a convenient, on-demand network. It also entails configurable computing resources to a share group network such as application, service, server and archive. With the minimal managerial effort, cloud computing can be rapidly provided and released for higher productivity.

The EHR system can be integrated into cloud computing. Basically, the smaller hospital and clinic have limited resources. Cloud computing can facilitate these smaller HCOs with adequate electronic medical record storage space to provide the exchange and sharing of electronic medical records (Charalampos *et al.* 2010). Cloud computing has a high impact on parallel distributed grid computing systems. The flexibility for the further development of these techniques is recommendable. Cloud integration is based on service-oriented architecture applications and HCOs can take advantage of the ability to combine all cloud applications as well as systems the traditional and existing systems. The large-scale cross-platform including the cloud terminology of EHR system architecture overview is shown in figure 3.2 (Saiod *et al.* 2017), as follows:

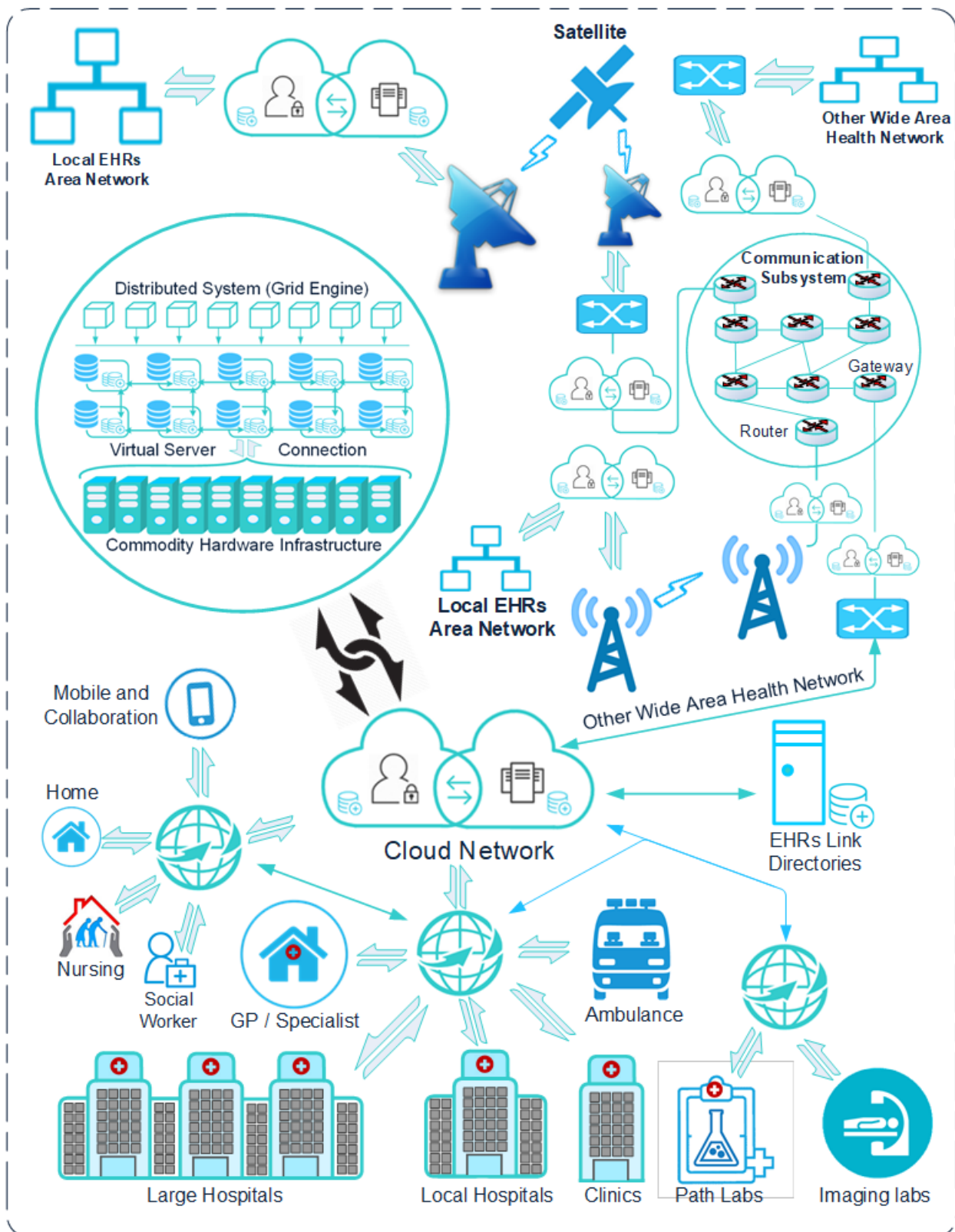


Figure 3.2: Large-scale cross-platform EHR system architecture overview (Saïod *et al.* 2017).



This is a very effective location independent technology and also enhance the user experiences over the Internet. These days, it can provide services for various application scenarios. More and more applications are migrated onto the cloud platform (Hsu *et al.* 2011).

### 3.2.3.1 Classification of Electronic Health Records systems

Identifying disease gene opens the door for clinical testing for individuals at risk of having a gene mutation. The two categories of EHR systems are:

- a) Cloud-based technology;
- b) Client-server based technology;

In a computer with Internet connection in order to access via the web, the online data can be stored externally. The cloud computing application allows users on-demand access and provide by the third-party organisation using the Internet.

### 3.2.3.2 Mobile Electronic Health Records systems

The medical record has evolved dramatically, from the paper base to electronic which was the initial stage of evaluation. Now it is being changed in the physical space from occupying walls of shelving to server rooms to the cloud and currently an independent location and even to the pocket. Mobile health service is another innovative technology in EHR systems that can provide a wide range of location independent services. **Switching from the paper base was, however, not an easy task as many doctors have been hesitant to switch for many reasons, such as trust, reliability, cost and time-consuming factors.**

Mobile health can provide great benefits to both patients and providers for healthcare service and includes monitoring, telemedicine, location independent medical services, emergency management, response and pervasive access to

healthcare information (Maglogiannis *et al.* 2009). Figure 3.3 (researcher source) describes Mobile eHealth Record Systems (MERS) architecture, as follows:

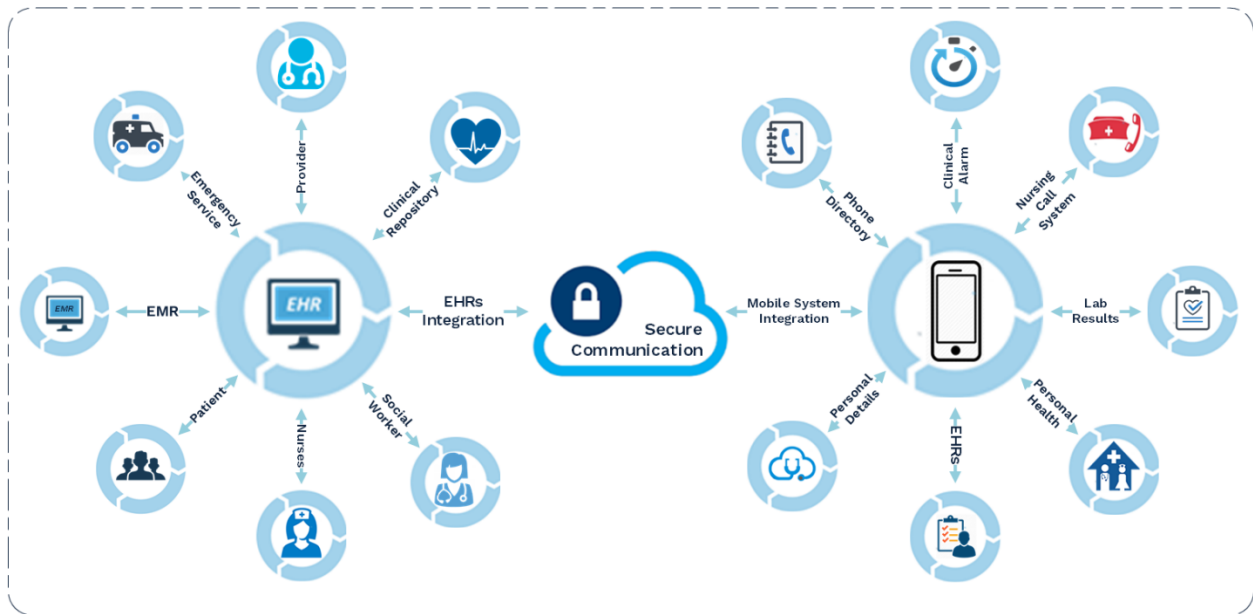


Figure 3.3: Mobile eHealth Record Systems (MERS) architecture (researcher source)

Most office-based providers (78%) are using EHRs, which means the infrastructure for the EHRs are almost complete. The next step is to integrate EHRs with HCOs to mobile devices, for the following reasons:

- a) Better decision-making from anywhere;
- b) A more convenient way to provide healthcare service and communication;
- c) The availability location independent of EHRs applications and information;
- d) The anytime and the invisibility of computing;

### 3.2.3.3 Cloud Electronic Health Record systems

Ease of data or record sharing at will has compelled most of the physicians to adopt EHRs for the record-keeping of patients. This is also convenient for the other

stakeholders of the healthcare ecosystem such as nurses, specialists and the patient.

Due to lower costs and the scalability of the application, the cloud is becoming the infrastructure for most of the EHRs but without comprising the privacy of data. Figure 3.4 (Saïod *et al.* 2019b) describes cloud Electronic Health Record (EHR) systems, as follows:

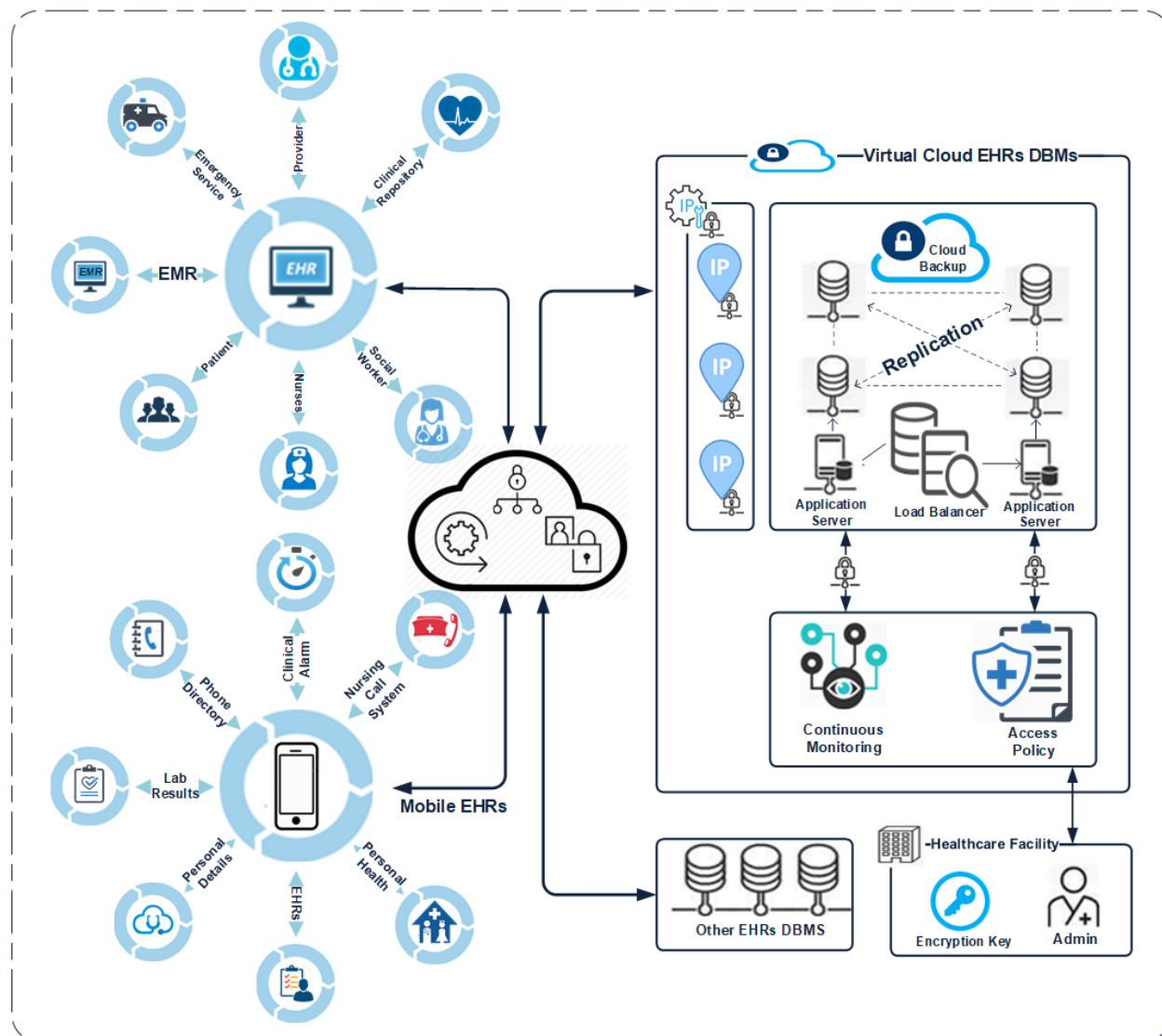


Figure 3.4: Cloud Electronic Health Record systems (Saïod *et al.* 2019b)

The IT benefits of Cloud-Based EHRs include that when choosing EHR systems, providers have the choice of hosting the software on their own network (client-

server) or EHR systems where the software is hosted on a remote server accessed through the Internet (cloud-based).

**Advantages:** As the pressure to lower healthcare costs, improve communication and adopt systems that will support EHRs rises, cloud computing is quickly becoming an important part of the healthcare industry for the following reasons:

- a) Reduced start-up costs:** The cost of setting up a client-server EHRs is a significant hurdle to a small practice. The start-up costs can range upward of R570396.00 South African Rand (equivalent to \$40,000) just for a single practice. With cloud-based EHR systems, practices benefit from economy of scale. Since many providers use the same system, redundant costs are minimised or eliminated.
- b) Lower infrastructure and IT costs:** Client-server EHRs require the practice to purchase or lease expensive hardware. Practices must hire IT staff or pay for the services of IT personnel to set up, test, maintain and upgrade the hardware and software. With a cloud-based EHR, all the costs of running the system are covered by the EHR vendor or hosting company. This means no hardware, network or maintenance costs to the practice over the typical equipment setup required to run a medical care business.
- c) Cost predictability:** The costs of a client-server system can lead to unpredictable costs. If the server crashes or an upgrade goes wrong, the practice's emergency fund takes a hit or worse. Cloud-based EHR systems have consistent costs that allow the practice owner to feel confident in their financial projections. The practice simply pays a monthly or quarterly access fee, similar to the fee for phone or Internet.
- d) Simpler implementation and scalability:** The process of setting up and testing a client-server EHR is more complex than cloud-based systems and scaling up. As one's practice grows it usually requires additional equipment or licensing costs. Under a cloud-based EHR, the practice personnel accesses the system through a secure website or client software installed

on their computers. Gaining capacity is simply a matter of contacting the EHR vendor and adding more users.

e) ***Better patient data security:*** If the practice currently relies on paper records for storing patient data, imagine what could happen if one had a fire, flood or another disaster. Insurance covers new equipment, but patient data is irreplaceable. Although practices with client-server EHRs generally have off-site backups, the data is vulnerable during transport and the practice must pay extra for storage costs. Cloud-based EHR records are transferred using secure encryption and backed up in multiple locations automatically at no extra cost.

***Disadvantages of cloud computing:*** The significant risks that each HCO will face when transitioning to cloud-based hosting. The main disadvantage of cloud computing is that all data, security, availability, maintenance and control domain, is a third party. Therefore, HCOs have absolutely no control over these matters. Trusting the third-party service provider is one of the important factors for cloud computing and it takes on a whole different meaning (Rodney 2012). Despite all the barriers, it is important to remember that cloud health computing paradigms are still under development, but with a lot of chances of being a revolution in numerous fields. In the near future, there will be more services on offer and the development will be larger.

### **3.2.4 Existing Electronic Health Records terminology**

Existing literature shows that several techniques and major EHR systems currently exist to deal with DQ issues, which historically have faced DBMS. After a profound analysis of various cutting-edge commercial accomplishments existing on the software market and an intensive review of the literature, some limitations still appear to the practical tools for EHR systems. Physically access to diverse information sources of robust support is provided, but only if these are standard database structure tables.

At the moment, no automatic mechanisms exist to solve existing integration problems (Nirase *et al.* 2016). Peer-to-Peer (P2P) topology is used when system-individual participants contact a localised server to search other data and to contact other participants directly, to exchange information or share resources. However, Gribble *et al.* (2001) stated that “the generic P2P systems often do not take care of the semantics of the data exchanged. This is a serious drawback, especially considering that when the network grows, it becomes hard to predict the location and the quality of the data provided by the system”.

Tania *et al.* (2004) propose a mediated query service, which is a system used for configuring mediation systems for building and maintaining multi-dimensional multimedia data warehouses. Considerable disadvantages are, however, involved in moving data from multiple, often highly disparate data sources, into a single data warehouse. This translates into a long implementation time, high cost, and lack of flexibility, outdated information and limited capabilities.

A subsequent representation by Gilson *et al.* (2005) for data integration was proposed using middleware architecture. The middleware can encompass dynamic scheduling, performance management and transport services for distributing scientific visualisation tasks in a grid environment. Middleware, however, has a high development cost, the implementation thereof is time and resource consuming, few satisfying standards exist, its tools are not good enough and often threatens the real-time performance of a system and middleware products are not very mature. Load-balancing issues, limited scalability, low levels of fault tolerance and limited programmer access area, for example, are some of the main disadvantages of middleware.

Combining Aggregation Operators (AO) and fuzzy Description Logics (DL), Vojta. (2006) presents a fuzzy DL with general AOs. The expressiveness of the logic is, however, very limited. An additional evaluation of this strategy was also done for data and multimedia sources using an ontology-based data integration system. A Mediator Environment for Multiple Information Sources (MOMIS) data integration system was proposed using a single ontology approach to overcome this limitation.

The system combines the MOMIS framework with the STASIS framework. MAFRA (Maedche *et al.* 2002) is an ontology mapping framework for distributed ontology, which supports an interactive, incremental and dynamic ontology mapping process in the semantic web context. Using ontology integration, the conflicts in the result can be solved by satisfying the consistency criterion. An approximation technique has been identified as a potential way to reduce the complexity of reasoning over ontologies in expressive languages such as OWL 1 DL and OWL 2 DL (Reb *et al.* 2010).

A vast amount of research concerning EHR mechanisms has been carried out over the last few years. Fuzzy-ontology is moving forward to express fuzzy properties, membership functions and linguistic hedges (Carlos *et al.* 2016). The fuzzy-ontology definitions found in the literature are quite naturally influenced by fuzzy set theory, fuzzy logic and existing ontology languages. Shaker *et al.* 2015 performed an exercise using fuzzy-ontology integration to solve the problem of equivalently matching concepts to avoid pairs of mismatching concepts and conflicts regarding multiple entities to reduce data inconsistency. Another related work was performed by Sanchez *et al.* (2006), which considers Fuzzy-Ontology with general quantifiers that could be used for some type of quantifier-guided aggregation.

Cristiane *et al.* (2010) used a DISFOQuE system to analyse the fuzzy-ontology to perform semantic query expansions. This is an ontology-based data integration system for data and multimedia sources, which is essentially performed manually by the integration designer. A few studies handle fuzziness and give support for uncertainty in their conceptual models for multimedia materials. The studies by Aygün *et al.* (2004) and Özgür (2007) tried to deal with this uncertainty by supporting Fuzzy attributes. Different types of databases exist, but the type most commonly used in healthcare is the Online Transaction Processing (OTP) database. For the most part, healthcare databases are used as the foundation for running the many transactional system databases, which structures accommodate the creation of a wide range of transactional applications such as EHRs lab systems, financial systems, patient satisfaction systems, patient identification, data tracking, administration, billing and payment processing and research.

The EHRs database servers are to replace the old paper-based documents, files, folders and filing cabinets. Data is therefore now more convenient and current. It's obvious that the benefits of EHRs are equal to the benefits of the applications that run on them. Significant advances in automation and standardisation of business and clinical processes can be attributed to these applications and databases.

With EHR databases, data can also be stored externally and backed up in a secure place to prevent data loss. Because front-end software can provide tip text and enforce data integrity, the back-end data can, therefore, become more standardised and accurate. Lastly, because the data is electronic, it allows for quicker processing of typical transactions such as lab results and payment claims. One of the biggest benefits of all these databases is the amount of data healthcare organisations have been able to capture.

They now have huge data stores that can be used to inform better and more cost-effective care. EHRs focus on strategies for combining data residing at different heterogeneous sources and providing users with a unified view of the data. A vast amount of work has been developed in the EHRs area and some interesting results have shown the effectiveness of this approach. It has, however, not been extensively evaluated with regard to ease of data access to dynamic EHR systems and their wide implementation over HCOs.

The aim is to avoid the theoretic pitfalls of monolithic ontologies, facilitate interoperability between different and independent ontologies and provide flexible EHRs capabilities. In addition, not all the existing EHRs integration techniques are sufficient, as many healthcare organisations are still capturing their data in spreadsheets and often mismatch information and formats, which cause incorrect report generation and reduce the quality of the data. There is thus a need to develop or use an efficient EHR system, using a template screen, which is efficiently mapped to the online transaction-processing database. **An important objective of EHR systems is the way they need to be adapted to address the DQ issue to achieve all possible benefits and address, all the problems described above.**



### 3.3 Electronic Health

According to Eysenbach (2001), “eHealth refers to a concerted effort undertaken by leaders in healthcare and hi-tech industries to fully harness the benefits available through the convergence of the Internet and healthcare”. eHealth is an emerging field in the intersection of medical informatics, public health and business referring to health services and information delivered or enhanced through the Internet and related technologies.

WHO defines eHealth as the use of Information and Communication Technologies (ICT) for health. An overview of eHealth architecture is shown in figure 3.5 (Saïod *et al.* 2017), as below:

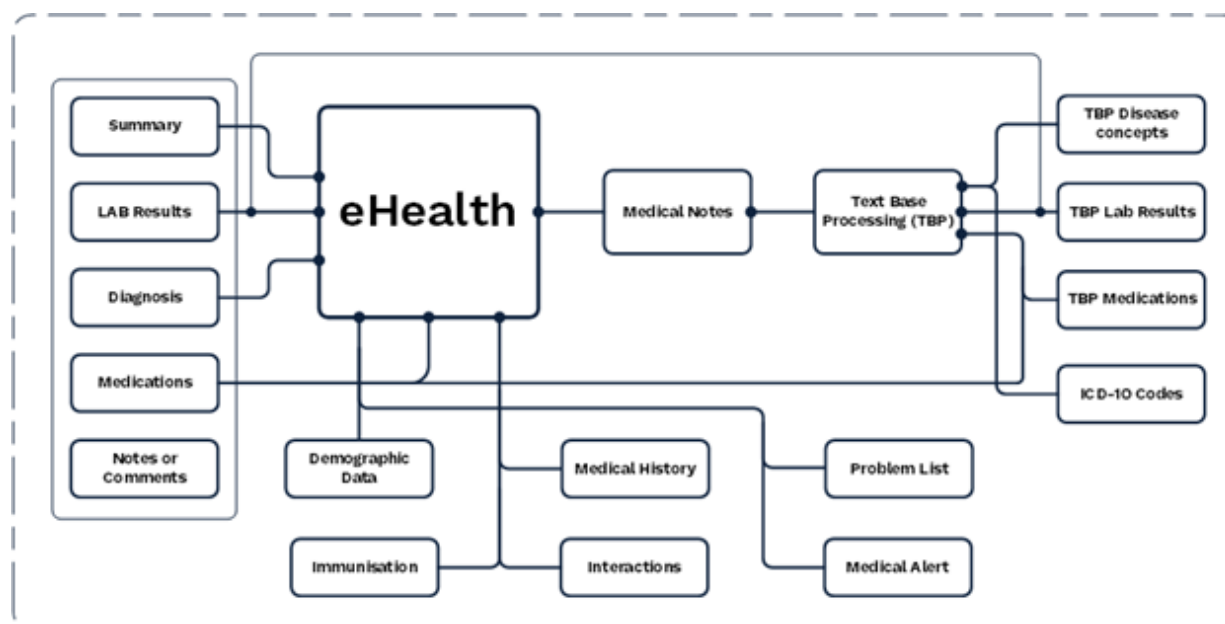


Figure 3.5: An overview of eHealth architecture (Saïod *et al.* 2017)

The eHealth is bringing to the delivery of healthcare around the world today and the way it is making health systems more efficient and more responsive to people’s needs and expectations. In its broadest sense, eHealth is about improving the flow of information, through electronic means, to support the delivery of health services and the management of health systems.

In other words, eHealth is the cost-effective and secure use of ICT in support of health and health-related fields, including health-care services, health surveillance, health literature and health education, knowledge and research.

### 3.3.1 The Electronic Health Records

Electronic Health Records (EHRs) refer to implemented structured digital manifestations of real-time, patient-centred health records (Abdel *et al.* 2015). EHRs are considered as one of healthcare's innovation heuristic items and are widely adopted over HCOs and are becoming an important mechanism to perform their daily services (Jinyuan *et al.* 2010; Matthew *et al.* 2016). The secure EHR systems provide information available instantly and accurately to the authorised users, therefore the user can coherently create new consistency of data sets (Illhoi *et al.* 2012; Weng *et al.* 2013). The EHR systems framework is shown in figure 3.6 (Saïod *et al.* 2017), as below:

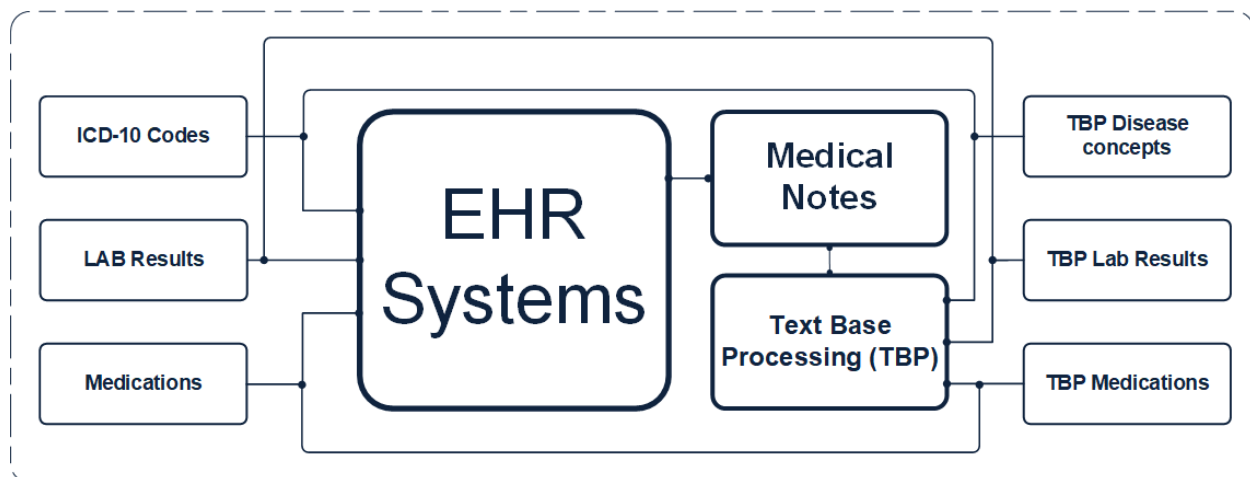


Figure 3.6: EHR systems framework (Saïod *et al.* 2017)

In general, the EHRs are faced with the problem of combining health data that reside at different sources and providing an accurate, comprehensive up-to-date patient history (Terence 2015; Ibrahim 2016). Improvements in the DQ have brought about efficiency, scalability and safety in the implementation of a large-scale healthcare DBMS (Jens *et al.* 2012).

Health data can, therefore, be composed and managed by the authorised user and consulted by authorised providers from across multiple HCOs nations or global wide and can be shared across them. The EHRs include an enormous range of patient data set, including patient details, history, references, medication, immunisation, allergies, radiology report including images, laboratory data and test reports, admission and discharge details, personal statistics such as Body Mass Index (BMI), blood pressure, and sugar level. These datasets are electronically stored in the database as narrative (free text) or encrypted data.

The EHR databases are structured to store accurately and securely health information over time. It can reduce the data replication risk as all access points are retrieving data from the main data server and reduce much lots of paperwork. The principle of data replication is to share information between multiple resources. The replication reduces fault tolerance and increases high accessibility and reliability. Many distributed database systems are using replication to avoid single access point failure and high traffic. It can be possible to dynamically improve load-spreading and load-balancing performance by providing replication (Yeturu *et al.* 2016).

Replication supports the restoring of replicated databases to the same server and database from which the backup was created (Andrew *et al.* 2015). Backup is one of the important processes of database server routine maintenance plans, namely to copying and archiving data to an external device. Therefore, backup data can be used to restore the original information after any data loss event. Now-a-days, electronic data is searchable even from heterogeneous sources and it is possible to combine them into a single data set. EHRs are even more effective when analysing long-term patient medical history (Gombert *et al.* 2015).

Due to EHRs data is tractable and it is easy to identify patient preventive visits or screening information and monitor the overall progress more effectively than the paper-based record in HCOs. EHRs improve patient care, increase patient participation, improve care coordination, improve diagnostics and patient outcomes, practice efficiencies for cost savings and allow more case studies for

research purposes. Despite the many advantages and functionalities of EHR systems, a considerable number of disadvantages are still associated with this technology (Brent 2014). One of the key concerns is the quality of the data, which includes inconsistency, privacy protection and record synchronisation, lack of standardised terminology, system architecture indexing and deficient standardised terminologies. The productivity may drop temporally with associated EHRs adaptation as workflows have changed.

Several long-standing consequences are emerging from the critical issue of EHRs adaptation (Andrea *et al.* 2005). Therefore, it is the utmost importance to advise healthcare organisations to choose the correct EHR systems and provide a proper setup to establish the complete system to become successful users of EHR systems (Nir *et al.* 2011). Healthcare organisations using tangible augmented EHR systems in their facilities can make better decisions based on the comprehensive information available to them. Improving healthcare distribution systems are becoming the most consequential technology for medical innovation of all the times.

EHR systems exhibit promising potential, which will play a crucial role in HCOs to ensure the provision of excellent patient care service, quality management, accurate information, perfect diagnosis, patient information safety, disease management and investigation as advancing innovation deftness (Sumit 2014). In particular, the chapter focuses on large-scale DBMS, the DQ introduction of smart interfaces and perfect data mapping in traditional EHR systems as well as mobile and cloud computing. This implies that integrated adoptive EHRs can show inconsistencies because the data structure and standard from the various HCOs are different.

### **3.3.2 Electronic Health Records in Large Scale Databases**

Although there is apparently no official or significant definition, Large Scale Databases (LSDB) refer to these decision support systems that process

transactions with magnetic storage in the terabyte range, containing billions of table rows and serving large numbers of users. In addition, LSDB may store every transaction for billions of patient and handle millions of queries per second over HCOs. Therefore, EHRs have to be an extremely special environment capable of handling everything the HCOs needs in terms of scale, performance and availability. Figure 3.7 (Saïod *et al.* 2017) describes the electronic health record systems in large-scale DBMS, as follows:

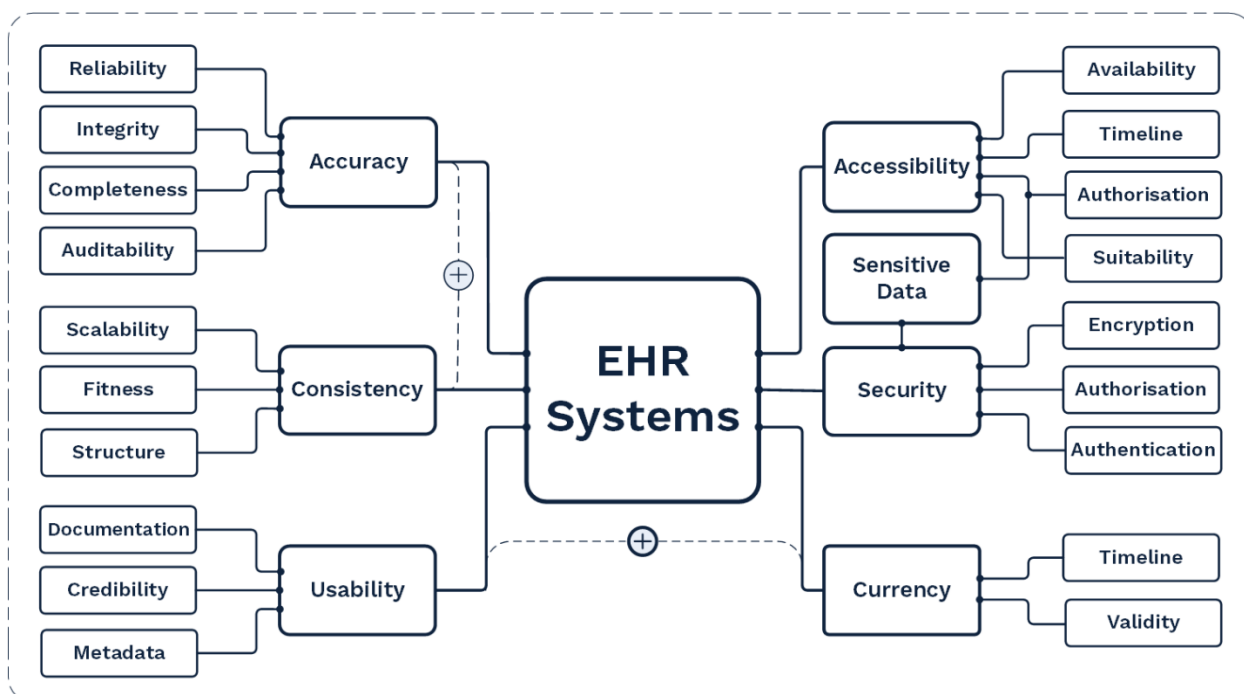


Figure 3.7: The electronic health record systems in large-scale DBMS (Saïod *et al.* 2017)

To keep up to the performance, the LSDB design should be in such a way that every EHR has to be a ripple effect that spreads beyond that specific user. It is not just about accessing some health data, it is also about analysing and ranking over HCOs network including billions of queries per seconds and nearly millions of row changes per second. The EHRs datasets are made even larger by taking into scope, alert and time. The EHRs may shard or splits its database into numerous distinct sections, to the sheer volume of the data it stores and it caches extensively to retrieve all these queries in a shorter time.

The most queries may never hit the database at all and only touch the cache layer as well as custom-built Flash-cache module for caching data on solid-state drives. LSDBs are faster, more reliable, allow replication and can do more complex queries. Therefore, the interface application is amenable to "sharing" and the database can cluster and be administered more easily or when everything is needed in one massive set of linked tables. Such is a complex authentication set up or it is a simple "one user" web application or the bulk of the data in binary objects or simple numbers and strings.

The benefits of EHRs are numerous when compared to the physician's time and finances, the health benefits for patients and the impact on the environment. The sparse health data may have multi-dimensions and it is practically challenging to investigate and analyse for different reasons, such as the heterogeneous features of the system, encompassing quantitative data as well as the categorical information.

This results in the random systematic error affecting badly and reducing the DQ. Most data integration methods are sufficiently robust to random systematic error for large datasets of input and process. This is commonly identical to bring them on the same scale when using pre-processing principal component analysis and the data simplification algorithm (Peter *et al.* 2012).

### 3.3.3 Concept of Electronic Health Data Level

Electronic Health Data (EHD) refers to health information converted into a binary form that is efficient for processing and level refers to a classification that describes the nature of information within the values assigned to variables. The EHRDL can be grouped into two categories as follows:

1. **Administrative level data:** Administrative data refers to EHRs generated by Healthcare Systems (HCSs), whether through providers, HCOs, physicians, laboratories, pharmacies, Government Medicare and Healthcare Risk Managers (HCRM). Often this is the administrative data used to evaluate the quality of

healthcare including the demographics, diagnoses and procedures code. With some exceptions, administrative data allows limited insight into the quality of processes of care, errors of omission or commission and the appropriateness of care.

**a) Demographic data:** Demographics health data is defined as a patient identity record that specifies the patient's particular medical record, such as the identification number, name, surname, age, gender, telephone number and address, which represent specific geographic locations and are often associated with the time. When the census assembles data about patient ages and genders, this is an example of assembling information about demographics. This is the first encounter patient record captured by administrative staff. Demographic data uses for various analysis for bivariate associations between demographic and health characteristics and mean medical scepticism scores for each question and the summary medical scepticism scores.

**b) Diagnoses data:** Health diagnoses data is a specific type of data used in the investigation and determination of identifying a particular disease or condition and explains a person's symptoms and signs. It is most often referred to as diagnosis with the medical context being implicit. The information required for diagnosis is typically collected from a history and physical examination of the person seeking medical care. Often, one or more diagnostic procedures, such as diagnostic tests, are also done during the process. Sometimes the posthumous diagnosis is considered a kind of medical diagnosis.

**c) Procedures code:** Procedure codes are a sub-type of medical classification used to identify specific surgical, medical or diagnostic interventions. The health procedure code is a set of healthcare procedure codes based on the American Medical Association's Current Procedural Terminology (CPT). The structure of the codes will depend on the classification; for example, some use a numerical system, others alphanumeric (HCPCS 2017).

**2. Clinical data:** Clinical data refers to health data collected during course ongoing patient care or collected during the formal clinical trial programme. Clinical data falls into six major types:

- a) Electronic Health Records (EHRs):** The purest type of electronic clinical data which is obtained at the point of care at a medical facility, hospital, clinic or practice. Often referred to as the Electronic Medical Records (EMRs), the EMRs are generally not available to outside researchers. The data collected includes administrative and demographic information, diagnosis, treatment, prescription drugs, laboratory tests, physiologic monitoring data, hospitalisation and patient insurance.
- b) Administrative data:** Often associated with electronic health records, this is primarily hospital discharge data.
- c) Claims data:** Claims data describes the billable interactions (insurance claims) between insured patients and the healthcare delivery system. Claims data falls into four general categories, namely inpatient, outpatient, pharmacy and enrolment. The sources of claims data can be obtained from the government (for example, Medicare) and/or commercial health firms (for example, United HealthCare).
- d) Patient/disease registries:** Disease registries are clinical information systems that track a narrow range of key data for certain chronic conditions such as Alzheimer's disease, cancer, diabetes, heart disease and asthma. Registries often provide critical information for managing patient conditions.
- e) Health surveys:** To provide an accurate evaluation of the population health, national surveys of the most common chronic conditions are generally conducted to provide prevalence estimates. National surveys are one of the few types of data collected specifically for research purposes, thus making it more widely accessible.



- f) **Clinical trials data:** Clinical research data may be available through national or discipline-specific organisations. The level of access is likely restricted but available through proper channels.

### 3.3.4 The Electronic Health Records data exchange

DQ issues might include a patient incorrect unique identification number. Other examples include misplaced name, incorrect gender, incorrect date of birth, numeric diagnosis code written in text or saved wrong radiology image, incorrect inserting standard code, such as the National Drug Catalog (NDC) for drugs and derailing bulk analysis (for example. ICD10 code: International Classification of Diseases Tenth Revision or CPT code: Current Procedural Terminology). The EHRs information exchange systems architecture is shown in figure 3.8 (Saïod *et al.* 2017), as below:

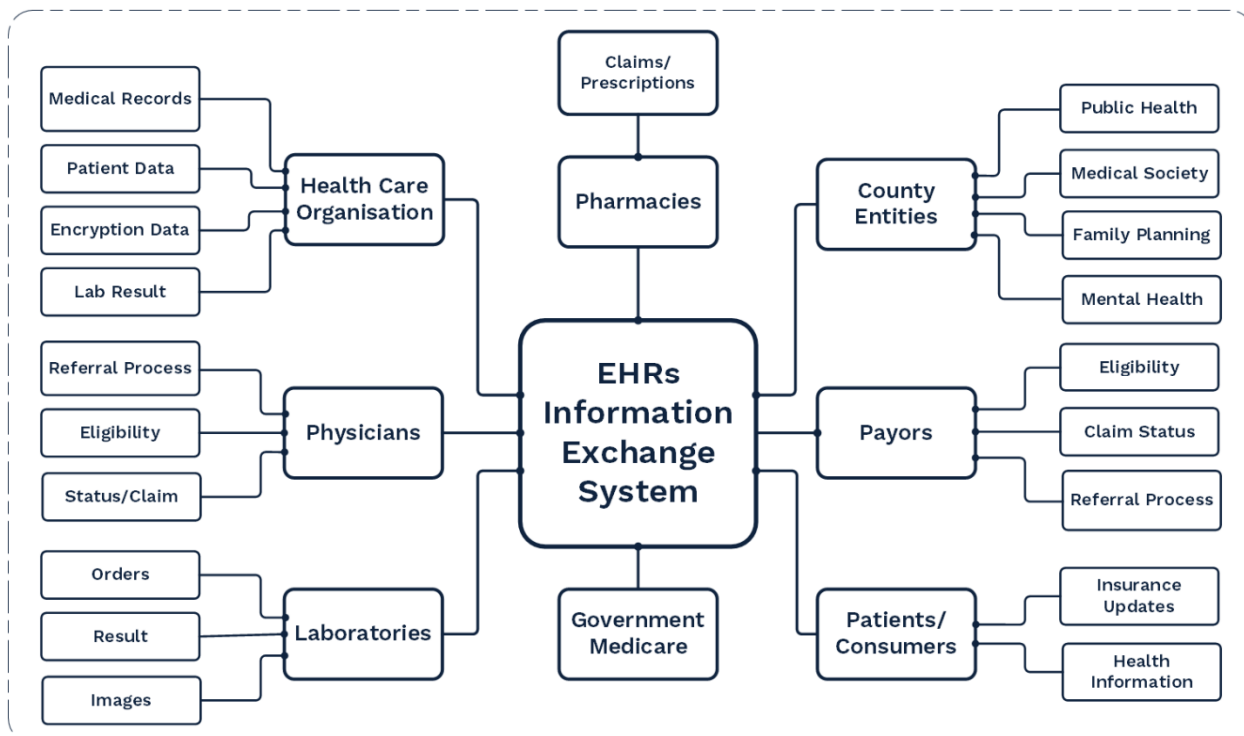


Figure 3.8: EHRs information exchange systems architecture (Saïod *et al.* 2017)

DQ refers to the concepts with immensely large-scale multi-dimensional in DBMS, which include not only data search, validation, extract and verification, but also the

appropriateness of use to take one even further beyond the traditional concerns with the accuracy of data. The EHR systems design, data structure, aggregation algorithm, simplification methodology and reporting mechanisms highly reflect on DQ.

### **3.3.5 EHRs Data Quality and benefits**

Quality data, appropriate for use, comprise characteristics that include completeness, uniqueness, consistency, accuracy, validity, correctness and accurate timelines. The quality of data can be analysed from multiple dimensions. One such dimension is a measurable DQ property that represents some aspect of the data accuracy and consistency that can be used to guide the process of understanding quality (Nuno *et al.* 2015). Though EHRs data quality is often only considered within the narrow scope of data verification and validation, it should also concern equally critical aspects of assuring that EHRs data is appropriate for a specific use. Alternatively, the quality of data is comprehended as in high demand, according to this denomination, as the volume of data increases and the question of internal consensus within data become significant, regardless of its appropriateness for use for any particular external purpose. Even when discussing the similar set of data used for the same intention, confluence's prospect on DQ can often be in uniqueness. Some information quality problems may arise from when the raw data is collected until it becomes useful information.

The majority of EHRs data is captured by a large number of individuals from heterogeneous sources and data exchange accessed these days to index text object use data rescue systems devised. This is due to unit measurement without different definitions and it may be captured in the EHR system. It will be absolutely impossible or may not be a comparative and assessment to interpret that which is being reported by another clinician when validated psychometric scales to assess patient status are not used. The objective deficiency of these problems classifies idiosyncratic DQ features.

The data inconsistencies can be identified directly, which can lead to inaccuracies and bias, as the data is collected geographically and over time and might be adjusted to differences over to account for unequal measures over time (Bruce *et al.* 2013). Schaal *et al.* (2012) motivate the adoption of accessible data based on its definition of data that comprises clarity and consistency.

Despite the intimidation posed during data storage and transmission, EHRs are seen as a hopeful accomplishment to problems in EHR management. One of the key barriers is to optimally use routinely collected data, as the increasingly poor quality remains in the data. This raises the need for automating the mechanisms used to measure DQ and semantic interoperability. This framework is a result of filtering the existing DQ dimensions in many research sources and checking its suitability to the nature of e-health systems.

Many research sources verify its praiseworthiness to the behaviour of e-health systems, this skeleton is an outcome of percolation subsisting of DQ dimensions.

### **3.4 Electronic Health Records data structure**

Most HCOs data is highly structured and heavily depend on claims data, but the prosperous scope provided by health data is absent. Furthermore, leveraging health data basically depends on vendor-delivered implementation, communication, such as a Continuity of Care Documents (CCDs), which are the few analytics applications. They also stick limitations via both design and integration that make them insufficient for populating health and productivity analytics, until CCDs offer a consolidated and expedient way to implement electronic health data.

Methods for data capturing in EHR systems include direct capturing, capturing on the screen template, scanning handwritten documents or importing transcribed data from other information systems in different data exchange formats, such as JSON, XML, CSV, TXT, REBOL, Gellish, RDF, Atom, YAML and other data exchange technologies. Each one of these methods has strengths and weaknesses that may

have an impact on DQ. Only the quality of the data can ensure that healthcare providers have confidence in EHR systems to deliver the best service possible. Figure 3.9 (researcher source) shows the EHRs database structure, as below:

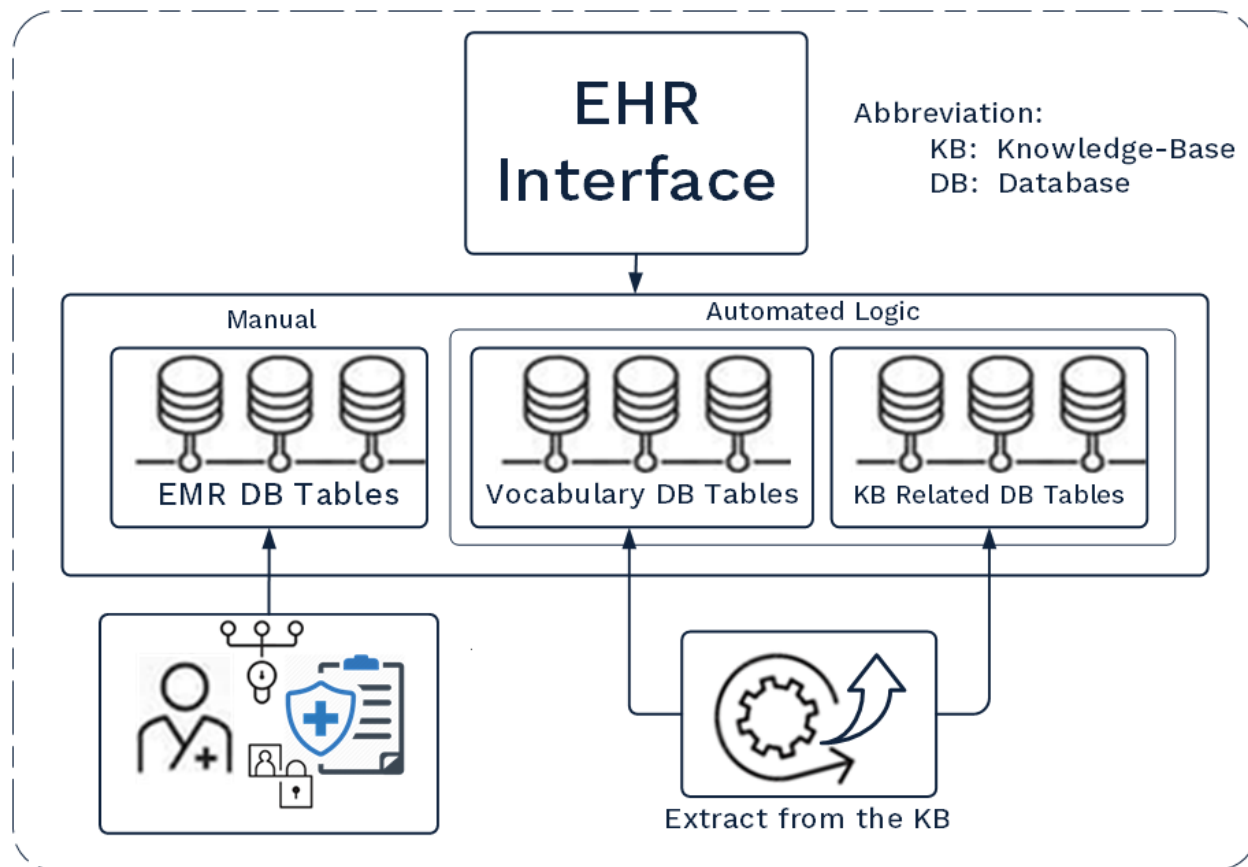


Figure 3.9: EHRs Database Structure (researcher source)

To capture, store as well as develop and implement structured health data to avoid DQ gaps, the integrated analysis programme must be used. To effectively solve the challenge of DQ gaps, further relevant points need to be discussed. Valid data capturing techniques require that when a clinical encounter takes place and a provider and/or automated systems insert information into an EHR system, it is, for example, captured accurately into the EHR of a patient. Valid data structures need to be used both in the way in which data is captured as well as the storage of the data in an appropriate format and location.

If an integer is captured in a VARCHAR field, its feasibility for reporting, analysis and quality will be reduced, even if it is captured in a structured field. If the

template or screen structure is not properly mapped or configured in the database, the value may still be stored in an incongruous location. The analysis or reporting purposes for this information is extracted from the server and made instant and available to the authorised user.

There is no need to include the extraction of all pertinent information when it back ends the database connection. The way in which data will extract or data will select from the query table, are the key factors in how the exchange of creating the dataset will take place and the exchange tackle impacts on the application quality of the outgoing DQ. It is of importance to identify the point at which DQ gaps are introduced. This will in-turn, lead to focused initiatives to eliminate such gaps.

Data security is a key concern in healthcare interoperability whether paper-based or electronic health records. According to human rights, every individual can keep personal data confidential and not being disclosed for surveillance or interference from another organisation or even to the government.

All confidential information should be protected and encrypted, whereas data that is shared is a result of the clinical relationship (Rinehart *et al.* 2006). Patient data can only be released when the patient gave his/her consent or when stipulated by law. Information may disclose information sharing only if the patient is unable to do so because of age or mental incapacity; then the data sharing decision should be made by the legal representative or legal guardian of the patient. The information is considered confidential and must be protected when a result is queried in clinical cooperation. The identity of the patient cannot be ascertained when information is populated; for instance, the number of patients with HIV in a government hospital does not fall in this denomination (Rinehart *et. al.* 2006).

### **3.5 Electronic Health Records data synchronisation**

The gradual harmonisation of the data over time, so-called data synchronisation is the procedure of establishing consistency between information from a diverse source to the destination data server and vice versa. Considering large-scale

computing, the data flow between multi-user clinicians and one central server definitely entails a multi-way synchronisation model, as the server must send a client the data previously created by other clients. Figure 3.10 (Saïod *et al.* 2017) describes the two-way data synchronisation workflow model, as follows:

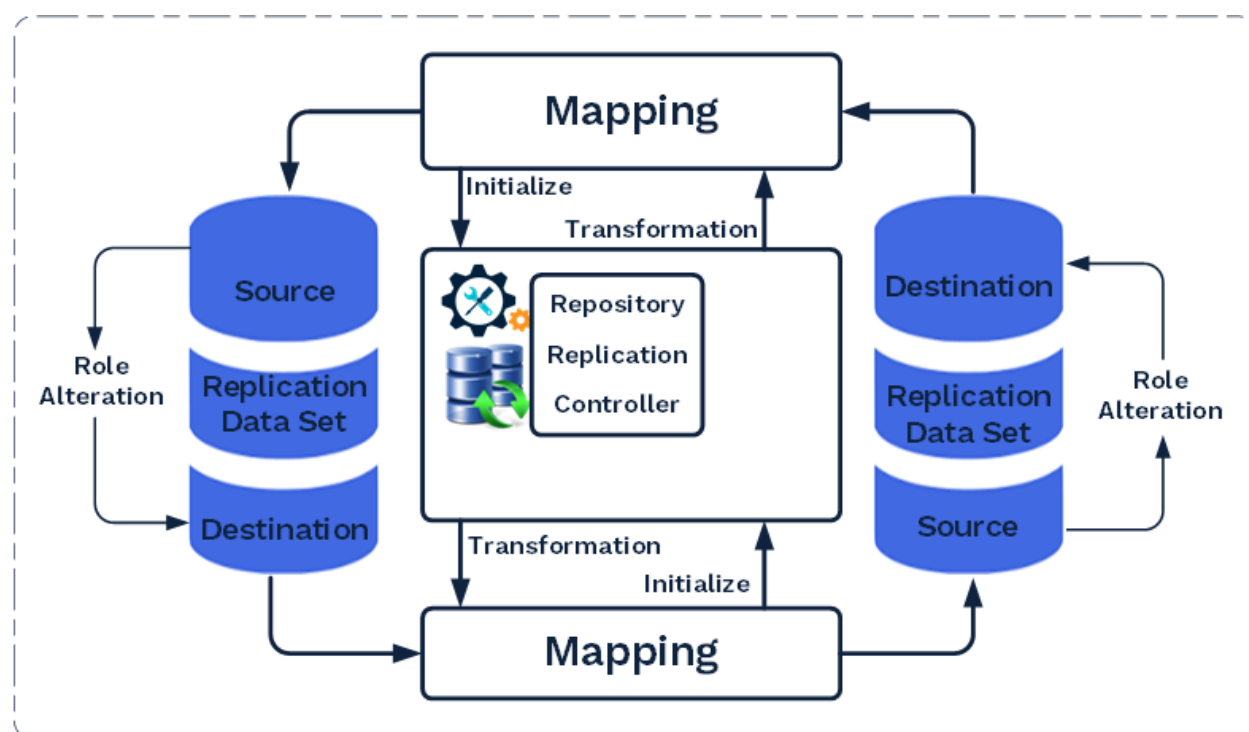


Figure 3.10: Two-way data synchronisation workflow (Saïod *et al.* 2017)

The patient is not conscious of the entire data structure as this is the server data structure algorithm to figure out the modifications of so-called rights management rules. In many cases, data should be available in more than one directory server using three different techniques for achieving this. This includes a directory replication protocol, direct synchronisation between a pair of directory servers and indirect synchronisation between two or more servers. Replication has the best operational characteristics but the lowest functionality, which can differ between the various techniques.

The majority of replication techniques entail indirect synchronisation, which has the highest functionality and the poorest operational characteristics, whereas direct synchronisation is intermediate. Data in one directory server needs to be

made available in another directory server for numerous reasons. These include availability, load sharing, locality, reaching data on other servers, data access restrictions as well as data mapping.

Generally, synchronisation between a client and a server follows five steps:

- a)** The data administrator rules prepare the data for a “go/no-go” response when the authorised user initialises the request;
- b)** The server algorithm rules check the user authentication to accomplish whether synchronisation is required and finally checks for all possible conflicts;
- c)** The authorised user submits the data trees;
- d)** Before to nodes and stores data, the server assigns new IDs to trees;
- e)** The server uniquely identifies the data in the network to these collective IDs and the collective database;

It should be noted that only the authorised user allows viewing the sent data to the client according to the accurate management rules. Finally, before replacing the local IDs with collective ones and storing the new trees to the server, the authorised user updates to a local database (Consuela *et. al.* 2005). The overall practice shows that EHRs and the ability to exchange health information electronically can help HCOs to provide higher quality and safer care for patients while creating tangible enhancements for healthcare. EHR systems thus enable healthcare providers to not only improve the care management plan for their patients but to also provide improved healthcare through accurate, up-to-date and complete information sets about patients.

This enables quick access to patient records for an improved and coordinated care plan. This is achieved by securely sharing electronic information with patients and other clinicians that in turn helps providers to diagnose more effectively and thus reduce medical errors. It contributes to the provision of safer care, improved

patient and provider interaction and communication. Add to this healthcare convenience, more reliable prescribing, promotion of legible and complete documentation supported by accurate, streamlined coding and billing. Figure 3.11 (Saïod *et al.* 2017) describes a typical model of the data synchronisation architecture of HCOs, as follows:

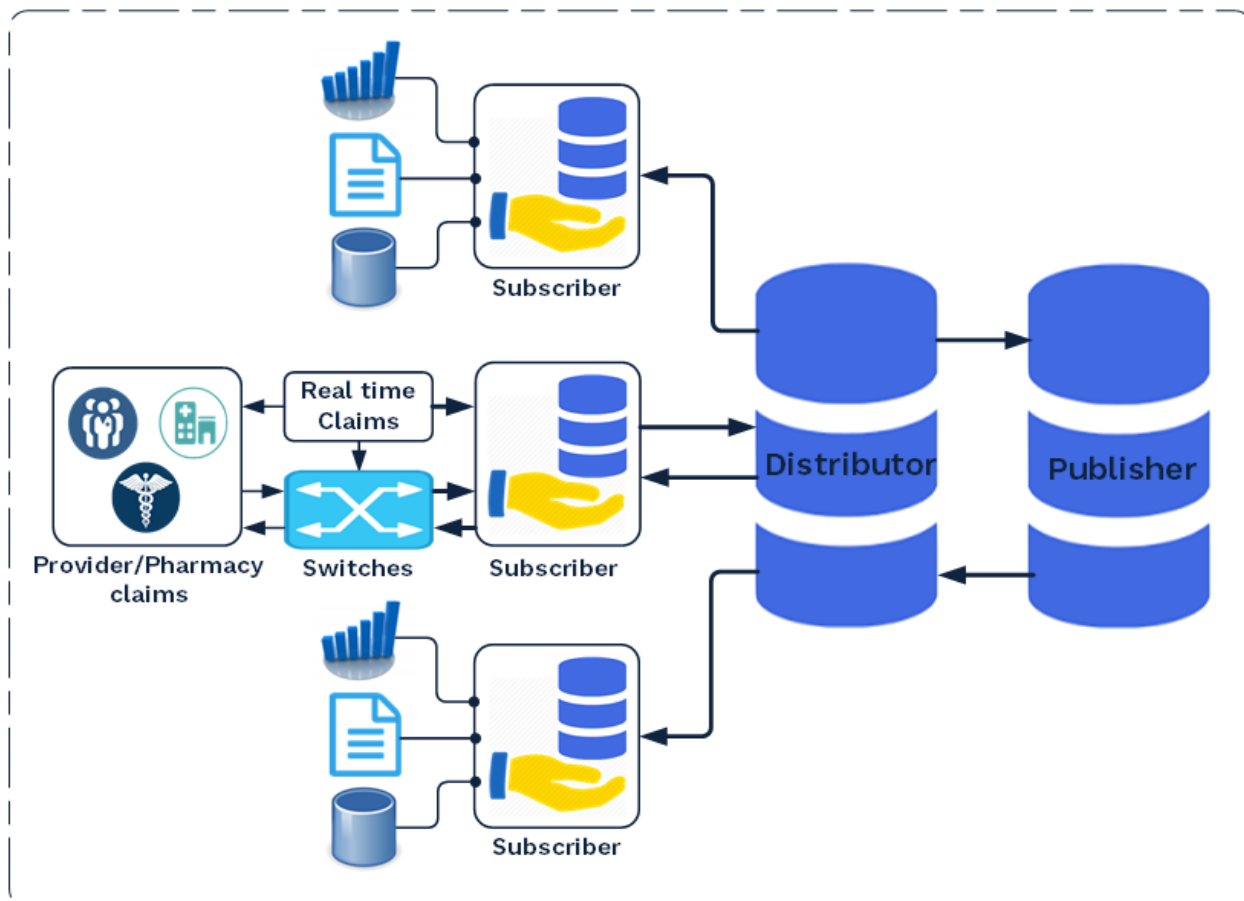


Figure 3.11: A healthcare risk manager's organisation data synchronisation architecture (Saïod *et al.* 2017)

Other improvements include enhancing privacy and security of patient data, helping providers improve productivity and work-life balance, enabling providers to improve efficiency and meet their business goals, reducing costs through decreased paperwork, improved safety, reduced duplication of testing and improved healthcare services.

It, therefore, became necessary, to implement adaptive, interoperable EHR systems to improve the quality of data, which addresses the current EHRs



challenges. As a result, the proposed solution has focused on a novel approach based on different methods and existing systems, to reduce the challenges of EHRs and DQ. EHRs technology will be applied to not only perform the function of receiving and displaying information but to automatically and accurately extract information from diverse heterogeneous data sources using healthcare services.

For data to equivalently match two concepts across different data sources and automatically resolve any inconsistency arising from multiple data entities is the challenge of EHRs. The important expected contribution of this study had to realised a method to improve EHRs data quality from heterogeneous and inconsistent data sources. The key outcome of EHRs has discovered a new merged concept by finding consensus among conflicting data entries.

### **3.6 Electronic Health Records data collection**

Data collection is defined as the on-going, systematic assembling and measuring of information, analysis and illustration of health data necessary for integration, implementing, designing and evaluating public health prevention programmes, which then enables one to answer relevant questions and evaluate outcomes (WHO 2016). The HCOs collect data to observe health to handle subsidies and services as well as inform bankroll and resource allocation, identify and appraise healthcare services, inform the development of health policies and interventions, assist clinical decisions about patient care and meet legislative requirements.

Surveillance is undertaken to inform disease prevention and control measures, identify health emergencies as an early warning system, guide health policies and strategies and measure the impact of specified health interventions. Few people are, for example, dying from infectious diseases, but due to changing patterns of physical activity and the expenditure of drug, tobacco, alcohol and food more people are suffering from chronic diseases. The survey process is conducted to maximise accuracy and participation to generate statistics.

Using a different data collection algorithm, these statistics are generated from diverse sources, including household surveys, routine reporting by health services, public registration and censuses and disease observation systems. HCOs involve a different civil dataset and private data collection systems, including clinical surveys, administrative enrolments, billing records and medical records used by various entities, including hospitals, physicians and healthcare plans. The possibility of each to facilitate data on patients or enrolled data on race, ethnicity and language are also collected to some extent by all these suggested entities (Citro *et al.* 2009).

Data breaches in healthcare come in a variety of forms such as different healthcare capturing and storing methods as well as technology used such as excel, access, SQL and Oracle. Manual data collection from ward-based sources captured only 376 (69%) of the 542 in-patient episodes, which were captured by the hospital's administrative electronic patient management programme. Administrative data from the electronic patient management programme had the highest levels of agreement with in-patient medical record reviews for both lengths of stay (93.4%) data and discharge destination (91%) data (Sarkies *et al.* 2015). Currently, fragmentation of data flow occurs because of the silos of data collection. In HCOs data is often collected by clinical assistants, clinical nurses, clinicians and practice staff.

Prospective observational studies compare the completeness of data capturing and the level of agreement between three data collection methods:

- a) *Manual data collection from ward-based sources or paper-based:*** Paper and pencil, surveys, chart abstraction and weekly return card;
- b) *Administrative data from an electronic patient management programme:*** Dedicated electronic data collection systems, EHRs-based, images and audio and video recording (qualitative research);
- c) *Historical data:*** Inpatient medical record review for hospital length of stay and discharge destination;

With specific diseases in clinical and genomic research, the objective of EHRs is to generate large cohorts of patients. The electronic phenotype selection algorithms are to find such cohorts a rate-limiting step as a development.

This study evaluated the portability of a published phenotype algorithm to identify Rheumatoid Arthritis (RA) patients from electronic health records at three different institutions, using three different EHR systems. EHR systems are seen by many as an ideal mechanism for measuring the quality of healthcare and monitoring ongoing provider performance. It is anticipated that the availability of EHRs-extracted data will allow quality assessment without the expensive and time-consuming process of medical record abstraction.

A review of the data requirements for the indicators in the Quality Assessment Tools (QAT) system, suggests that only a third of the indicators would be readily accessible from EHRs data. Other factors such as the complexity of the required data components, provider documentation habits and EHRs variability make the task of quality assurance more difficult than expected. Accurately identifying eligible cases for quality assessment and validity scoring, those cases with EHRs extracted data will pose significant challenges but could potentially lower costs and therefore expand the use of quality assessment. Improving the data collection process across the healthcare system is one of the key challenges to improve DQ.

### **3.6.1 Improving the Electronic Health Records data collection process**

Overreaching opportunities abound for increased quantities of EHRs, for improved quality of data and for new data components that were once considered too burdensome or expensive to capture. This wealth of EHRs can be used to validate or calibrate health demand models, for inpatient care information systems analysis and for modelling mobile source emissions across a healthcare network. These data collection and processing advancements are, however, costly and should be implemented with caution. The focus of healthcare is on data accuracy issues pertaining to the mechanism chosen for data collection and data processing, using

EHRs technology. Vast amounts of technology spreads exist throughout the transportation field, which are automating numerous manual data collection processes. Figure 3.12 (Saïod *et al.* 2017) depicts the data flow control system in large-scale DBMS, as follows:

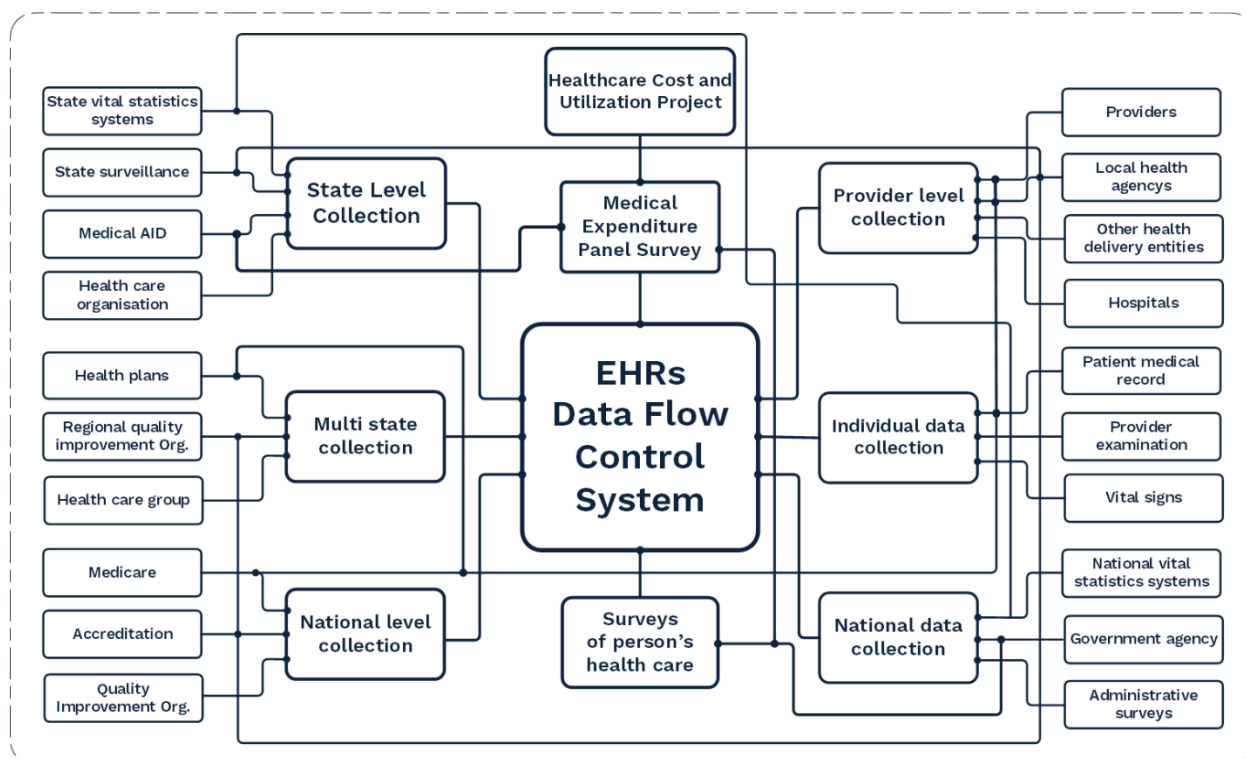


Figure 3.12: The data flow control system in large-scale DBMS (Saïod *et al.* 2017)

These advances generally reduce labour costs and manual capturing errors. Automation of survey data collection allows EHR systems to collect new data streams without increasing respondent burdens. When data is combined from diverse heterogeneous sources, the data that are syntactically identical (same format, same units) can show important inconsistencies, as data components that supposedly represent the same concept, actually, represent different concepts at each site. The term semantic variability expresses the data variability caused by differences in the meaning of data components. Differences in data collection, abstraction and extraction methods or measurement protocols can result in semantic variability.

Failure to distinguish between fasting and random blood glucose, finger-stick or venepuncture sampling or serum or plasma measurements would, for example, result in glucose values that do not represent the same concept. Semantic variability is difficult to detect using single-site data alone because data semantics tend to be consistent within an institution. Only when the data is combined from multiple heterogeneous sources can such semantic differences be detected. The above discussion regarding the challenges faced by various healthcare professionals and healthcare institutions highlights the importance of accurate data capturing and DQ to overcome HIT constraints and minimise respondent and organisational resistance.

The integration of data systems has the potential to streamline collection processes, so that data can be reported on easily and that an individual would not need to self-identify race, ethnicity and language requirements during every health encounter. Integrating the various data systems, enhancing legacy HIT systems, implementing staff training and educating patients and communities about the reasons for and importance of collecting these data can help improve data collection processes.

Not all data systems capture the method through which the data were collected and some systems do not allow for data overrides. The interoperability of data systems may, for example, prohibit a provider from updating a patient's data, which were provided by the patient's healthcare plan. Self-reported data should, therefore, trump indirect, estimated data or data from an unknown source. Ways of facilitating this process logistically warrant further investigation. Data overriding should be used with caution, as overriding high-quality data with poor-quality data reduces the value for analytical processes.

Currently, one specific data collection effort under evaluation for automation is the patient update survey, which traditionally has been administered to obtain a comprehensive up-to-date patient history. The fundamental concern associated with the need to change medical practice tendencies and the way of interacting with patients created barriers to EHRs implementation and use. The adaptation to

EHR systems was also considered a major threat to practitioner professionalism, because of the corresponding requirements for providers to adhere to the requirements of the EHRs, including electronic documentation and compliance with standardisation guidelines.

Even though current data collection methods are subject to numerous errors, the survey data collected are used to forecast regional health data such as demographics, hospital admissions and discharge notes, medical history of patients, improvement notes, outpatient clinical health notes, medication prescription records, medication and allergies, immunisation statuses, radiology reports and images, laboratory data and test reports, essential symptom, personal statistics such as BMI, blood pressure, age and weight information.

In addition, the availability of EHRs databases makes the automated processing of such data feasible. With the application of these technologies, however, care and caution should be applied when using and interpreting the datasets obtained from the data collection method used.

### **3.7 The barriers and threats of Electronic Health Records**

The overarching barriers to the EHRs framework are to tackle the indigent quality of data to provide a single, centralised and homogeneous interface for users to efficiently integrate data from diverse heterogeneous sources. DQ issues may arise when capturing raw data into the EHR systems. The data flow process has several factors that influence the quality of information obtained from such datasets at a later stage. The purpose of the data collection processes is DQ management functions which include the data flow process application, as well as data, accumulate, warehousing process systems used to archive data and analysing the process of translating data into meaningful information.

The DQ may seriously affect patient care and even could lead to the death of the patient. This is the key challenges of eradicating treatment errors in the health service process. As patient safety is the key issue in healthcare service, using

effective EHR systems integration and implementation can improve the DQ to reduce medical error. The main consideration for health data includes data accuracy and accessibility, as well as data comprehensiveness, currency, consistency, granularity, precision, relevancy definition and timeliness. DQ will empower the tendency of EHR systems, this emphasises the magnificence of implementing a design-oriented definition. The dimensions of the existing EHRs framework are basically based on historical reviews, understanding intuitive and comparative experiment.

The EHRs structures of orientation usually vary from framework to framework. For example, the actual use of the data depends on the definition of DQ. It, therefore DQ also depends on the application type and that which may be deliberated in one application as good quality but may not be good for another. DQ has emerged as a crucial issue in many application domains. The objective of DQ becomes even more important in the case of patients who need to be identified and notified about important changes in drug therapy or in the case of merging systems of different and similar organisations. The consolidation of information from diverse sources to provide a unified view of an organisation's data assets is technically challenging. This difficulty involves the way to practically combine data from disparate, incompatible, inconsistent and typically heterogeneous sources.

The other difficult objective in EHR systems is that data has a structure, which is usually complex and cannot be treated as a simple string of bytes. Often data inconsistency occurs because the data structures may depend on other structures, therefore on a distributed system, this kind of data management is very difficult. Another significant aspect of a health data integration system is data mapping. The system must be able to materialise data that are mapped from a diverse source.

Optimally using routinely collected data increases poor quality data, which automatic mechanism would raise the need of the semantic interoperability as well as quality data measurement (Liaw *et al.* 2012). Quality improvement and error reduction are two of the justifications for healthcare information technologies. Despite their concerns, HCOs are generally very interested in adopting and

implementing EHR systems. A major concern of the success of implementation is the large gap between planning for the introduction of EHR systems and medical maintenance systems in hospitals. The primary purpose of the successful EHR systems implementation depends on these application systems and the maintenance of the application significantly to achieve the desired and expected benefit.

The real barriers causing this gap may not be the availability of technology to the HCOs, as information systems are actually becoming available almost everywhere, but the deficiency in providing proper support before, during and after the implementation of the EHR system. The financial constraints are another important matter of the migration from the paper-based health record to an EHR system. Generally, two principal barriers and challenges in the method of prosperous EHR system integration are, namely:

*a)* Human barriers (for example, professional and beliefs);

*b)* Financial barriers (for example, available money or funding opportunities);

The human factors become even more important as the benefits are only anticipated after the successful integration and implementation of the EHR systems. Information security is most important for quality healthcare service. It improves the potential of EHRs as well as accuracy, accessibility, productivity, efficiency and to reduce the costs of healthcare and medical errors. Most HCO administrators are aware that it is time-consuming to migrate from a paper-based record system to an EHR system.

It is important to change the provider and healthcare practitioners' behaviours with regard to electronic healthcare systems, but time is needed. A few factors things also need to be addressed regarding the successful implementation of an EHR system, such as attitudes, impressions and beliefs. The most important factor is that it is essential to understand the reasons for and the purpose of the implementation of EHR systems in the whole subject (Pagliari *et al.* 2005). Research and statistics showed EHRs estimated potential savings as well as the costs of the



widespread adoption of EHR systems. Important health and safety benefits were modelled and concluded that the effective EHR system implementation and networking, could improve healthcare efficiency and safety.

It also showed that Health Information Technology (HIT) could enhance the prevention and management of chronic diseases, which could eventually double the savings while increasing health and other social benefits. The feasibility of introducing an EHR system to improve DQ is the meaningful association between the heterogeneous data source and the integration into HCOs to improve healthcare service.

The integrity constraints are specified in the global scheme of data mapping, which can be used to promote EHRs data quality as well. The uncertainties are the other important integration aspect in EHRs that should be minimised to improve DQ. The most important barriers and constraints to high-quality datasets to be solved is the integration of EHR systems and the electronic health record, achieving the maximum benefit of the healthcare services. Finally, it is noted that query answering in the context of data exchange, contributes to DQ.

### **3.8 Summary**

This chapter provided an overview of various EHRs concepts including the advantages and disadvantages relevant to the research presented in this study. The chapter began with an overview of the general concept that is related to eHealth and EHRs. This was followed by an overview of various definitions of EHRs and its application areas. A detailed discussion of different application systems and platforms were provided. The chapter also looked at the advantages and disadvantages of the different EHR systems. The chapter concluded by highlighting the issues to maximise DQ to achieve all possible benefits of EHRs.

### 3.9 Conclusion

The DQ issues hold huge potential for EHRs integration in HCOs as well as transforming the healthcare domain from traditional and paper-based health record to electronic-based health records. The DQ contributes significantly to eHealth through the use of mobile, distance health nationally and globally. Sharing and accessing EHRs can be developed significantly to improve the quality healthcare service.

Based on the EHRs landscape in LSDB, as presented in this chapter, **the lack of DQ issue in EHRs integration has identified as no common methodological approach is currently in existence to effectively solve every health data integration problem.** The challenges of DQ raise the need in this study for interventions to overcome these barriers and challenges, including the provision of EHRs as they pertain to DQ. These will combine features to search, extract, filter, clean and integrate data to ensure that users can coherently create new consistent data sets. Figure 3.13 demonstrates the outcome of Chapter Three, as below:

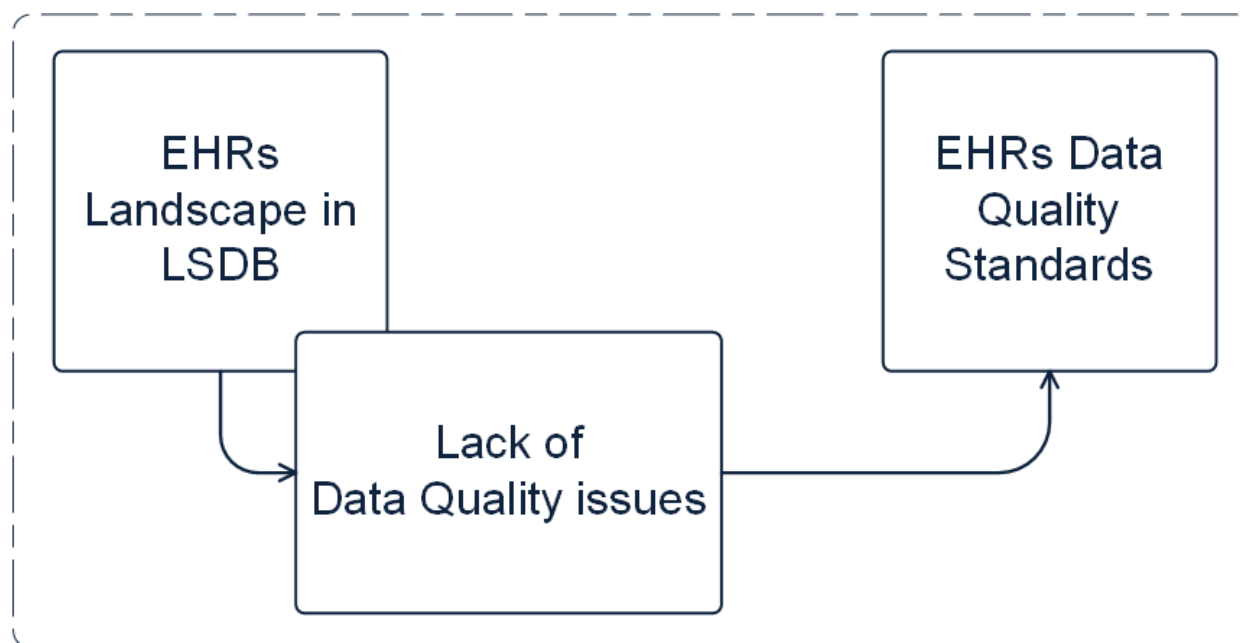
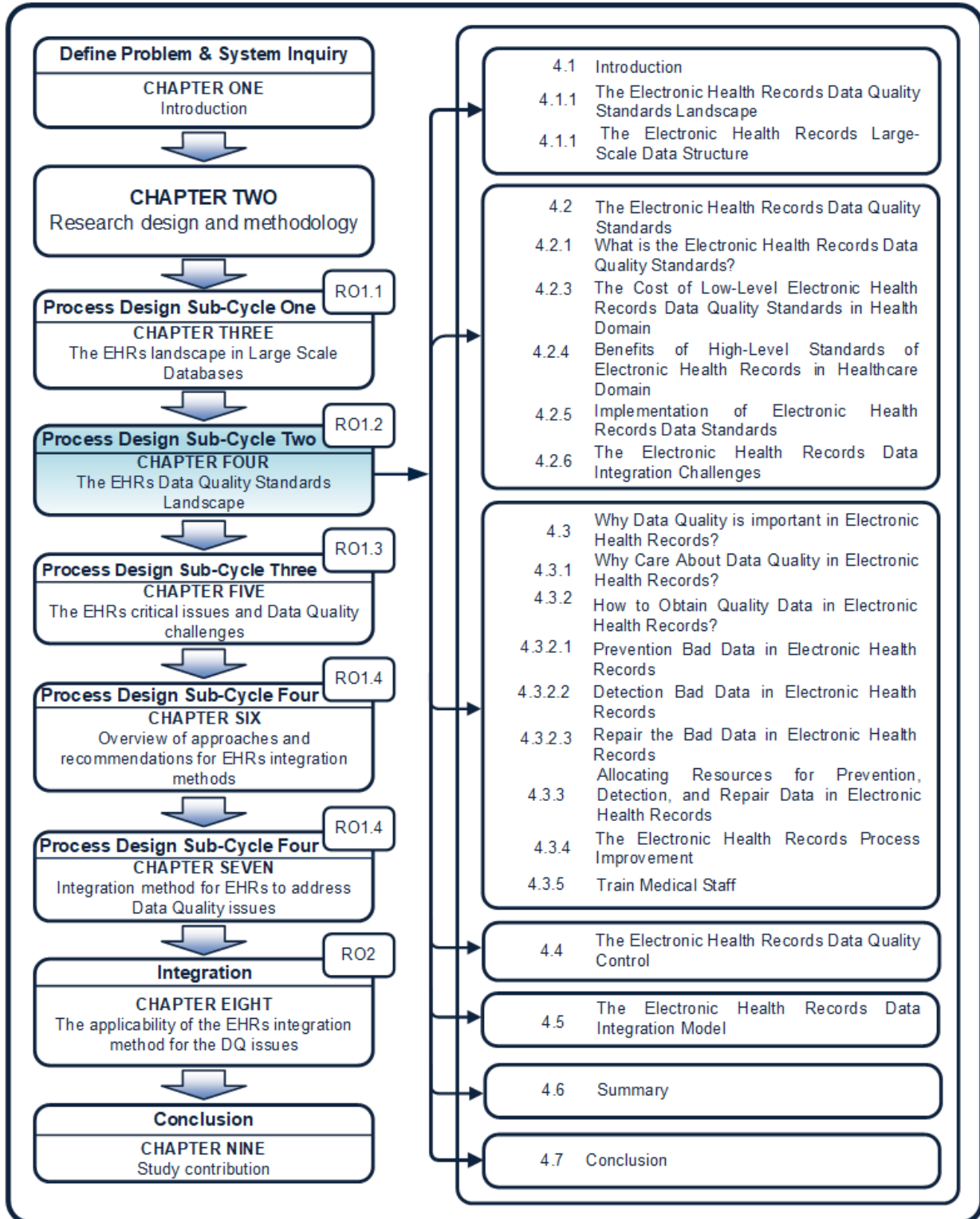


Figure 3.13: Outcome of Chapter Three

Therefore, as illustrated in figure 3.13, the review of the DQ issue is the next logical step to perform in this study. Chapter Four presents a review of the EHRs DQ standard to investigate DQ issue in EHRs.

## CHAPTER FOUR: The EHRs Data Quality Standards Landscape



Outline of the Chapter Four

## CHAPTER FOUR

### 4.1 Introduction

Caring about DQ is key to safeguarding and improving the quality of healthcare services. Data standards are the principal essential component for DQ necessary for data flow through the EHRs infrastructure. Although diverse health data is needed for clinical care, patient safety and quality improvement for integration, there is as yet no common methodological approach to integrate this data easily and economically from diverse heterogeneous sources, despite the availability of the ICT to support such data exchange. The purpose of this chapter is to provide a foundational material to determine the DQ standards, as well as ways to safeguard and improve the DQ entails in the healthcare domain. It maps to Sub-Cycle Two of the DSR process, described in section 2.3.2.2.2 and highlighted in figure 4.1.

Section 4.2 provides a detailed discussion of Data Quality Standards (DQSs) by defining its meaning, levels of standard, benefits, difficulty and complexity of EHRs and the associated DQ challenges that make the EHRs integration difficult. Section 4.3 discusses the necessity of DQ, the way to obtain quality data and prevent bad data.

After identifying when data are of high quality, the reasons are discussed for one to care about DQ. Section 4.4 discusses for one to care the DQ control of the EHRs whereas section 4.5 presents the EHRs integration model. The summary and conclusion of the chapter are provided in section 4.6 and 4.7 respectively.

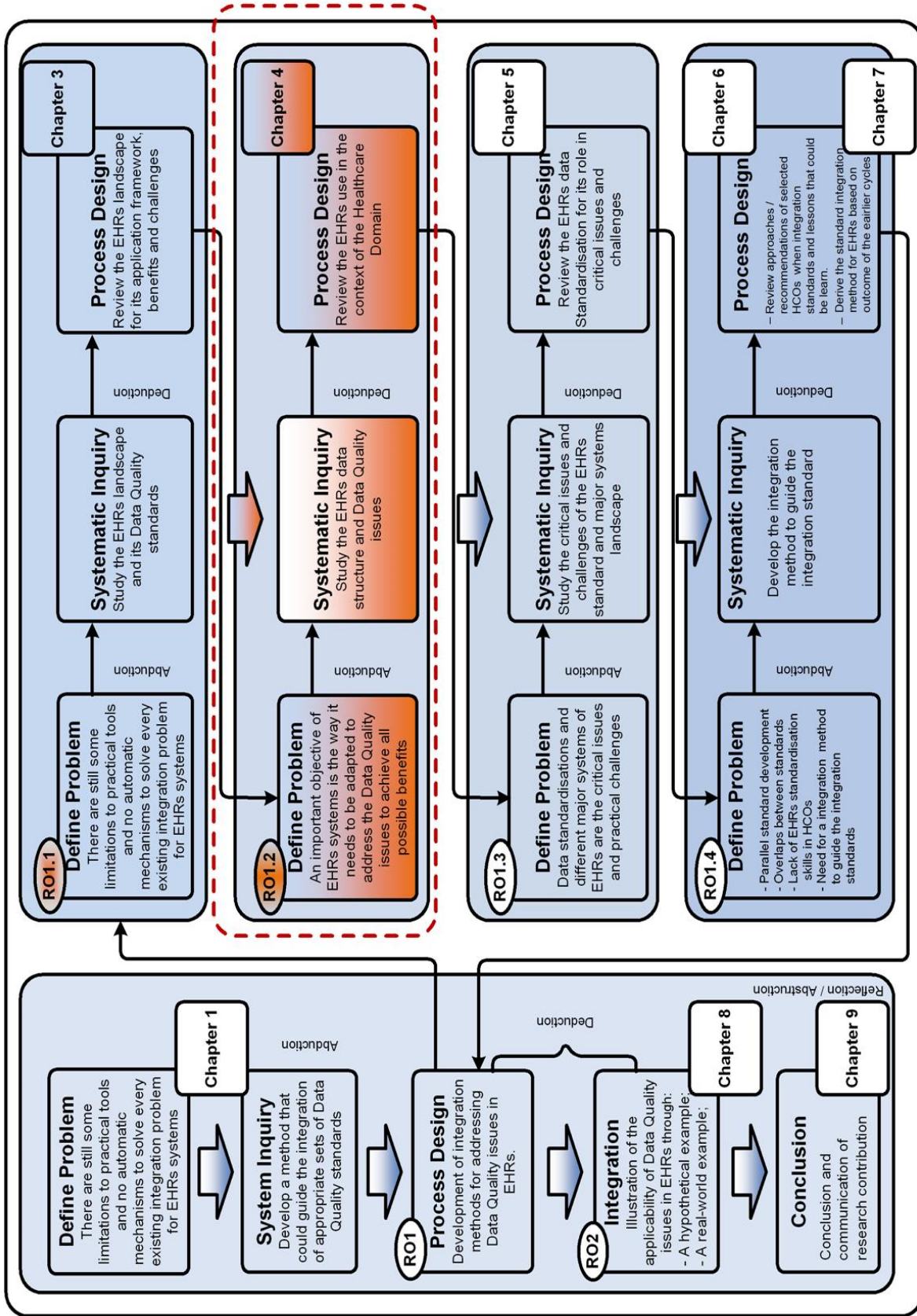


Figure 4.1: The position of Chapter 4 in the design science research process used in this study

### 4.1.1 The Electronic Health Records Data Quality standards landscape

DQ refers to a perception or an assessment of the condition of a set of values of qualitative or quantitative variables. Numerous definitions of DQ exist but data is generally considered high quality if it is “fit for [its] intended uses in operations, decision making and planning” (Baškarada *et al.* 2014). Complete and accurate data is called quality data that is appropriate for use in intended operational, decision-making and other roles. Therefore, DQ improvements the intervention that involve specific training on the importance of health data is regular reviews, audit information and feedback. The EHRs data quality standards landscape of the study are presented, followed by an outline of this chapter.

### 4.1.2 The Electronic Health Records Large-Scale Data Structure

Large-scale data becomes relevant for more and more organisations and move to new fields of applications where massive volumes of data are automatically generating continuously from diverse data sources and applications. When dealing with big datasets organisations face difficulties to be able to create, manipulate and manage the large-scale. Large-scale data is particularly a problem in business analytics because traditional approaches and procedures are not designed to search and analyse massive datasets.

Existing literature shows that several techniques, software, applications and major approaches currently exist to deal with the big data, which historically have faced DBMS. After a profound analysis of various cutting-edge commercial accomplishments existing on the software market and an intensive review of the literature, some limitations still appear with practical approaches for big DI. Integration is the exchange between different organisations and other important players that brings big data or functions from one data platform to another. Integration is important because when one looks at the quantity and diversity of data involved in the large-scale data domain systems, it would be virtually

impossible to process or analyse without breaking through the data silos. Figure 4.2 (Saïod *et al.* 2019a) describes the large-scale advanced analytic platform architecture, as follows:

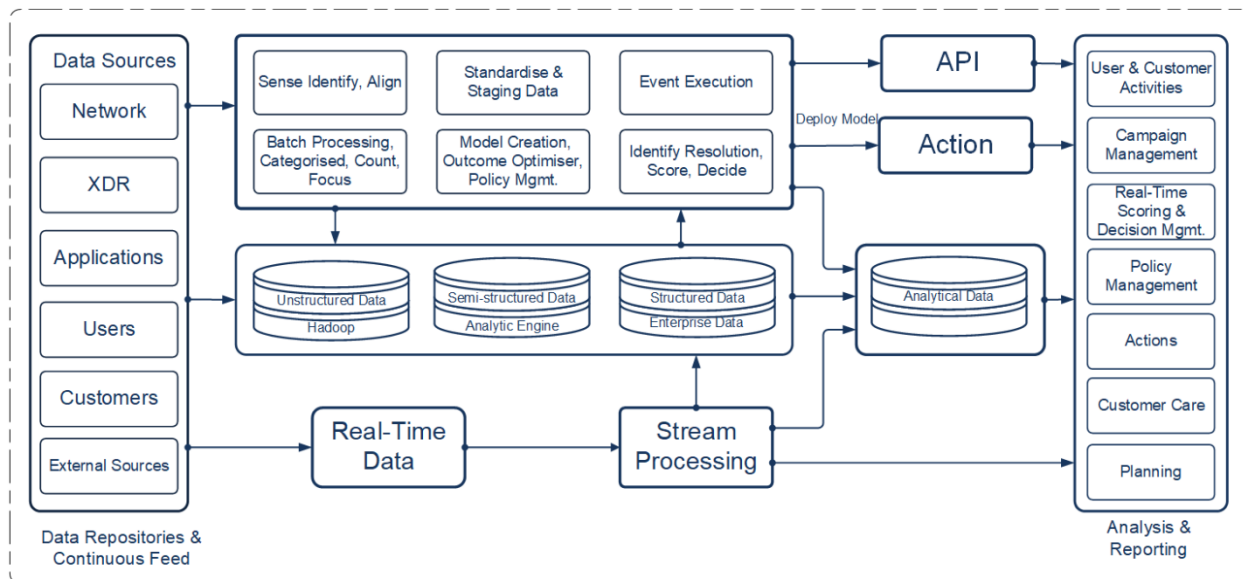


Figure 4.2 Large-scale advanced analytic platform architecture (Saïod *et al.* 2019a)

The same as traditional IT systems, the large-scale data uses completely different technical and semantic standards to depict and manage data. This makes it extremely difficult to correctly and simply integrate data from various conflicting systems. Big data stores in a linear record of each data info. It includes data types, pattern, length, demographics, progress notes, vital signs, history and indexes. By offering hybrid approaches, it helps to automate and streamline their workflow. Given that a large part of the picture exists within the big data, one would have expected that integrating into and out of the DBMS would have been easier for other sources.

However, for the majority of HCOs, it requires a skilled team of experts to bring the idea to fruition. This team typically includes a project manager, operational owner, systems administrator/network engineer, interface engine analyst, big data /application interface analyst, big data web service analyst, big data analyst and



support staff. In addition to orchestrating all the involved parties, it is also important to ensure that the HCOs is meeting all the electronic data standards interpretations.

## 4.2 Electronic Health Records Data Quality standards

Data Quality Standards (DQSs) refers to the convention defining the quality that has been set as “**the Conformance to Standards**”, so that appropriate use is achieved. DQSs belong to two criteria, namely HCO’s staff doing their work (**Conformance to Standards**) and in the way that the patient receives the quality healthcare service (**appropriate for use**). When these two criteria meet together, these two can yield efficient EHR systems that achieve the desired accuracy level and other specified quality attributes to achieve the maximum benefits of EHRs. As different HCOs have different standards and different major systems, this is one of the main challenges in developing efficient EHR systems, namely the inherent difficulty to coherently manage diverse heterogeneous sources.

Unfortunately, many HCOs data does not meet either of these criteria and EHR systems using the existing DBMS, present challenges, because of their incompatibility and sometimes inconsistency of data structures. With the wide availability of techniques and EHR systems in the software market, including many well-trained data analysts, a keen desire exists to analyse such DBMS in-depth.

As stated in section 3.7, the uncertainties are the other important integration aspect in EHRs that should be minimised to improve DQ. The most important barriers and constraints to promote high-quality datasets must solve the integration of EHR systems and electronic health records to achieve the maximum benefit for healthcare services. This chapter provides a detailed discussion of that which is meant by DQ and its issues in EHRs, its benefits and challenges, as well as the risk of having poor quality EHRs.

## 4.2.1 What are the Electronic Health Records Data Quality standards?

DQ refers to the pervasive appropriateness of a dataset as a possibility of its calibre to be spontaneously processed and analysed for related uses, generally by a DBMS, data warehouse or data analytics system. DQ is one of the important essential parts to HCOs for numerous reasons. High-quality data can be an essential quality healthcare service, which is a reputation for world-class healthcare services. In contrast, poor-quality data can reduce patient contentment, and even lower healthcare staff job contentment too, prominent to an obsessive turnover, that will result in the loss of key process wisdom. Poor-quality data can also breed organisational mistrust and make it hard to mount efforts that lead to needed improvements. To improve DQ, the primary key step is to evaluate a new way to collect and analyse quantitative data that belongs to improving the work processes by first understanding the basic procedures.

Quality data is suitable data. To be of prosperous quality, the data must be compatible and unequivocal. DQ issues are often the outcome of DBMS merges or systems/cloud integration procedures, in which data fields that should be compatible are not according to design or distribution inconsistencies. Data that is not high quality can sustain data abstergent to the enhancement of its DQ.

DQ performances employ information rationalisation and affirmation. DQ striving are often desiderate while integrating diverse applications happen in the course of amalgamation and achievement performances, but also when soloed DBMS within a single HCO are enhanced along for the first time in a data warehouse or LSDB lake. DQ is also ticklish to the efficiency of HCOs level applications, such as Enterprise Resource Planning (ERP) or Customer Relationship Management (CRM).

In the area of EHRs exchange, data standards are needed for data format, as well as the document model, healthcare report templates, user application interface and consumer EMRs linkage, as follows:

1. *Communication writing format standards:* Communication writing format standards simplify interoperability via the use of generic encipher depiction, EHR patterns for determining links among EHR elements, document models and healthcare patterns for modelling information as they are interchanged.
2. *Document model:* A method for introducing electronic health data, such as discharge compendiums or progressive records and consumer safety information, demand standardised model record architecture. This need stems from the longing to access the considerable gratified presently stored in free-text health notes and to authorise similitude of content from documents generated on EHR systems of extensively analysed representatives.
3. *Clinical templates:* EHRs represent the arrangement to differentiate further obligations on the variation of the data elements through the use of templates that can be applied against a version of messages or QA documents. The EHRs messages maintain moderate optionality, although they also provide some constraints. For greater precision in the standardisation of clinical data, more objective descriptions of the admissible components for the data components must be applied.
4. *User application interface:* The clinical apparatus organisations are well learned in constructing the user application interfaces that make the apparatus certain, more feasible and simpler to use, by introducing a deliberate standard for human multiplier model introduced by the Association for the Advancement of Medical Instrumentation (AAMI) and approved by the American National Standards Institute (ANSI).  
This standard, the ANSI/ AAMI HE74 Human Factors Design Process for Medical Devices, establishes devices and mechanisms to support the analysis, model, experimenting and appraisal of both simple and perplexing systems. These mechanisms and tools have been introduced for many years in the engineering of consumer applications, military applications, aviation arsenal and nuclear power systems. Discretion of the HE74 standard may present acumen into the processes introduced for

modelling and constructing user-friendly EHR systems, including electronic consumer safety as well as reporting systems.

*5. Consumer EMRs linkage:* While not an EHRs standard in the common interpretation, being conventional to associate a consumer EMRs from one HCOs or hospital to other unequivocal, is indispensable for handling the integrity of consumer EMRs and performing safe healthcare. The management palliation standards of the Health Insurance Portability and Accountability Act (HIPAA) initially obligated the development of a unique demographic to ascertain for consumers. However, the American Congress held back subsidisation of the development and postponed sufficient federal privacy conservation. Now that the HIPAA privacy regulations have been accomplished nationwide in the USA, this concept determined that consumer medical record over HCOs should be overlooked.

When EHRs represent with excellent quality, it can be simply processed, explored, experimented and analysed, governing to acuteness that assists the HCOs to make better decision-making. High DQ is indispensable to business intelligence endeavours and other appearances of data analysis, as well as a better implementation proficiency.

Data standards are the primary informatics element indispensable for data flow stream to the HCOs data infrastructure.

With common standards, clinical and patient safety systems can share an integrated information infrastructure whereby data are collected and re-used for multiple purposes to meet more efficiently the broad scope of data collection and reporting requirements.

The conventional EHR standards also facilitate the magnificent implementation of the new era into the decision-making components (for example, alert for the new medication contra-indication and decontamination of the treatment procedure). Due to the complexity of efficient data interexchange between different HCOs as a vast amount of EMRs demanded the healthcare, consumer safety and DQ enhancement reside on the healthcare domain, despite the availability of the ICT

to facilitate EHRs integration. It's becoming more complicated and prevents the EMR exchanges among laboratories, pharmacies, providers, HCOs and stakeholders for reimbursement when no common standards exist at the level of the HCOs (Hammond 2002).

In terms of patient care, the EHR standards terminology embedded generic methods, rules, data collection specifications, interexchange, terminologies, storing, EHR integration associated systems, including EMRs, drugs, laboratory results, remuneration, health-associated tools and data governance process controls (Washington Publishing Company, 1998). EHR systems standardisation employs four different indications as follows:

- a) *Specification of EHR components*: Identification of the EHR content to be collected, integrated and exchanged;
- b) *EHR exchange standards and formats*: EHR exchange standards and formats for the data components include process flow controls and trouble handling (Hammond 2002). EHR exchange standards can also include system models documentation for architecting data components as they are interchanged and EHR architectures that identify the links between EHR components in information.
- c) *EHRs specific technicalities or terminologies*: The clinical specific rules and concepts implemented in the EHR systems represent, categorise and encode the data components and information manifestation languages and modelling structure that represent the links between the clinical specific rules, including the concepts.
- d) *Knowledge description of EHRs standard*: EHRs standard methodologies for clinical composition, health guidelines, such as for decision-making support tools and *information representing standards*.

## 4.2.2 Levels of the Electronic Health Records Data Quality standards

Several researchers and data analytics have defined various Data Quality Standards (DQs). However, after profound analysis and literature reviews, currently, no consensus exists on that which the various DQs are. Figure 4.3 describes the high-level DQ process architecture framework, as follows:

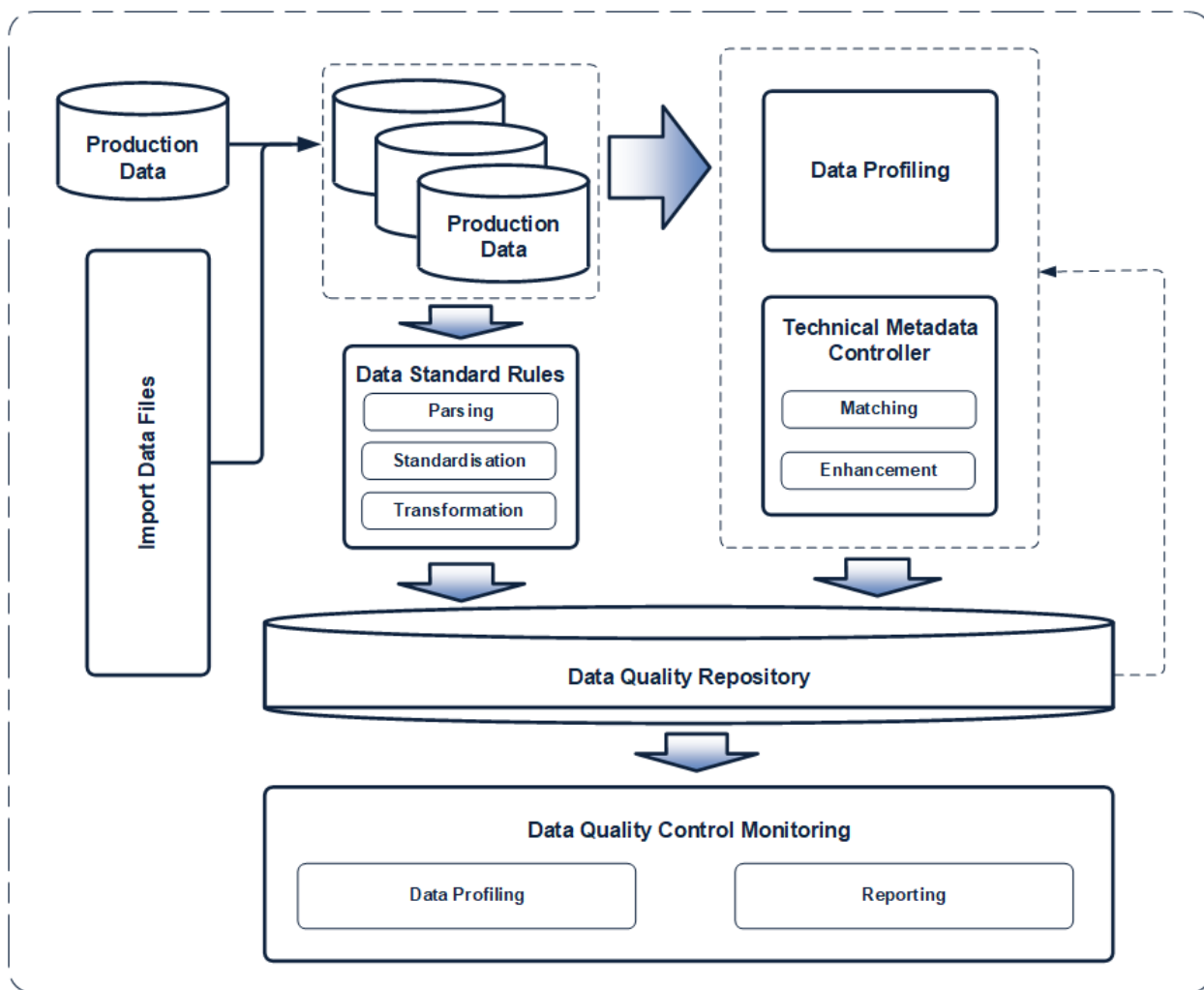


Figure 4.3: High-level data quality process architecture framework (researcher source)

Poor DQ can pervert principal clinical financial data, which can make it inconceivable to identify the financial situation of an HCO and it is essential for all levels in healthcare services. Generally, the provider requires high-level DQ for all

its procedures, specifically for diagnosis, identification of allergies and identifying the risk level, among others. At the indigenous stage, high-level DQ is essential so that patients are appraised properly with their own information and can communicate with the provider. The WHO DQ standard can be an ideal role model for DQ assessment and analysis.

Poor DQ can render it impossible for healthcare risk management to obtain an accurate estimate, which may miscalculate the claims. After profound analysis and the data analyser's efforts, it appears that too vague data still exists to analyse without principle data abstergent. Therefore, DQ remains the extensively recognised general properties and cannot typically be used without further elaboration to describe specific substances of DBMS that could affect experiments and modelling. The nine most commonly DQ standard cited properties are, as follows:

- 1) *EHRs topicality*: Three essential terms to the EHRs topicality are as follows:
  - a) The EHRs must meet the fundamental requirements for which they were procured, captured in a DBMS and employed;
  - b) The DQ and its property must be flexible to add additional features (for example, data analysis) and use them for several different purposes;
  - c) The possibility to change their role from primary to secondary or otherwise to create a subset of patients to determine the risk level for the healthcare service;
- 2) *Accuracy*: It is more likely impossible to protect against all the errors in every field while capturing or importing into the health database. To reduce these types of risks, data needs to be checked and verified and data types with standard constants, while capturing, prefer staging before importing data and ensure the proper data mapping. The following questions need to be set up and addressed to implement the standard and checks, as follows:

- a) How to determine the risk level according to the member's age to set up the scheme, benefit, tariff and the price level in the healthcare risk organisation.
- b) Which medication for certain diagnoses are more effective according to quality and price level?
- c) Determine the certain disease record to analyse the probable reason for consumers being deceased. The EHRs context can be used for the clinical experiment in which providers are experimenting with the effectiveness of a new medication. The EHRs data fields must include the consumer age, health condition, disease level, dosage level, genetic relations and consumer global location.
- d) The accurate measurement for the dosage level is needed to set what exactly the dosage level should be.
- e) How to determine all the factors that need to be measured (for example, including other medication or general health level) as these might be capable to replace the effectiveness of the new medicine.
- f) Are all related factors being measured with sufficient correctness to develop a standard specimen to effectively enumerate the effectiveness of different dosage levels of the new medication?
- g) What are the most stringent DQ standards for financial data essential for managerial or survey data?

**3) Compatibility:** Comparability consists and specifies the designate of the analytical DQ, in observing the variation in performance critical measures to the Triad as the Triad emphasises collaborative data sets. Various analytical methods show three different types of divers that can be identified in collaborative data sets. These variations are:

- a) Individual-level variations (for example, comorbidities, age and sex);
- b) Provider level variations;
- c) Random/residual variations;

It is obvious that challenges in data availability and comparability issues are numerous among national and international comparisons. Data should be applicable to assemble multiple DBMS into one DBMS to contribute to EHRs



to employ for empirical analysis, statistical or modelling estimation. The identity field must be present and should be accurate to merge easily a patient (for example, Social Security Numbers) or provider (for example, Practice Numbers) across different databases and some other fields could be included to enable the proper merge.

**4) *Completeness:*** Completeness is defined as the extent to which all data components are integrated. It also describes that no EHRs are absent and that no EHRs have absent data components. One of the most important aspects of the integration result is that completeness is the guarantee of the appearance of all ingredients when integrating. Each data component should be captured in EHR systems so that a provider could create a dataset for a patient's characteristics. It will, however, not be possible for a provider to create an accurate diagnosis for a patient characteristic with incomplete data, which cannot provide accurate diagnosis even when the corresponding information is supplied by the EHR systems. The data will remain incomplete until the providers are completed with the approximate group value associating to semantic categories. Finally, the inconsistencies of the patient diagnosis with appropriate values, with the inconsistency of any value in these appropriate values, are the intention acceptability of total designation for the EHRs completeness. In the survey health domain, if entire information is missing it is referred to as:

*a)* Unit non-response (UNR) and missing item referred;

*b)* Element non-response (ENR) can determine a deficiency of DQ;

In health databases, such as an economic database, the occurrence of UNR or ENR is considered as catastrophic considerations; in the observation and managerial database, it can have significant considerations if this information amalgamated with health LSDB or with a large patients' symmetry. To prevent this these issues must be audited to determine, as follows:

*a)* All healthcare staff have additional training in the use of EHR systems;

*b)* The EHR systems are sufficient, user-friendly and reactionary;

*c)* A particular procedure for refurbishing the database of the DBMS is sufficient and no oversight exists;

d) The servers are well maintained including creating a backup and replicating data regularly;

5) **Consistency:** Consistency is defined as the absence of any inconsistencies in the EHRs data and all appearing conflicts among components have been solved when integrated. When stream data appended dependent interpretative variables from diverse heterogeneous sources, often integrate data refers to a similar subject but apprehends different inconsistency information. The time (T) dimensional observation data could be large and is asymptotically valid for a certain time. The data could be repeated approximately at the same value over time; such a situation is called a conflict.

6) **Identification:** Identification is defined as structural similarity among the entity sources and the EHRs result. The EHRs domain apprehensions often contain entities that have interrelated among the property value. This defines that each EHRs entity is accompanied by certain concepts. If identical characteristics are associated with the identical apprehension in various ontologies, the conflict in the EHRs are also associated with the different associated EHRs values; this is also manifested as conflict.

7) **Timeliness:** Timeliness is defined as an association between the registration and diagnosis entity and the determined time to the observation diagnosis of the occurrence statistical report. The EHRs statistics reports improve provider observation of patient outcome. Overall statistics indicated that the use of the EHR systems could sustain improvements to the productivity of healthcare services, such as timeliness statistic reports or invoices. The current EHRs requires to enumerate, which subset of patient data is more similar to use in a particular situation. Some health data has to be released/edited daily with the survey information to patents, and it needs to be considered how the delay affects the employ of the EHRs in, as follows:

a) General notice for the patient;

b) Using result or other information for healthcare purposes;

*8) Accessibility and clarity of results:* The accessibility and clarity are referred to as the dimensions introduced in most statistical DQ environments. They are basically charted edgewise several degrees, such as topicality, correctness and compatibility, for which there are completely implemented and elaborate indications of measurement (Steven 2008). Approachability is defined as in terms of to meet the needs of the user. The word accessibility is primarily considered in denominations of tabular information; although there is an intensively flourishing consideration in discovering efficient contrivances to current EHRs graphically. The concept of clarity is defined as a term for which a consensual repercussion is required on that which the consumer actually demands.

*9) Coherence:* Coherence in EHRs is defined as when everything fits together well, in other words, logical and complete with numerous supporting facts. HCOs are always looking for coherence in EHRs to support their service.

### **4.2.3 The Cost of low-level Electronic Health Records Data Quality Standards in the health domain**

HCOs are constantly challenged to maintain the right level of DQ. The DQ may seriously affect patient care and even could lead to the death of the patient. These are the key challenges of eradicating treatment errors in the health service process. As patient safety is the key issue in healthcare service, using effective EHR systems integration and implementation can improve the DQ to reduce medical errors. The main consideration for health data includes data accuracy and accessibility, as well as data comprehensiveness, currency, consistency, granularity, precision, relevancy definition and timeliness. Although the adoption of electronic health record (EHR) systems promises a number of substantial benefits, including better care and decreased healthcare costs, serious unintended consequences from the implementation of these systems have emerged.

Poor EHR systems design and improper use can cause EHR-related errors that jeopardise the integrity of the information in the EHR, leading to errors that

endanger patient safety or decrease the quality of care. These unintended consequences also may increase fraud and abuse and can have serious legal implications. Whether a provider accesses EHRs via a cloud service or administrators' access data during normal hospital facility operations in a data centre, the regulators require that the data is accurate and maintained with the proper level of standard. Table 4.1 describes the comparison between systematic and random, as follows:

Table 4.1: Comparison between systematic and random

Systematic Issue	Random issue
Unclear data definitions	Illegible handwriting in data source
Unclear data collection guidelines	Typing errors
Poor interface design	Lack of motivation
Programming errors	Frequent personnel turnover
Incomplete data sources	Calculation errors (not built into the system)
Unsuitable data format in the source	
Data dictionary is lacking or not available	
Data dictionary is not adhered to guidelines or protocols are not adhered to	
Lack of insufficient data checks	
No system for correcting detected data errors	
No control over adherence to guidelines and data definitions	

The poor DQ in demographics results in duplicate and confused patient entries on EHR systems. In other words, one patient with more than one Patient Identification Number (PIN) or the same PIN number is assigned to more than one patient or a record in place of an update is inserted, when data existed. The consequences can result in incorrect and mixed medical records, missed screening requests and even cancelled operations. Many HCOs are upgrading their outdated technology and implementing new data capture systems (enter the EMR). However, it is important to note that technology alone will not fix this problem of poor DQ. DQ is not a technology problem, as much as it is a people and process problem.

Danielle, DeKeizer *et al.* (2002) divide potential causes of poor DQ into two areas, systematic and random. If one takes a careful look in DQ issues, very few of these issues are technology related.

#### 4.2.4 Benefits of high-level Standards of Electronic Health Records in the healthcare domain

EHRs are defined as “a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting. Included in this information are patient demographics, progress notes, problems, medications, vital signs, past medical history, immunisations, laboratory data and radiology reports” (HIMSS 2017). Clinical outcomes include improvements in the quality of care, a reduction in medical errors and other improvements in patient-level measures that describe the appropriateness of care. Organisational outcomes, on the other hand, have included such items as financial and operational performance, as well as satisfaction among patients and clinicians who use the EHRs. Lastly, societal outcomes include being better able to conduct research and achieving improved population health.

1) *EHRs and clinical outcomes:* Many clinical outcomes that have been a focus of EHR studies relate to the quality of care and patient safety. Quality of care has been defined as “doing the right thing at the right time in the right way to the right person and having the best possible results” and patient safety has been defined as “avoiding injuries to patients from the care that is intended to help them”. Quality of care includes six dimensions, but most EHR research has focused on the following three as below:

- a) Patient safety;
- b) Effectiveness;
- c) Efficiency;

In the following paragraphs, some of the studies are summarised that examine the way in which EHRs or various related components impact these three quality

dimensions. More research is needed on the other three components, namely patient-centeredness, timeliness and equitable access.

2) *The EHRs organisational outcomes and societal benefits:* The recognition of the EHRs importance in organisational outcomes and social benefits are as follows:

a) *Organisational outcomes:* Studies examining organisational outcomes have focused on EHRs use in both the inpatient and outpatient settings. Such outcomes have frequently included increased revenue, averted costs and other less tangible benefits, such as improved legal and regulatory compliance, improved ability to conduct research and increased job/career satisfaction among physicians. Increased revenue comes from multiple sources, including improved charge capture/decrease in billing errors, improved cash flow and enhanced revenue. Several authors have asserted that EHRs assist providers in accurately capturing patient charges in a timely way (Schmitt *et al.* 2002).

b) *Societal benefits:* Another less tangible benefit associated with EHRs is an improved ability to conduct research. Having patient data stored electronically increases the availability of data, which may lead to more quantitative analyses to identify evidence-based best practices more easily (Aspden 2004). Moreover, public health researchers are actively using electronic clinical data aggregated across populations to produce research that is beneficial to society. The availability of clinical data is limited, but as providers continue to implement EHRs, this pool of data will grow. By combining aggregated clinical data with other sources, such as over-the-counter medication purchases and school absenteeism rates, public health organisations and researchers will be able to better monitor disease outbreaks and improve surveillance of potential biological threats (Kukafka *et al.* 2007).

## 4.2.5 Implementation of Electronic Health Records Data standards

Implementing data standards is just as important as developing and selecting standards. In preparing to implement standards, several issues tend to arise that should be considered when establishing a mechanism for compliance. These issues include vendor readiness, organisational readiness, cost of compliance tools, unresolved issues related to terminologies and coding, identifiers for providers and patients and interpretation of the implementation guides and standard specifications. Help in dealing with these issues is critical.

In addition to the establishment of an oversight organisation and a national implementation plan, a mechanism for assessing conformance with the data standards is needed. Conformity assessment, an integral part of the utilisation of standards, is the comprehensive term for measures taken by manufacturers, their customers, regulatory authorities and independent third parties to evaluate and determine whether products and processes conform to particular standards (National Research Council, 1995). The National Institute of Standards and Technology (NIST) could perhaps serve as the body supporting the implementation process as the developer of protocols for conformance tests, information assurance and certification procedures, to verify vendors' compliance with the standards.

Because, the core terminology group for the EHR and other health-related applications will be housed for public availability within the UMLS, NLM will play a vital role in the coordination, mapping and dissemination of the terminologies for national adoption. NLM will share responsibility for the maintenance and regular updating of the terminologies with the terminology developers. As the chief standards development organisation for the EHR, HL7, in collaboration with government organisations (for example, Centers for Medicare and Medicaid Services), will develop the specifications for the actual implementation of the terminologies.

As stated in Chapter Three, the EHRs integration system can facilitate the standards adoption process by functioning as a coordinating body and provider of technical assistance for the efforts in the area of data standards and for the Quality Interagency Coordination Task Force, evidence-based practice centres, specialty societies, academic institutions and professional organisations involved in the determination of best practices, which become translated into electronic data systems. EHRs should be fully funded to function in this capacity.

Assessing the costs related to the development, implementation and dissemination of data standards, will involve a coordinated set of evaluations by EHRs integration systems. These would most likely have the responsibility for estimating the costs related to the establishment and operation of a data entity for standards implementation and conformity assessment. EHRs would have the responsibility for estimating the costs related to the development and maintenance of the core terminology group and mappings to supplemental terminologies. Together, these organisations should engage in a comprehensive evaluation of the costs to provide the data standards needed for the NHII and patient safety systems.

#### **4.2.6 The Electronic Health Records data integration challenges**

Electronic Health Records (EHRs) play a critical role in the translation of genomic information into clinical care. Tremendous strides have been made in making pooled health records available to data providers for healthcare services; yet still, more must be done to harness the full capacity of big data in healthcare. Despite the widespread adoption of electronic health records, the integration of healthcare data remains a critical challenge for the industry as it strives to achieve interoperability.

Many EHRs are kept in a variety of unstructured formats, making it difficult to query directly via digital algorithms. Paper health records are standalone, lacking the ability to integrate with other paper forms or information. The ability to integrate health records with a variety of other services and information and to share the



information is critical to the future of healthcare reform. Digital, unlike paper-based healthcare information, can be integrated with multiple internal and external applications:

- a)* Ability to integrate for sharing with health information organisations (another chapter);
- b)* Ability to integrate with analytical software for data mining to examine optimal treatments, etc.;
- c)* Ability to integrate with genomic data as part of the electronic record. Many organisations have begun this journey;
- d)* Ability to integrate with local, state and federal governments for quality reporting and public health issues;
- e)* Ability to integrate with algorithms and artificial intelligence;

### **4.3 Why Data Quality is important in Electronic Health Records?**

DQ is important to HCOs for numerous reasons and it is the major EHR's essential asset and a unique source of competitive advantages. Poor quality data can lower employee job satisfaction too, leading to an excessive turnover and the resulting loss of key process knowledge. Poor-quality data can also breed organisational mistrust and make it hard to mount efforts that lead to needed improvements. Further, poor-quality data can distort key corporate financial data. In the extreme, this can make it impossible to determine the financial condition of a healthcare business. In construct, the following reasons impact the DQ in EHRs:

- a)* Incorrect data may seriously affect patient care;
- b)* Poor quality data reduces patient satisfaction;
- c)* Death of the patient;
- d)* A reputation for healthcare world-class quality service is profitable;
- e)* Incorrect or missing data can have disastrous consequences financially for HCOs;

### 4.3.1 Why care about Data Quality for Electronic Health Records?

The ability of EHRs to share and exchange the quality data seamlessly offers numerous benefits. The use of EHRs is expected to provide quality healthcare services that improve the health outcomes for patients. The benefit will, however, only realise if the data in the EHRs are of sufficient quality to support these uses. The EHR systems that provide quality data with interoperability in mind, enable timely access to necessary health information. Proper EHR integration systems are able to share information and reduce the need to recapture the same information over and over in every EHR system. An EHR integration system can improve the quality healthcare service. For example, when EHR systems access any local EMR through interoperability, the attending doctor receives more complete health information, including chronic family histories.

EHRs can improve the patient safety when using electronic prescriptions with the capability to alert when a medication is prescribed that could interact with other medication, which the patient is known to be allergic for or currently taking. This would reduce the possibility of a potentially fatal drug reaction and improve patient safety. Using EHRs as a multidisciplinary team of health professionals, who must interact with other providers to provide the best possible care to the patient, could reduce the healthcare cost from unnecessary duplicate diagnostic tests or investigations.

A proper EHRs integration system involves the genuine concern of regarding the threats of data exchange, including unauthorised access, identity theft or information alteration. The EHRs integration systems reduce the unauthorised access by using compatible security models, identification and authentication models as well as the proper encryption algorithm. Only the quality data can provide all those possible EHRs benefits discussed above, that implies the need to care about DQ.

### 4.3.2 How to obtain Quality Data in Electronic Health Records?

Many authors have defined a different way to obtain quality data. However, the DQ management processes support the observance of the DQ policies, such as standardised data inspection templates, operational DQ, issues tracking and remediation, manual intervention when necessary, the integrity of data exchange, contingency planning and data validation.

The fragmentation of EHRs and its inability to exchange health information, necessary to support continuity of care, makes it a difficult task to take to its full advantages. Those initiatives applied in determining the essential factor that is necessary to successfully drive interoperable healthcare service include, are as follows:

#### 4.3.2.1 Prevention bad data in Electronic Health Records

*Keep bad data out of the databases:* The first and preferable way is to ensure that all data entering the database are quality data. One thing that helps in this regard is a system that edits the data before they are permitted to enter the database that prepares the raw data in the staging stage according to the quality data standard before integration. Moreover, as Granquist *et al.* (1977) suggest, “The role of editing needs to be re-examined and more emphasis placed on using editing to learn about the data collection process, to concentrate on preventing errors rather than fixing them.”

Of course, there are other ways besides editing to improve the quality of data. Here organisations should encourage their staff to examine a wide variety of methods for improving the entire process. One way in a survey sampling environment is to improve the data collection instrument, for example, the survey questionnaire. Another is to improve the methods of data acquisition, for example, to devise better ways to collect data from those who initially refuse to supply data in a sample survey.

### 4.3.2.2 Detection of bad data in Electronic Health Records

*Proactively look for bad data already entered:* The second scheme is for the data analyst to proactively look for DQ problems and then to correct the problems. Under this approach, the data analyst needs at least a basic understanding of:

- a) The subject matter;
- b) The structure of the database/list;
- c) Methodologies that might be used to analyse the data;

If one has quantitative or count data, a variety of elementary methods can be used, such as univariate frequency counts or two-way tabulations. More sophisticated methods involve Exploratory Data Analysis (EDA) techniques. These methods, as described by Keith *et al.* (1994), are often useful in examining:

- a) Relationships between two or more variables;
- b) Or aggregates;

They can be used to identify anomalous data that may be erroneous. Record linkage techniques can also be used to identify erroneous data. Suppose two databases had information on the patients of HCOs. One of the databases had highly reliable data on patient health history, but only sketchy data on the diagnosis data on these patients. The second database, however, had essentially complete and accurate data on the diagnosis of the patients. Records in the two databases could be merged and the diagnosis data from the second database could be used to replace the diagnosis data on the first database, thereby improving the DQ of the first database.

### 4.3.2.3 Repair the bad data in Electronic Health Records

*Let the bad data find the problems and then fix things:* By far, the worst approach is to wait for DQ problems to surface on their own. Does a chain of grocery stores really want its retail customers doing its DQ work by telling the store

managers that the scanned price of their can of soup is higher than the price posted on the shelf? Will, a potential customer be upset if a price higher than the one advertised appears in the price field during checkout at a website? Will an insured person whose chiropractic charges are fully covered be happy if his health insurance company denies a claim, because the insurer classified his health provider as a physical therapist instead of a chiropractor? DQ problems can also produce unrealistic or noticeably strange answers in statistical analysis and estimation. This can cause the analyst to spend much time trying to identify the underlying problem.

### 4.3.3 Allocating resources for prevention, detection and repair in Electronic Health Records

The question arises as to how best to allocate the limited resources available for a sample survey, an analytical study or an administrative database/list. The typical mix of resources devoted to these three activities in the United States tends to be in the order of:

- Prevent: 10%;
- Detect: 30%;
- Repair: 60%;

According to our investigation and experience, we have strongly suggested that a more cost-effective strategy is to devote a larger proportion of the available resources to preventing bad data from getting into the system and less to detecting and repairing (for example, correcting) erroneous data. It is usually less expensive to find and correct errors early in the process than it is in the later stages. So, in our judgment, a much better mix of resources would be:

- ✓ Prevent: 45%;
- ✓ Detect: 30%;
- ✓ Repair: 25%;

#### 4.3.4 The Electronic Health Records process improvement

One process improvement would be for each company to have a few individuals who have learned additional ways of looking at available procedures and data that might be promising in the quest for process improvement. In all situations, of course, any such procedures should be at least crudely quantified – before adoption – as to their potential effectiveness in reducing costs, improving customer service and allowing new marketing opportunities.

#### 4.3.5 Train medical staff

Many HCOs may have created their procedures to meet a few day-to-day processing needs, leaving them unaware of other procedures for improving their data. Sometimes, suitable training in software development and basic clerical tasks associated with customer relations may be helpful in this regard. Under other conditions, the staff members creating the databases may need to be taught basic schemes for ensuring minimally acceptable DQ. In all situations, the HCOs should record the completion of employee training in appropriate EHR systems and if resources permit, track the effect of the training on job performance. A more drastic approach is to obtain external hires with experience/expertise in, as follows:

- a) Designing databases;
- b) Analysing the data as they come in;
- c) Ensuring that the quality of the data produced in similar types of databases is *“fit for use”*;

### 4.4 Electronic Health Records Data Quality control

Data that is an accurate representation of the part of the “real world”, are models. DQ challenges include transparency, reputation, compliance, decision-making and cost reduction. Organisational rules identification, implementation of business rules, monitoring results and feedback loop includes:

- a) Study of the existing documentation of the Organisational Rules (OR);
- b) Review of reports on DQ;
- c) Own knowledge of the environment;
- d) Classification of the OR according to DQ dimensions;
- e) Data analyses on the contract data;
- f) Identification of the key business rules based on results

*Define prioritisation criteria:* Agreeing on prioritisation criteria for the implementation of the business rules. Implementing the EHRs to automate the verification. Implementing a reusable framework to measure and monitor the DQ. Providing a tool for the visualisation of the results by the business data owners. The DQ controls are prioritised based on the following criteria:

- a) *Feasibility:* What is the complexity of implementing a business rule to control the data? How long will it take to develop the control?
- b) *Materiality:* What is the impact of not respecting the business rules in man-workdays? What is the efficiency loss in man-workdays due to bad DQ? Because of this error, will a long time be needed to fix the report?
- c) *Reputation:* What is the impact on the DQ reputation? What will happen when one sends the wrong figure or report to external parties?
- d) *Compliance:* Which regulation does one need to follow? Is one compliant with the financial regulation, with the accounting rule?

*Implement organisational rules:* Includes the design of document functionalities of the framework (use case and activity diagrams) and defining a template for manual DQ checks results. Building the framework using the EHRs integration model, which includes the HCO's rules to a set of data; producing the results of the quality assessment through a dashboard and validating the implementation.

*Monitor results:* Refers to designing the dashboard, contents and layout, DQ results per dimension, the evolution of the results across time, as well as building the dashboard and validating the implementation. In this way the system can feedback loop to the identification of the issue by the data owner, evaluating the issues, prioritising, communicating to the data encoder for correction,

communicating to the support team for potential bug/defect correction, communicating to the IT DQ team for post-processing and communication.

## 4.5 The Electronic Health Records Data integration model

A term that is often, but incorrectly used interchangeably with interoperability, is integration. EHRs are particularly important to the advancement of integrated clinical systems because they provide the backbone for the next set of standards needed for the EHR. These include those required for the use of concept-oriented terminologies, document architectures, clinical templates, alerts and reminders and automated clinical guidelines, all which would result in improved interoperability and structuring of clinical and patient data. Figure 4.4 (Saïod *et al.* 2019b) describes the EHRs integration model, as follows:

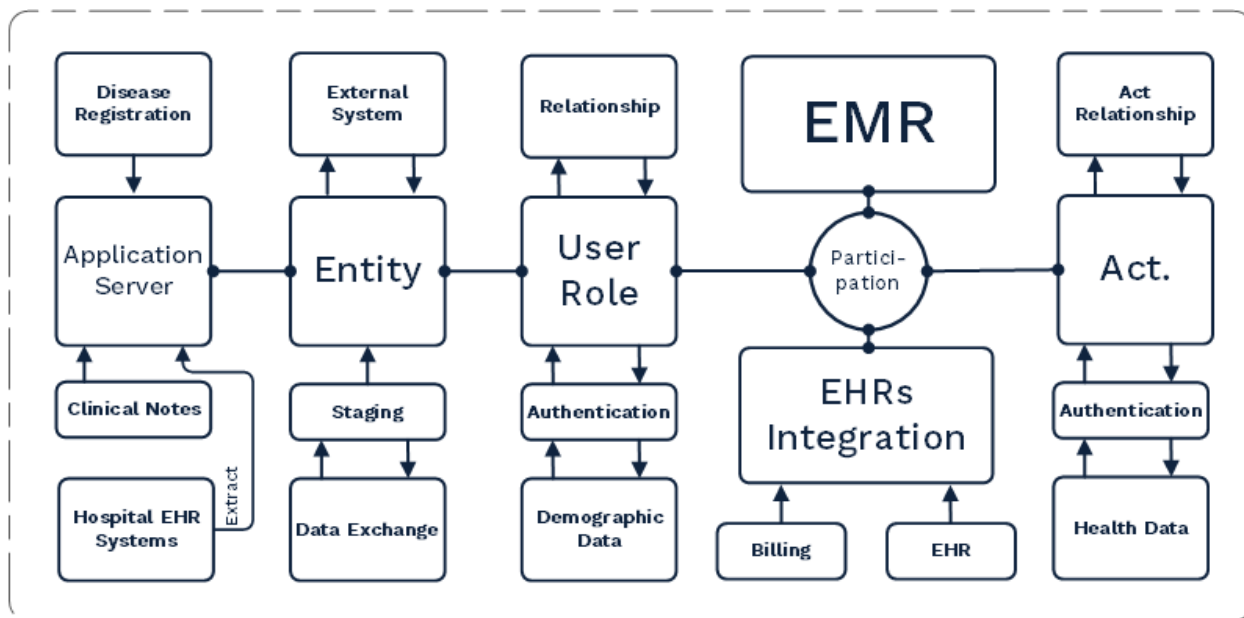


Figure 4.4: The EHRs integration model (Saïod *et al.* 2019b)

A method for representing electronic clinical data, such as discharge summaries or progress notes and patient safety reports, requires standardised document architecture. This need stems from the desire to access the considerable content currently stored in free-text clinical notes and to enable the comparison of content from documents created on information systems of widely varying characteristics (Dolin *et al.* 2001). The architecture should be designed as a layout standard (Dolin



*et al.* 2001) so that clinical documents can be revised as needed or appended to existing documents.

It should also be able to accommodate the desire for rich narrative text that makes up a significant portion of patient safety information from voluntary and mandatory reports. The DQ framework components are as follows:

- a) Profiling:** Checking-up the database by screening all tables and columns in scope through several criteria;
- b) Metadata comparator:** Comparing the technical constraints on a database with the profiling results.
- c) Business validation:** Meetings organised with data owners and data stewards to identify pain points and categorise DQ checks by priority.
- d) Business rules:** Implemented in EHRs DBMS to logging and error handling, integration of manual checks and DQ issue exclusion;

## 4.6 Summary

This chapter discussed what data standards entail in the healthcare segment. It provided a brief definition of EHRs data quality and various levels of DQ standards. The complexity of the healthcare domain and the associated challenges that contribute to the difficulty of attaining the DQ challenges in there was discussed including the DQ standard, the reason for caring about DQ, the way to obtain the quality data, and the way to prevent and improve the data processing. The chapter also gave an overview of EHRs components that control DQ. One of the components is DQ standardisation that focuses on the research presented in the study. In addition, the chapter described the EHRs integration model that should be implemented to enable seamless health data integration.

## 4.7 Conclusion

Although the use of EHRs in the healthcare domain has great potential to improve the healthcare quality, including the service efficiencies, the realisation of achieving

the maximum benefit is limited to DQ. However, achieving the full benefit using the EHRs in healthcare service is practically challenging for several reasons and DQ remains the key issue in the EHRs to enable quality care service. The review in Chapter Four revealed many ways to improve the DQ and prevent bad data. These include the EHRs DQ control and using the proper data integration model. Therefore, the primary driver to prevent the DQ issue is data standardisation. Figure 4.5 demonstrates the outcome of Chapter Three and Chapter Four, see below:

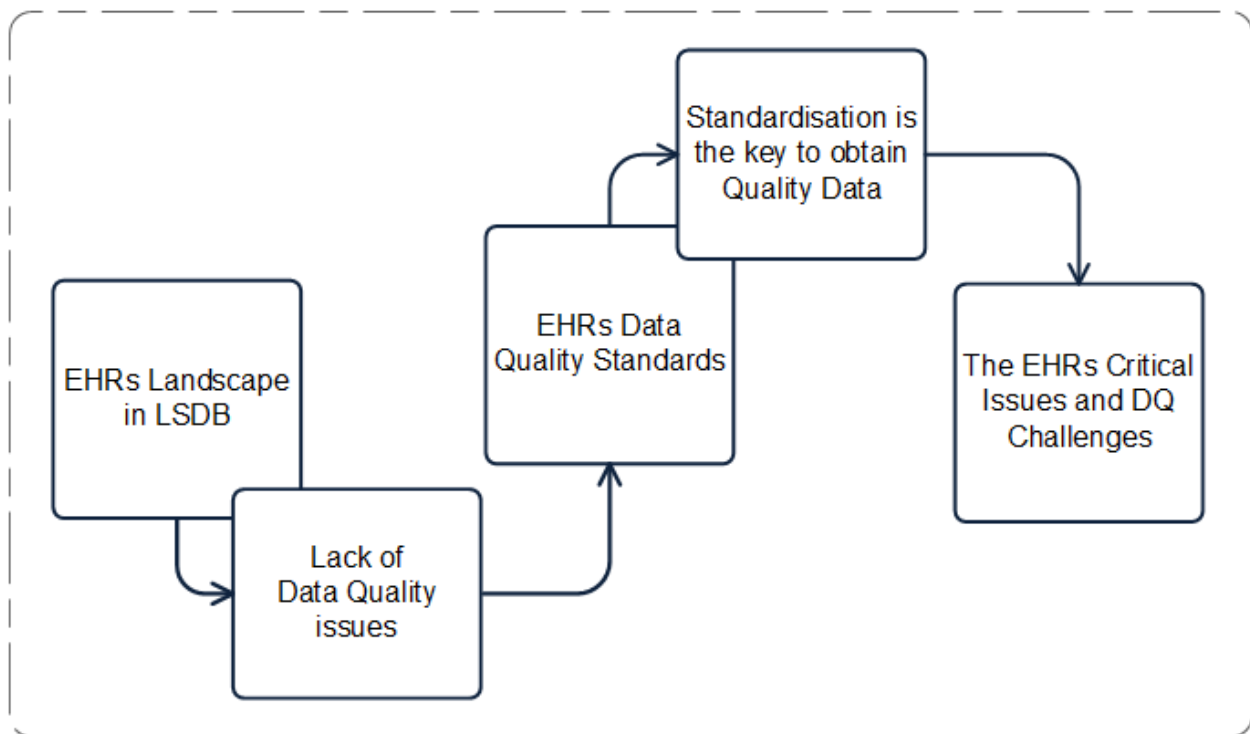
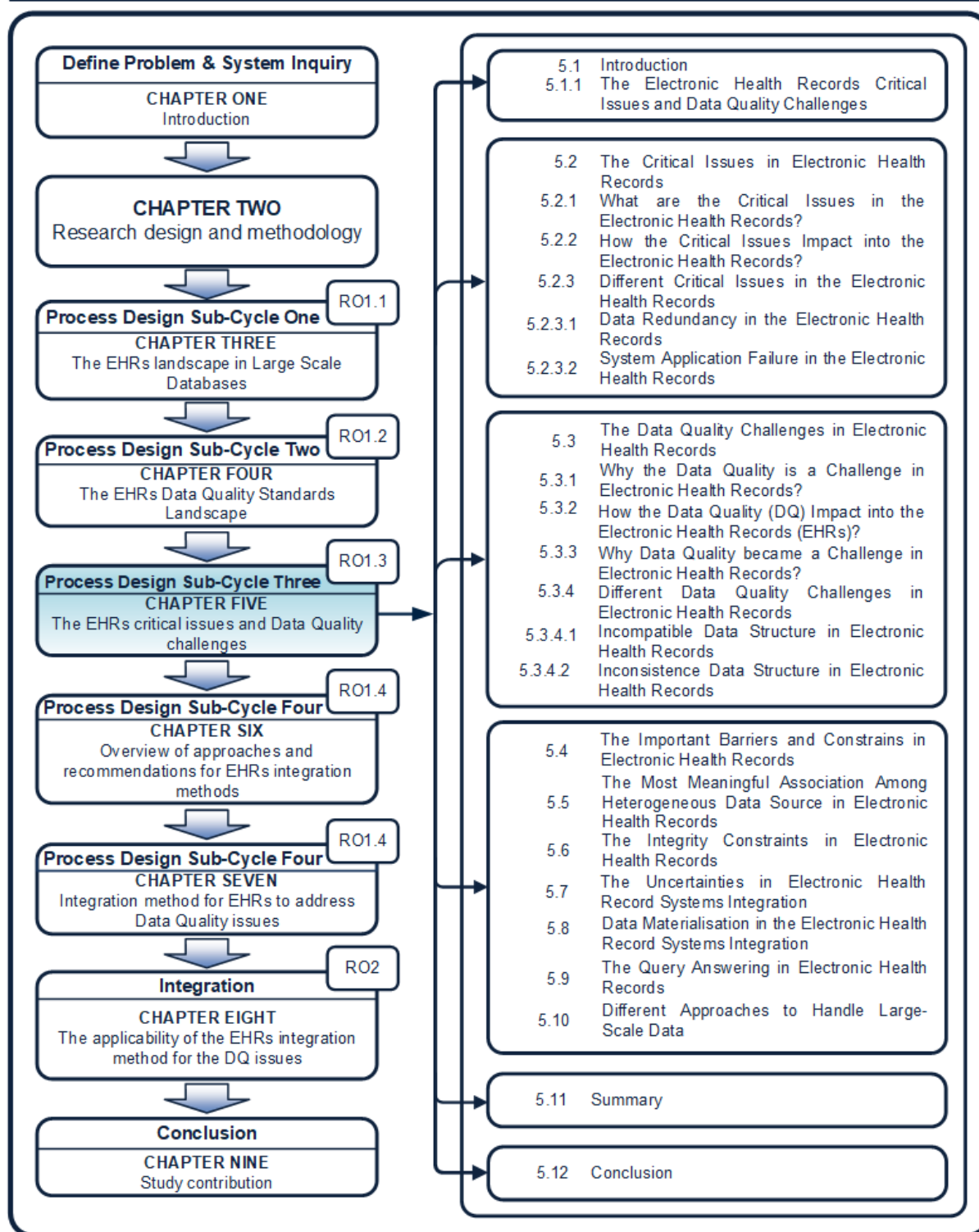


Figure 4.5: The combined outcome of chapter Three and chapter Four

As illustrated in figure 4.5, the next logical step in this study is, therefore, the EHRs critical issues and Data Quality challenges. This is done in Chapter Five.

## CHAPTER FIVE: The EHRs critical issues and Data Quality challenges



Outline of the Chapter Five

## CHAPTER FIVE

### 5.1 Introduction

DQ is an integral essential component of EHR systems. Quality assurance for the systems not only identifies the current issues in the EHRs but also aims to minimise the risk in the health service process. EHRs are usually used for purposes other than healthcare delivery, namely research or management.

This fact has an important impact on the way in which data are introduced by healthcare professionals, on the way the data are recorded on the databases and also on the heterogeneity found when trying to integrate data from different Information Systems (IS). The purpose of this chapter is to describe the critical issues and DQ challenges in the healthcare domain, the way to detect them and overcome those challenges and barriers. It maps to Sub-Cycle Three of the DSR process, which is described in section 2.3.2.2.3 and highlighted in figure 5.1.

Sections 5.2 and 5.3 give a formal idea about the EHRs critical issues of the historic events leading up to EHRs data integration. Section 5.2.3 reviews the different critical issues, including data redundancy and system application failure in sections 5.2.3.1 and 5.2.3.2. Section 5.3 reviews the DQ challenges. To make sense of what exactly the DQ challenges are in EHRs, sections 5.3.1 and 5.3.2 look at the different ways in which the DQ is a challenge and the way this impacts in EHRs. Section 5.3.3 provides a detailed discussion of the reasons for DQ to become a challenge. Different DQ challenges including the incompatibility and inconsistency of data structure are discussed in sections 5.3.4, 5.3.4.1 and 5.3.4.2, and the maximum benefit that can be achieved from DQ.

The important DQ barriers and constraints in EHRs are discussed in section 5.4, whereas section 5.5 gives an overview of the most meaningful association among heterogeneous data sources in EHRs. Section 5.5 provides an overview of the most

meaningful association among heterogeneous data sources in EHRs and section 5.6 gives an overview of the integrity constraints in EHRs. The uncertainties in EHRs data integration are discussed in section 5.7. Section 5.8 gives a detailed overview of data materialisation in EHRs integration. The query answering in the context of data exchange is an important issue discussed in section 5.9. The summary and conclusion of the chapter are provided in sections 5.11 and 5.12 respectively.

### **5.1.1 The Electronic Health Records critical issues and Data Quality challenges**

Improving DQ to achieve benefits through EHRs is neither low-cost nor easy. However, different HCOs have several standards and different major systems, which have emerged as critical issues and practical challenges. One of the main challenges in EHRs is the inherent difficulty to coherently manage incompatible and sometimes inconsistent data structures from diverse heterogeneous sources. The DQ may seriously affect patient care and even could lead to the death of the patient. These are the key challenges of eradicating treatment errors in the health service process. As patient safety is the key issue in healthcare service, using effective EHR systems integration and implementation can improve the DQ to reduce medical error. The main consideration for health data includes data accuracy and accessibility, as well as data comprehensiveness, currency, consistency, granularity, precision, relevancy definition and timeliness.

## **5.2 The critical issues in Electronic Health Records**

The EHRs all have the expectation to enhance the quality of service delivery, competency and consequences of healthcare through excellent implementation, integration, handle worldwide, presence and exchange of clinical information. Successful and effective EHR systems integration depends on a technical factor of the host operator at the HCOs, provider and consumer levels that continue condign of important erudition.

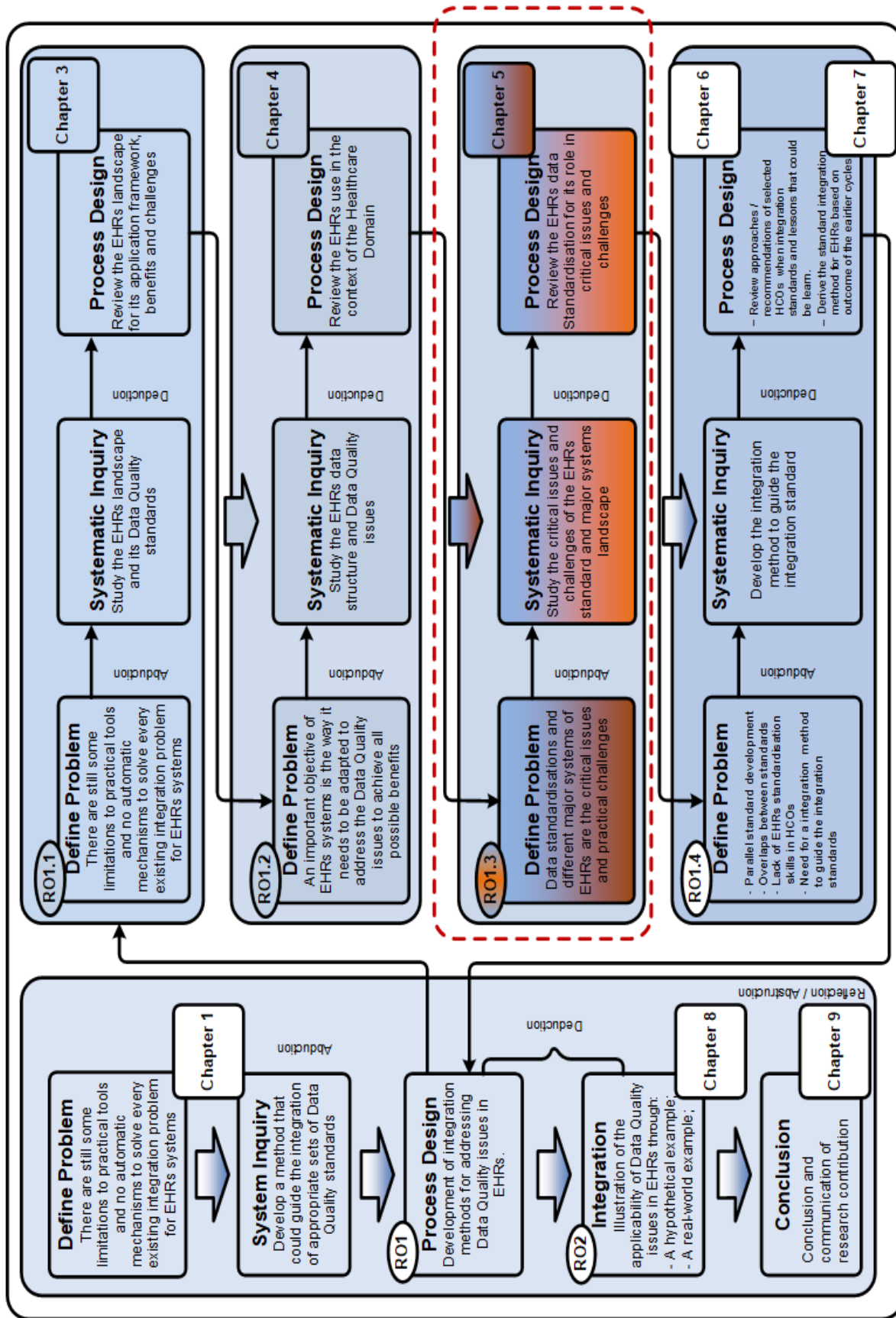


Figure 5.1: The position of Chapter 5 in the design science research process used in this study

One of the important aspects to overcome the critical issues in EHRs is to excellently comprehend the bracing between integration systems, the so-called human artefact. The healthcare services (real world) requires a definition of the phase to be introduced to create a human artefact. Hogan *et al.* (1997), “argued that EHRs were not properly evaluated regarding data accuracy”.

Generally, to accomplish all the DQ issues to provide quality care services and securely attain accurate and irrefutable EHRs, the data redundancy needs to be reduced, system application failures prevented and decision-making ameliorated. In this way, healthcare can be conducted through the decision to support EHR systems. Efficient EHR systems provide professional healthcare support and reduce variation beyond any suffocating healthcare service or any social substantiality.

Introducing the IT and healthcare professional implements a clear and reciprocal intellect of the EHR systems, which will assist with the adaptation of the new system, including HCOs work tendency. The security concerns and confidentiality not only focused on efficiently managing the EHRs but also to ensure the reassurance of the consumer about being verbatim introduced to secure access. This concept requires implementation in the system flow control architecture and the way these will be implemented and analysed to EHRs, together with the nimble and considerable analysis in the routinely DQ measurement process.

### **5.2.1 What are the critical issues in Electronic Health Records?**

According to Peter *et al.* (2009), “the common issues were engaging the public, so as to support the move to EHRs without erosion in trust, engaging practitioners sufficiently to motivate changes in work processes necessary for adoption and engaging policy-makers to maximise access to patient databases for care quality improvement and formal research”. The technical level of the integration system, including the semantic level, should be modelled in such a way that the EHRs will not miss during the data exchange between the system and healthcare placing.

The EHR standardisations are diametrical accomplishments, whereas DQ remains the principle issues rather than amounts. Other key issues also involved in the successful EHRs implementation include:

- a) The inappropriate fact that influence the EHRs;
- b) Insufficient training, knowledge and accomplishments;
- c) Poor DQ measurement;
- d) Consumer confidence and belief;
- e) Cultural interchange;
- f) Handling the changeable statements, technological systems and EHRs apprehension;
- g) Governing the HCOs expense;
- h) Assuring the EHRs availability and secure access;

### **5.2.2 How the critical issues impact into the Electronic Health Records?**

Enhanced quality and safety are the key issues in EHRs; these are considered the specific decision-making endorsement that alleviates or prevents issues regarding the availability of EHRs incidental wisdom. These concepts enable the possibility to achieve the maximum benefit of EHRs for quality care services and efficient EHRs access, which fructuously improve the EHRs meaningful use. The EHRs competency is based on consumer participation to authorise consumers according to their own care, particularly for chronic conditions. It is significantly effective when consumers have direct access to their own EHRs. In this way, they can monitor their health improvements, such as sugar level, blood pressure, lab report and BMI.

### **5.2.3 Different critical issues in the Electronic Health Records**

One of the common issues in EHRs is to introduce the consumer, to the EHRs adaptation and implementation besides deterioration (for example, consumer belief). The providers' work habit, including all clinical professional service



processes, needs to indispensably change for the EHRs adaptation and introduce policy-makers to have sufficient access to EHRs for healthcare service amelioration, general purposes and analysis. Other EHRs anxieties and challenges are the implementations and then maintenance for the consumer belief achievement (for example, consumer's EMRs safety, efficient accessibility and privacy). This concern has often been nibbling challenges and barriers for the EHR systems adaptation, whereas other numerous emergency healthcare systems have been accomplished and adopted without any kinds of discernible firmness.

This study identified the reason as emergency healthcare service systems being an incidental and not ongoing continuous process as EHR systems included as filament and gains are easy to impress. The cultural reevaluation in terms of EHRs adaptation is another challenge in the need to change the service process and set different new priorities. This concern became a drastic challenge for the EHRs adaptation as the central process control changes, which impacts the entire healthcare services compared to simple IT shaping. This study was concerned that the national EHRs projects are still far from being vindicated as cost-effective systems associated with healthcare service improvement, due to the complete transformation to EHRs, interexchange and efficient integration disabilities. Therefore, a number of consequences have been exposed and raised to actualise the advantage of EHRs associated with the issues, as follows:

- a)* Availability of pecuniary fund opportunity, the essential infrastructure and times;
- b)* Diaphanous deficiency of EHRs DQ standards;
- c)* The probability of social and political impact into efficient EHRs delivery;
- d)* The necessity to replenish the EHRs to achieve DQ;
- e)* The lack of training and clear knowledge of clinical stuff;
- f)* The indispensable requirement for feasible administration to meaningful use of EHR systems to fulfilment;
- g)* The requirement to tackle divergent inducement and simplify changes for healthcare professionals' strength;

- h)* The probability to enhance healthcare distinction when granting additional access to EHRs which already have sufficient access;

The EHRs prospective and priority changes over time, but the efficient maintenance requires a combination of consumer expectation and healthcare professional service improvements, as the study has identified the fundamental variation between EHRs and the paper-based health record.

Due to the healthcare industry improving rapidly, EHR systems, therefore, need to improve faster to reduce all possible inequality for efficient delivery. EHR systems can provide distance eLearning (electronic learning) healthcare systems and present a diverse degree of healthcare procurement. EHRs are sophisticated systems, which introduced into intelligent inoperable healthcare service systems, require a wide adoptable method to exposition and appraisal and cannot be treated as an ordinary IT intervention. More consumer introduction into the EHR systems might impressively turn the EHR concepts, together with the improvement of the healthcare delivery.

Generally, the type of EHR systems may differ according to the social, political, economic aspects. For example, in a developed country the EHRs service delivery will absolutely differ from the developing country as all IT and healthcare infrastructures are in place. Therefore, according to these concepts, the efficient integrated EHRs and knowledge-based system highly impact on healthcare delivery and safety.

### **5.2.3.1 Data redundancy in the Electronic Health Records**

Data redundancy (DR) in the database is defined as when similar data fields are unnecessarily repeated in the database. This concept is clear when similar data fields exist within a single database. Where ever data repeats within a single database this generally constitutes as data redundancy. This might happen due to different circumstances, such as accidental design during the recovery and backup

process. According to Carol *et al.* (2016), “DR is a common issue in computer data storage and database systems”.

This data redundancy may appear due to some other incident, such as when similar fields are repeated unnecessarily between multiple tables. Yes, data may repeat for different reasons and will not be considered as data redundancy. For example, when the blood pressure or sugar level reading is captured to monitor the patient health improvement and a similar value may repeat multiple times at different times. A few different disadvantages of data redundancy, include:

- a) Unnecessary enhancement of the database volume;
- b) EHRs inconsistency occurring
- c) Abated EHR systems competency;
- d) Deformation of EHRs may occur;

Several reasons and ways for the data redundancy to occur are detailed below:

*When copy-paste:* Copy-paste is one of the common data capturing tendencies around users if the information is available in any previous record. Often healthcare professional copy-paste EHRs from other providers or previous notes when capturing present patient assignment. In this way, between a brief EMRs, one anticipates accomplishing massive DR. In our belief, traditional conversation and expression of massive datasets provide preferable results in text excavation. Therefore, the DR might aggrieve the text excavation that can affect massive textual datasets. In the EHRs migration process, it is essential to determine if there is any DR to diffuse a preference, which will corrupt the data structure or introduce DR on the text excavation.

The EHRs contain beneficial electronic medical records captured by providers. Beyond the straight use, EHRs introduce as a clause of care, EMRs repository, analysis, research and medical invention. In particular, EMRs refer to copious amounts of clinical history, treatments, syndrome, prognostication, as these are often not inserted in the modelling part of the EMRs. These historical data can be discovered in the formation of free text semi-structured narrative distribution. The

emergence of the copy-paste capability between notes, along within the EHR systems, provided advantages for meaningful documentation. Although this functionality improves the documentation quality, it also may provide faults in the documentation process. According to Raphae *et al.* (2013), four types of patient comments with possible redundancy are:

- a)** Consumer comments (78%);
- b)** Improvement notes (54%);
- c)** Demographic comments (30%);
- d)** Release notes (30%);
- e)** Analysis of DR over these comments across the time;

Several dimension matrixes for determining DR are available for narrative text. Sequence classification methods, such as the one proposed by Zhang *et al.* (2011) are, “accurate yet expensive due to the high complexity of string alignment even when optimised”. The lower scathing DR dimension metric introduces collective words, apprehension or significant coincidence bigrams. Although semantic similarity could be determined by using these methods, they do not provide particular vindication for copy-paste performances that can regenerate the entire contexts.

The majority of bioinformatics semantic similarity algorithms are founded on detruncating inconsequential smaller substrings to introduce optimistically deliberate in progressive process classification for sequential consistency to interexchange efficient sub-sequences. Bigger datasets yield preferable consequences in the narrative excavation according to our traditional knowledge. Intrinsically truth is a bigger dataset containing more precise data dimension for the integration process. For efficient integration, the large dataset must be maintained in such a way to integrate records from corresponding sources. According to the literature, the crosspiece health domain’s large EHR batches yield impoverished language structures. This issue often happens due to the EHR systems adaptation when introduced to compensate for the impoverished quality

of EHRs from heterogeneous sources, by associating assiduous strings from other domains or statistical systems (for example, automatic translation).

*EHRs redundancy on text mining:* Generally, redundant data batches demonstrate diverse statistical models, rather than non-redundant EHRs, consequent to their data generation allocation. Therefore, it is important to identify whether this issue reduces data mining performances, such as arrangement determination and theme designing. Redundancy is simulated by indiscriminately data patterns and replicating them while the next redundancy regulator is identified.

*Controlling strategies for EHRs redundancy while migration:* The principal aims to reduce EHRs redundancy in the migration process is to identify the biggest probable subset from the EHR batches where the upper is constrained on the significant subset. In this procedure, EHRs migration strategies rely on the fundamental metadata rules to identify and address the redundancy, where the fundamental metadata rules satisfy the health standards of the organisations. The fundamental metadata rules control and produce the latest EMR batches. The substance-founded mitigation strategy generally confides on biometric identification and can consequence in data batches to identify the data redundancy ranges.

*Quantifying redundancy in the EHRs batches:* Compared to paper-based records, electronic notes have many advantages, but it may be time-consuming when capturing and enhancing the redundancy probability during integration. According to Stetson *et al.* (2008), “Not all of the information in EHRs is likely to be useful to the clinician”. However, some types of data redundancy are momentous and disinterested or detrimental, but inconsiderable is conversant about the amount of redundancy present. Various contentions could be made according to the EHRs redundancies advantages and disadvantages (for example, a provider’s iteration notes may reassess and ameliorate upon composite visits).

Other cognitive advantages of these procedures are to ensure the periodic appraisal of the providers. However, data redundancy is concerned as high risk, because it is difficult to balance the actual data dimensions and the ranges that are determined as disadvantages. These issues can increase the data inefficiency, increase the data error's proclamation and can completely damage the data integrity capacity. According to Codd 2009, "a commonly used database design heuristic is that preservation of data integrity is contingent on the elimination of redundancy". The most common issue for the EHRs redundancy is when the clinical staff are just copy-pasting data from one source to another source without even doing the minimum DQ checks.

One is cognizant of the fact that copy-pasting saves much time, in the same way as typing or capturing commonly takes more time. Siegler *et al.* (2009) and Donnell *et al.* (2008) stated that "unfortunately, studies suggest that there are significant risks associated with copy and paste. These risks include the introduction of inconsistencies in the record and error propagation".

Data redundancy, however, remains to be very common and extensively acknowledged across EHRs, as few algorithms are present to identify and address the redundancy in the narrative EHRs (for example, sequence classification algorithms are one of them). Syntactic duplications can serve only as a representative for the analysis of speculative redundancy.

One of the common calculations of redundancy is to measure the actual size of new EHRs measured as the amount of string, which did not assimilate with the previous entry divided by the length of the strings. The outcome can be populated as an aggregated percentage value of the EHRs that been repeated from the previous entry.

***Measuring the data matrix level for redundancy analysis:*** Provided the data batches, the statistic of the data dimension is considered for the analysis, whether the EHRs abide by Zipf's law. Zipf's law is defined as data dimension conditions that have assignment chains between them. Extremely few conditions happened

along with entire data batches and these conditions can be single strings or semantic apprehension.

*Mitigation strategies for handling redundancy:* The metadata-based mitigation strategy leverages the note creation date, the note type and the patient identifier information and selects the last available note per patient in the data batches. This baseline ensures the production of non-redundant data batches, as there is one note per patient only.

### 5.2.3.2 System application failure in the Electronic Health Records

Integrating EHR system is a challenging and expensive undertaking (Karan *et al.* 2016). EHR systems do not only entail the IT integration, but it is also a business transformation for healthcare systems and clinical users from the paper-based system. EHR systems involve many moving parts, including a staggered implementation and transitioning off of traditional systems. For example, one of the EHR challenges HCOs encountered concerning provider numbers. When importing provider practice numbers into a spreadsheet or from spreadsheet to database, if a provider number had a zero at the beginning, sometimes that zero would be dropped. Therefore, it is very important to pay attention to simultaneously staging data for data retrieving and transformation, to do all the spot checking and provide training for the new EHR systems for meaningful usage.

The EHR systems allow providers to easily capture new patients' electronic records and refurbish all new entries. The EMRs may include health history, including the family history for any chronic diseases. Some other additional details are included, such as the initial reason for the issue, lab result, diagnosis, medication, allergies and other important details that help in decision-making and the possibility to share information across HCOs.

To prevent medical errors, EHR systems play a vital role as part of the overall healthcare processes to ensure the EHRs are secure and dependable. According

to Win *et al.* (2002), “the identification of safety requirements of EHR systems would also help to reduce errors”.

Determining unspeakable occurrences which can generate from EHR systems would support in exploring the risk. Different EHR systems use a different algorithm to determine certain risk analysis and confide on different incidents, such as correctness, accurateness, comprehensiveness, timeliness and consistency.

Arts *et al.* (2002) stated that “DQ has been defined as the totality of features and characteristics of a data set that bears on its ability to satisfy the needs that result from the intended use of the data”.

Poor-quality data have a cabbalistic consequence on decision-making procedures on Health Information Management Systems (IMSS); therefore, there is no DQ compromise wherever applicable.

### 5.3 Data Quality challenges in Electronic Health Records

In recent years, DQ has become a considerable focus of healthcare programmes, due to EHRs accountability enhancement. Large-scale data integrations are a combination of technical and business processes used to combine data from disparate sources into meaningful and valuable information.

A complete large-scale data solution approach delivers trusted data from DHS and the term referring to the requirement to combine the data from multiple separate business systems into a single unified view is often called a single view of the truth.

Three primary components of DQ are:

1. **Data Profiling:** Data profiling is defined as the act of data component analysis;
2. **Data correction:** Data correction is defined as the act of data component correction when data does not comply with the DQ standards;
3. **Data monitoring:** Data monitoring is defined as the ongoing procedure of implementing DQ standards in a set of prosodies significant to the HCOs



service policy, analysing the outcomes in a re-occurring model and accepting the corrective operation, whenever one overcomes the adoptable commencement of DQ;

DQ assuagement procedures determined the data incompleteness, including fulfilling the objective for diminishing the risk of subsequent similar instances. According to the literature, the secondary uses of EHRs in research provided the diverse data incompleteness dimension and a different methodology has been suggested to tackle them. Although these methods were successful in small data dimensions, they comprised many remonstrances when implemented for the large-scale data dimension. In this chapter, these remonstrances have been identified for these proposed methods and measurement and a model is proposed for using present solutions to handle those remonstrances. Figure 5.2 (researcher source) describes the DQ challenges and benefits, as below:

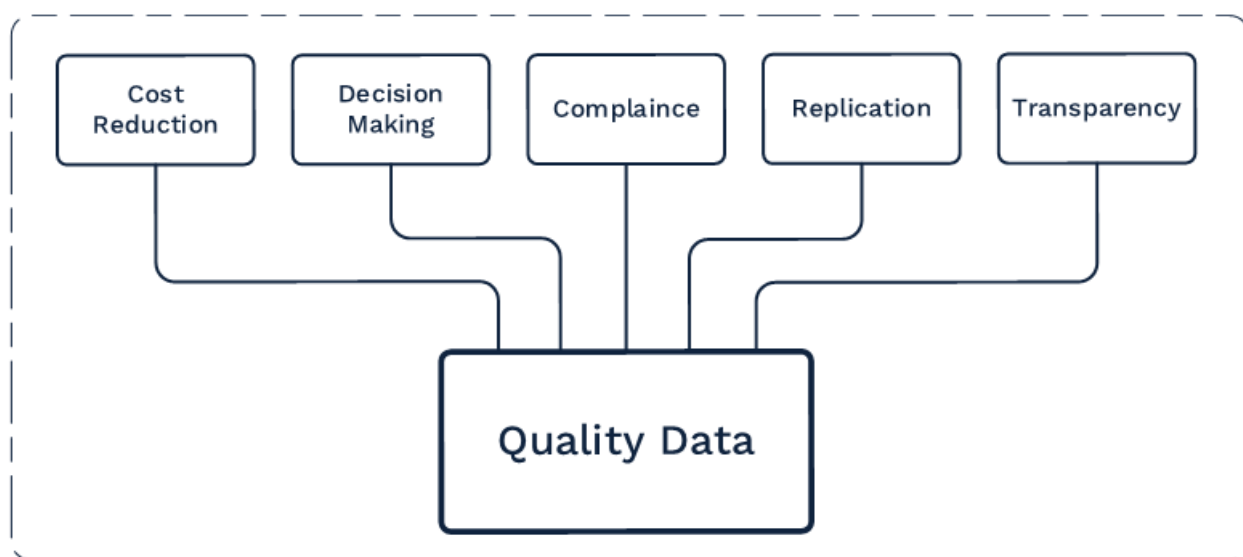


Figure 5.2: Data Quality challenges and benefit (researcher source)

In addition, it is the process of retrieving data from multiple source systems and combining it in such a way that it can yield consistent, comprehensive, current and correct information for the business work process, report and analysis. The objective of the big DI becomes even more important in the case of merging systems of different similar organisations. Among the key knowledge from literature is that big data initiatives, especially at the integration level, should

consider business logic system needs and challenges. It should also consider the data inconsistency challenge that specific big data initiatives will play in addressing the DQ issues in LSDB, the HCOs interoperability requirements and its priorities.

### 5.3.1 Why the Data Quality is a challenge in Electronic Health Records?

DQ is defined as an undivided component of the EHR systems. Several issues exist and remonstrance have been initiated, which have an extensive impact on the EHRs adaptation and inoperability in the Health Information Systems (HISs) ground. This concept has raised the remonstrance expression as the principal barriers for the meaningful use of EHR systems and cumulating poor-quality data.

*Quality Data is appropriate for use:* This concept remains as the best statement of the DQ. This determination concept ascertains even further beyond the general disturbance with the data accuracy, as it will raise other multi DQ dimension issues. Therefore, one can summarise that DQ is an apprehension of multi-dimensions.

The present data dimension structure was generally based on literature review, organisational wisdom and intuitional knowledge. Therefore, different DQ structures are present and the concept might differ according to different organisational structures. Alleviating prescription error is one of the core challenges in the medication process of HCOs. The development objective of the DQ framework consolation is to manage the low-quality data, which could exceptionally affect the healthcare services. These causes could even lead to the death of patients. The DQ concept relies on the organisational structure and actual use of the EHRs. In this way, the DQ concept relies on the specific application and this concept may not be meaningful to other application platforms.

Therefore, the key consideration is to present a design-oriented DQ concept, which will determine the nature of ISs. The other common issues of existing solutions as they are applied are a too generic adaptation and raised some attributes that

became inappropriate to EHRs. Figure 5.3 (researcher source) describes the EHRs data quality framework architecture, as follows:

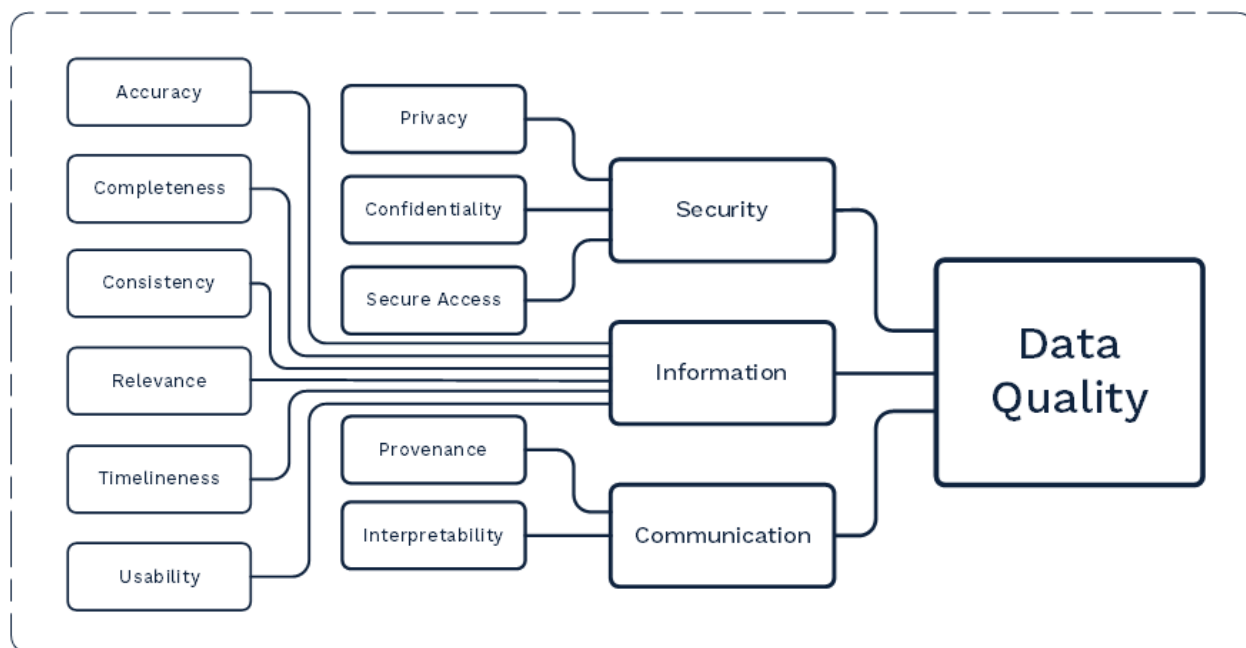


Figure 5.3: The EHRs Data Quality Framework architecture (researcher source)

Therefore, efficient EHR systems have been seen as a promising solution to issues in health ICT management, notwithstanding menaces that occur during the integration and transaction process. However, the fundamental issue of using the routinely integrated data meaningfully increases the low-quality data. These challenges raised the necessity for automatic DQ analysis technology and semantic EHR systems' inoperability.

### 5.3.2 How does the Data Quality impact into the Electronic Health Records?

EHR systems are designed to replace the paper-based record with electronic-based health records to deliver quality healthcare service. Since EHRs are now much more easily accessed than abstracting from the paper-based record, it is frequently used for other purposes, such as clinical effectiveness research,

predictive modelling, population health management and healthcare quality improvement. The use of EHR systems are expected to improve quality healthcare for the patient, but the benefit will only be achieved if the information that is captured in the EHR is of sufficient quality to be appropriate for usage.

According to Dean *et al.* (2009), the EHRs often contain errors that can impact research results, yet only 24% of clinical studies that use EHRs had a data validation section. For DQ measurement, there must be an understanding of the way that the data will be used. No generally accepted quantitative measure methodology of DQ is present. To summarise, high DQ are quality data that are fit for use in their ideal operable, decision-making, designing and operational character (Juran *et al.* 1999). Four high-level DQ dimensions are:

- a)** Correctness Measure;
- b)** Consistency Measure;
- c)** Completeness Measure;
- d)** Currency Measure;

These days, the EHRs landscape is rapidly innovating and often requires completely re-evaluation of the DBMS strategies. Due to the DQ assessment synchronisation methods, tactics, algorithms and report generation, can corroborate a general amicability of the possibility and limitation of EHR systems for strategic analyses, DQ improvement and experiment. The LSDB clinical information experimental diagram has arrived, along with high anticipation that new EHRs sources and DQ analysis methodology will provide results that cannot experiment in the traditional systems.

### **5.3.3 Why Data Quality became a challenge in Electronic Health Records?**

Today, perhaps the single greatest asset of healthcare organisations and EHRs is data to provide quality care service and without quality data, this would simply not be possible. The quality data or lack thereof has a direct and substantial impact

on the effectiveness of performing the daily service and only quality data drives operational effective service. The EHRs functions break down when the data has eroded and is no longer accurate and complete.

Unfortunately, and in most cases, organisations unknowingly operate with surprisingly high percentages of bad data that is outdated, inaccurate and incomplete or some combination of these issues (LexisNexis 2017). The integrity of EHR systems is at risk and present vulnerabilities in all healthcare programmes when data are incomplete, inconsistent and incorrect. Considerable progressive solutions are based on the four principal data substance grounds, as follows:

- a) *Completeness:*** Completeness is defined as the powerful spectacle of the consumer and data attributes support for HCOs structure and consent needs;
- b) *Consistency:*** Consistency is a data attribute standardisation process and an unseparated source of the view;
- c) *Accuracy:*** Accuracy is defined as dependable data sources, which accurately affect the actual meanings;
- d) *Governance:*** Data governance is the ongoing data standard maintenance procedure and management and observance. The data governance maintains three primary metrics for measuring quality to ensure data reliability:
  - a) *EHRs assimilate accuracy:*** The assimilate of the accurateness of each individual clinical attribute;
  - b) *EHRs Coverage:*** Data coverage has a two-dimension matrix:
    - ✓ Completeness of the healthcare services;
    - ✓ Completeness of each individual attribute rates, which provides a complete clinicians profile;
  - c) *EHRs Consistency:*** Consistency remarks the clinician's rate of changes;

Quality EHRs is indispensable for successful EHRs integration and implementation, even the anointing of HCOs. The EMR that many HCOs are using to run their daily

operation services is of a deteriorative quality, which is leading to undefined amounts of wasted time, money and resources and also keeps the HCOs permeable to a wide range of legal and security risks.

### 5.3.4 Different Data Quality challenges in Electronic Health Records

Today's HCOs are often overwhelmed by the DQ that should be helping them to stay abreast of patient interests and competitor tactics. Today HCOs are facing three main DQ challenges. These are caused by the increasing number of heterogeneous sources that must operate in as well as optimise those sources for the always-connected individual. To ensure the quality of the data collected through those heterogeneous sources, fully make sense of and leverage that data for healthcare purposes and combine the aforementioned into a single centralise of data set, the following is needed:

1. ***Collecting accurate data from heterogeneous sources:*** Data collection involves the usage of a number of sources and platforms. Examples include a website, import, call centres, capturing and tradeshow. Increasingly, this means managing the data collected through heterogeneous sources initiatives that take advantage of multiple patient touch points.
2. ***Leveraging data as an asset:*** Healthcare services must be able to connect the dots from the data they collect and use that to ultimately improve patient experience and drive revenue. However, HCOs are often limited by poor technology, poor processes and poor understanding of how the two work together. Scenarios include:
  - a) Legacy data environments that are unable to scale or keep up with increasing workloads;
  - b) Different internal processes regarding the organisation and standardisation of data;

- c) Not having the employee know-how to create proper processes or optimise the use of technology;

**3. *Creating a single centralised dataset:*** Perhaps the greatest challenge of all is creating a single centralised data set with the data collected. This requires both accurate and valid data as well as understanding how to properly leverage that data as a valuable asset.

A defective single source can generate DQ issues. The massive DQ challenges are raised, however, when EHRs integrate from diverse heterogeneous sources. Johnson *et al.* (2014) stated that “differences in how data are captured and stored in the original source-data system usually the institutional EHR or enterprise data warehouse, how data are extracted and transformed into the analytic data and the impact of data workflows or provenance can cause significant challenges in ensuring common data formats (syntax) and meaning (semantics)”.

#### **5.3.4.1 Incompatible data structure in Electronic Health Records**

Lack of timely communication between clinicians remain the substantial challenges in quality healthcare services. These challenges are becoming even worse in EHR management systems. As consumers, EHRs about healthcare is interred in oodles of unnecessary extender EHRs. Different HCOs have different standards and use incompatible formats according to others. Other issues, such as regulations are associated with privacy, timely care, over/under services and healthcare quality. The difficulty arises in EHR systems for smaller organisations, when changing the regulation and these individual providers need to customise the user interface to maximise the appropriateness. No centralised standardisation has been implemented for most of the EHR systems fundamental structures.

According to Mead (2006), Hammond *et al.* (2009) and Veli *et al.* (2009), “achieving broad-based, scalable and computable semantic interoperability across multiple domains requires the integration of multiple standards, which therefore must be mutually consistent, coherent and cross-compatible”. Unfortunately, the field

format has often been formulated in collateral and is therefore somewhat inappropriate with each other.

The EHRs could be immensely encrypted in a distributed system, similar to a “blockchain” model so that data could be accessed easily from anywhere using key values. This concept can reduce the cost associated with the privacy concern. The EHR systems may develop in such a way, which will allow both standardisations such as encrypted or non-encrypted data structures, controlled by the administrator and customisability of the user interface.

The exorbitant cost and waste of multiple incompatible proprietary systems could be eliminated and quality metrics would become more accessible to patients and payers. However, the study indicated initially that different HCOs have several standards and different major systems, which have emerged as critical issues and practical challenges. One of the main challenges in EHRs is the inherent difficulty to coherently manage incompatible and sometimes inconsistent data structures from diverse heterogeneous sources.

#### **5.3.4.2 Inconsistence data structure in Electronic Health Records**

Electronic health records (EHRs) have become a pervasive healthcare information technology. They replaced paper-based systems in many healthcare organisations and garnered rich health data, which holds great value for re-use. Inconsistency is defined as an information mismatch between various or within the same EHRs data source. EHRs integration systems must be able to control and maintain to provide quality data before extracting the data to the users. Dirty data is another key issue in DQ and this can be categorised into incomplete, redundancy and inconsistent data. In addition to EHRs design feature and function that can potentially contribute to suboptimal healthcare quality (AHIMA 2007), errors can result from the improper system use (Phillips *et al.* 2009).



The main contribution of EHR systems consists of a hybrid method based on Fuzzy-Ontology, mathematical modelling of a heuristic methodology, based on a combination of the perfect matching and similarity measurement for distributed concepts, unanimity techniques for inconsistency and conflict resolution performance to improve DQ. Creating a consensus between perfect matching and similarity measurement can be resolved to unite data inconsistency, mismatches and conflict ontology entity regarding diverse data sources. DQ, consistency, reliability, validity, accuracy, completeness and timeliness are the eventual significant list of EHR systems. Matching is a process of finding alignment between sets of correspondences with a semantic verification output of the matching process. This merging is a process of creating a new set of possibly overlapping data.

Usability errors occur as a result of system complexity, lack of user-friendly functionality (for example, confusing user interfaces), workflow incompatibility or limitations of the user (Hoffman *et al.* 2009). Faulty functionality could mislead clinicians, with a confusing screen display or incorrect values resulting from a programming error that incorrectly converts from one measurement system to another (for example, pounds to kilograms or Celsius to Fahrenheit) (Phillips *et al.* 2009).

However, the aim of the main task is to determine the best illustrate object to find a semantically fundamental equivalent motive in EHR systems. This provides a strong theoretical and practical framework to work with heterogeneous, complex, conflicting and automatic consensus methods for EHRs integration. A concept detection method could be semantically more meaningful, if the multimedia data integration process is a subset of the annotation process for trustable automatic data mapping, such as image, sound or video indexing by special code.

This means that each and every EHRs' amalgamated data is associated with a distinguishable adumbration characteristic. Therefore, conflict may happen on the EHRs data integration, if a diverse amalgamated data is associated with the same apprehension in the diverse EHR systems. Discovering the interrelation in EHR

systems is a significant phenomenon among entities manifested in diverse EHRs value. The similarity measurement often discovered that those conflicting entities are approximately identical among EHRs entities. Especially, with a chronic health circumstance, the EHRs statistics predict individual development and effectuation, since EHRs adoption better meets the needs of the growing modern community. Currently, five principles contribute to DQ. These principles are listed below:

- a) Formal concept analysis;
- b) Conceptual clustering;
- c) Generation;
- d) The grid-file for multi-attribute search;
- e) Semantic representation conversion;

This will be done to ease data access, extract information, search mechanisms, synchronise and establish semantic connections, filter data and provide different levels of security, provide data inconsistency solutions, resolve equivalently matching or conflicting information in multiple entities, resolve queries and achieve data compression and automatic EHRs integration simultaneously. Conflicts based on the occurrence of the same names or the same structures for different concepts were solved by using the concept of Potentially Common Parts (PCP) propagation. Other aforementioned conflicts, such as associated value conflicts and conflicts on a concept level, were also resolved using consensus methods. Specific criteria can be attributed to the representation. The criteria comprise:

*Comparability consists and specifies as designate of analytical DQ:* The variation in performance critical measures are observed to the Triad as the Triad emphasises collaborative data sets. Various analytical methods show three different types of divers can be identified in collaborative data sets. These variations are individual-level variations (for example age, sex and co-morbidities), provider level variations and random/residual variations. It is obvious that challenges in data availability and comparability issues are numerous in national and international comparisons.

**Completeness:** Completeness is defined as the extent to which all data components are integrated. One of the most important aspects for integration results is that the completeness is the guarantee of the appearance of all ingredients when integrating. Each data component should be captured in an EHR system so that a provider could create a dataset for a patient's characteristics. Otherwise, it will not be possible for a provider to create an accurate diagnosis for a patient characteristic. Incomplete data cannot provide an accurate diagnosis even when the corresponding information is supplied by EHR systems. The data will remain incomplete until the providers are not completed with the approximate group value associating to semantic categories. Finally, the inconsistencies of the patient diagnosis with appropriate values, with the inconsistency of any value in these appropriate values, are the intention acceptability of total designation for the EHRs completeness.

**Consistency:** Consistency is defined as the absence of any inconsistencies in the EHRs result and has solved all appearing conflicts among components when integrated. When stream data appended dependent interpretative variables from diverse heterogeneous sources, often integrate data refers to a similar subject but apprehends different inconsistent information. The time (T) dimensional observation data could be large and are asymptotically valid for a certain time. The data could be repeated approximately at the same value (an example: CPT code data) over time; such a situation is called a conflict.

**Identification:** Identification is defined as the structural similarity among entity sources and the EHRs result. The EHRs domain apprehensions often contain entities that have interrelated among the property value. This is defined that each EHRs entity is accompanied to certain concepts. If identical characteristics are associated with the identical apprehension in various ontologies, the conflict in the EHRs are also connected with a different associated EHRs value; this is also manifested as conflict.

**Timeliness:** Timeliness is defined as an association between the registration and diagnosis entity and the determination time to the observation diagnosis of the

occurrence statistical report. The EHRs statistics reports improve provider observation of the patient outcome. Overall statistics indicated that the use of EHR systems could sustain improvements to the productivity of healthcare services, such as timeliness, statistics report or invoices.

## **5.4 The important barriers and constraints in Electronic Health Records**

The overarching barriers to the EHRs framework are to tackle the indigent quality of data to provide a single, centralised and homogeneous interface for users to efficiently integrate data from diverse heterogeneous sources. DQ issues may arise when capturing raw data into EHR systems. The data flow process has several factors that influence the quality of information obtained from such datasets at a later stage. The purpose of the data collection processes is DQ management functions, including the data flow process application as well as data accumulation, warehousing process systems used to archive data and the analysis of the process of translating data into meaningful information.

The DQ may seriously affect patient care and even could lead to the death of the patient. These are the key challenges of eradicating treatment errors in the health service process. As patient safety is the key issue in healthcare service, using effective EHR integration systems and implementation can improve the DQ to reduce medical error. The main consideration for health data includes data accuracy and accessibility, as well as data comprehensiveness, currency, consistency, granularity, precision, relevancy definition and timeliness. DQ will empower the tendency of EHR systems; this emphasises the magnificence of implementing a design-oriented definition. The dimensions of the existing EHRs framework are basically based on historical reviews, understanding intuitive and comparative experiment.

The EHRs structures of orientation usually vary from framework to framework. For example, the actual use of the data depends on the definition of DQ. It, therefore,

the DQ also depends on the application type and that which may be deliberated in one application as good quality but may not be good for another. DQ has emerged as a crucial issue in many application domains. The objective of DQ becomes even more important in the case of patients who need to be identified and notified about important changes in drug therapy or in the case of merging systems of different and similar organisations. For the consolidation of information from diverse sources to provide a unified view of an organisation's data assets, is technically challenging.

The difficulty involves the way in which to practically combine data from disparate, incompatible, inconsistent and typically heterogeneous sources. The other difficult objective in EHR systems is that data has a structure, which is usually complex and cannot be treated as a simple string of bytes. Often data inconsistency occurs because the data structures may depend on other structures; therefore, on a distributed system, this kind of data management is very difficult. Another significant aspect of a health data integration system is data mapping. The system must be able to materialise data mapped from diverse sources. Optimally using routinely collected data increases poor quality data, which automatic mechanism would raise the need of the semantic interoperability as well as quality data measurement (Liaw *et al.* 2012).

Quality improvement and error reduction are two of the justifications for healthcare information technologies. Despite their concerns, HCOs are generally very interested in adopting and implementing EHR systems. A major concern of the success of the implementation is the large gap between planning for the introduction of the EHR systems and the medical maintenance systems in hospitals. The primary purpose of the successful EHR systems implementation depends on these application systems and maintains the application significantly to achieve the benefit desired and expected. The real barriers causing this gap may not be the availability of technology to the HCOs, as information systems are actually becoming available almost everywhere, but the deficiency in providing proper support before, during and after implementation of the EHR systems.

The financial constraints are another important matter of the migration from the paper-based health record to EHR-based systems. But, the human factors become even more important as the benefits are only anticipated after the successful integration and implementation of the EHR systems. Information security is most important for quality healthcare service. It improves the potential of EHRs as well as accuracy, accessibility, productivity and efficiency and to reduce the costs of healthcare and medical errors. Most HCO administrators are aware that it is time-consuming to migrate from paper-based records to EHR based systems. It is also important to change the provider behaviours and healthcare practitioners with regard to electronic healthcare systems, but time is also needed. A number of factors also need to be addressed regarding the successful implementation of EHR systems, such as attitudes, impressions and beliefs.

The most important factor is that it is essential to understand the reasons for and the purpose of the implementation of EHR systems in the whole subject (Pagliari *et al.* 2005). Research and statistics showed EHRs estimated potential savings as well as the costs of the widespread adoption of EHR systems. Important health and safety benefits were modelled and concluded that effective EHR systems implementation and networking, could improve healthcare efficiency and safety. It also showed that Health Information Technology (HIT) could enhance the prevention and management of chronic diseases, which could eventually double the savings while increasing health and other social benefits. The feasibility of introducing the EHR systems to improve the DQ is the meaningful association between the heterogeneous data source and the integration into HCOs to improve the healthcare service.

The integrity constraints are specified in the global scheme of data mapping, which can be used to promote the DQ on EHRs as well. The uncertainties are the other important integration aspect in EHRs that should be minimised to improve DQ. The most important barriers and constraints to promote high-quality datasets must solve the integration of the EHR systems and electronic health records to achieve maximum benefits for healthcare services. Finally, it is noted that the query answer in the context of data exchange, contributes to DQ.

## 5.5 The most meaningful association among heterogeneous data source in Electronic Health Records

The EHRs provide the possibility to ameliorate quality healthcare services, increase the clinical staff's performance, measure the HCOs performance base service and simplify health-related research. Concerns have been raised about the increasing recruitment challenges in trials, burdensome and obtrusive data collection and uncertain generalisability of the results. Leveraging EHRs to counterpoise this tendency is an area of extreme interest. A few indications have been raised to clinical practice about the three current states of cardiovascular clinical research (Jackson *et al.* 2016), which are as follows:

- a) The increasing recruitment challenges;
- b) Incommodious data collection;
- c) Uncertain generalisability;

These factors add to the increasing costs of clinical research (Eisenstein *et al.* 2008) and are thought to contribute to declining investment in the field (Jackson *et al.* 2016). The overall candid and widely accepted reasons for EHR systems are mensuration the experimental possibility and simplifying the consumer ingathering. EHR systems are nowadays implemented for this intention in many HCOs. Implementing the EHR systems to yield the consumer lists that could be suitable for research, is acknowledged as a way to convene the significant use of the EHRs standard in the United States (Blumenthal *et al.* 2010). However, incomplete EHRs should not be recognised for screening purposes for the complete list of acceptable criteria (Kopcke *et al.* 2013).

EHRs could, however, be promoted for pre-screening of consumers by age, gender and diagnosis, especially for the elimination of unmerited consumers and alleviating the overall screening compulsion in hospital experiments. Another excessive complex stratum comprises the re-use of EHRs from DBMS for general healthcare as a data model for decision-making. Using EHRs as the resource for

demographic health records, co-symptoms and chronic medications has many advantages over severally storing these records.

A focus on the significant use to achieve meaningful improvements in care is indispensable given the different testimony of the conveniences of EHRs. However, according to the literature, confined experimental research has concentrated on the conveniences of significant use. Recently, Jones *et al.* (2011) showed that “the use of Computerised Physician Order Entry Systems (CPOE) for electronic medication orders satisfying EHRs meaningful use criteria was associated with lower mortality rates for cardiovascular conditions”. Overall, the incompatible discovering and a deficiency of dynamic positive testimony increase concerns among anticipated EHR adopters. Quality of care exempted at clinics is affected by their functional representatives and the conditions in which they operate (Lehrman *et al.* 2010; Werner *et al.* 2011). The significant use of EHRs is defined in periods of the possibilities of the EHR system as a structure of hospital EHR systems. Time-varying disregard variables that happened simultaneously to EHRs adoption, such as quality enhancement implantations or clinical refunctioning.

In addition, data in the EHR systems compared databases did not include all process analysis for every HCO and the denominators for quality analysis did not always comply with the authenticity termination. The outcomes differed according to standard quality representation, with low-quality HCOs realising the most furtherance in quality.

## **5.6 The integrity constraints in Electronic Health Records**

The integrity constraints in EHRs are defined as the substance of defending the exactness, validity and thoroughness of EHRs in the healthcare environments. Availability is the substance that refers to the ability to use the EHRs or resources intended when required. These security substances, which one calls solid secure substances, are set by the legal domination that empowers and manages the EHR systems to conciliate its own security strategies.



Although the EHRs decontamination obligations can underpin the EHR semantics and confirm that it harmonises among its prospective data model, this approach does not ensure reliable measurements access for health data (Bertino *et al.* 2009). Integrity is considered to be the accountability of the proprietor of the EHRs and thus a somewhat demonstrable assertion. Confidentiality, as discussed earlier, is an assertion that provides to the consumer, not to HCOs, and consequently should be destined by legislation or other superficial obligations.

The EHRs must store or achieve at least as long as the consumer's life and apparently longer if queries are associated with the consumer's death. This burdens momentous obligations on the storage of the data. The data must be accessible even after archived of EHRs. But, more relevant to this article, the digital signatures must also remain indisputable for the corresponding epoch, so that data entered and digitally signed can be verified decades afterward.

## 5.7 The uncertainties in Electronic Health Record Systems integration

Healthcare service systems are perplexing systems functioned by reducible and irreducible uncertainty and clinicians face both types of incertitude when they care for patients (Han *et al.* 2011; Holly *et al.* 2013). The uncertainties are the other important integration aspect in EHRs that should be minimised to improve DQ. The non-linear interdependencies in perplex social systems are a principal accomplice to uncertainty, specifically uncertainty that cannot be diminished with information (Holly *et al.* 2013). This perplexes science and its concern of uncertainty in data analysis and exposition efforts. Drawing on theories from the medical and clinical literature, seven groups of appreciations of uncertainty, are as follows:

1. ***Stochastic Uncertainty:*** Stochastic uncertainty: Stochastic uncertainties are defined as random variability in outcomes between identical patients.
2. ***Reduction uncertainty:*** Reduction uncertainties are defined as reducible through data collection or data processing (example data provides a clear

overview of the risks). For uncertainty that is reducible, data collecting and processing are commonly feasible. When the EHR requires availability, uncertainty reduction is an essential technique for handling uncertainty. The encouragement for the uncertainty reduction denomination came from this traditional view of uncertainty and its management.

3. *Parameter uncertainty*: Parameter uncertainty is defined as the uncertainty in the estimation of the parameter of interest. This is second-order uncertainty.
4. *Heterogeneity*: Heterogeneity is defined as an inconsistency between consumers that can be attributed to the characteristics of those patients. In other words, variability observed or explained heterogeneity.
5. *Absorption uncertainty*: Absorption uncertainties is defined as irreducible or not resolvable with information. Irreducible uncertainty is present in the patient care processes, illness and disease trajectories and in the interactions between the two.
6. *Hybrid uncertainty*: Hybrid uncertainties is defined as an elective conspectus of the data revealed as an unanticipated third category of clinicians. Importantly, the hybrid category captured the uncertainty aspects of physicians that were not sufficiently narrated by the summarisation or exploitation categories. These clinicians observed information captured in the EMRs as critical to their work. At the same time, they observed the interchange of information and the creation of new knowledge with patients as a critical part of their practice.
7. *Structural uncertainty*: Structural uncertainties is defined as the assumptions inherent in the decision model. In other terms sometimes employed as model uncertainty.

Uncertainty reductionists stoop to demonstrator high levels of EHRs use; uncertainty absorbers stoop to demonstrate low levels of EHRs use and clinicians exhibit both aspects of uncertainty (hybrid), stopping to demonstrate medium levels of EHRs use. Complication knowledge accommodated a commencement pinnacle for determining the dissimilation in the way that physicians view the role

of information in caring for patients and the way they comprehend and manage uncertainty.

An uncertainty determination can be recognised as “fit for purpose” in terms of relying on the pronouncement of the designing ascertains to patronage. Uncertainty experimentation can serve two principle intentions, as follows:

- a) Evaluate dependence in a selected method of the process;
- b) Determine the importance of accumulating associated information to better decision-making;

The methodical investigation and efficient reporting of uncertainty are the blueprints of good designing strategy. All designing practices should include an uncertainty measurement as it pertains to the decision issues being traced. The decision-makers role should be deliberated when providing uncertainty measurements. The analytic perspective narration should include a circumstantial representation, associating that which is estimated about the decision-makers authority to obstruction or reconsider decisions and to intermediation or credential with research (Andrew *et al.* 2012).

In particular, stochastic (first-order) uncertainty is prominent from both parameters (second-order) uncertainty and from heterogeneity. Furthermore, each conception is disputed to have a corresponding form within a “**regression-type**” pattern in statistics. The term “**parameter uncertainty**” is not the same as the uncertainty around the accomplishment of a specific occurrence or consequences. Technicality is to narrate presumptions associated with parameter measurement and the re-narrative of uncertainty metamorphoses within the healthcare decision sampling field and in elimination, to associated fields. Stockholders should be acquainted of this and ascertain to cautiously identify their use of methodology to eliminate feasible conundrums (Andrew *et al.* 2012).

## **5.8 Data materialisation in the Electronic Health Record Systems integration**

The EHRs manifested healthcare LSDB is instrumental to enhance the entire quality of the healthcare service and to manage the financial data. Challenges to

using EHRs in the healthcare experiment have been determined, associated with which are DQ data affirmative, data capture completeness, heterogeneity between actual work process, system processes and constructing a developing knowledge across systems. DQ and validation are key factors in identifying whether EHRs might be compatible with data resources in healthcare experiments. Figure 5.4 (researcher source) describes the data materialisation in large-scale EHR systems, as follows:

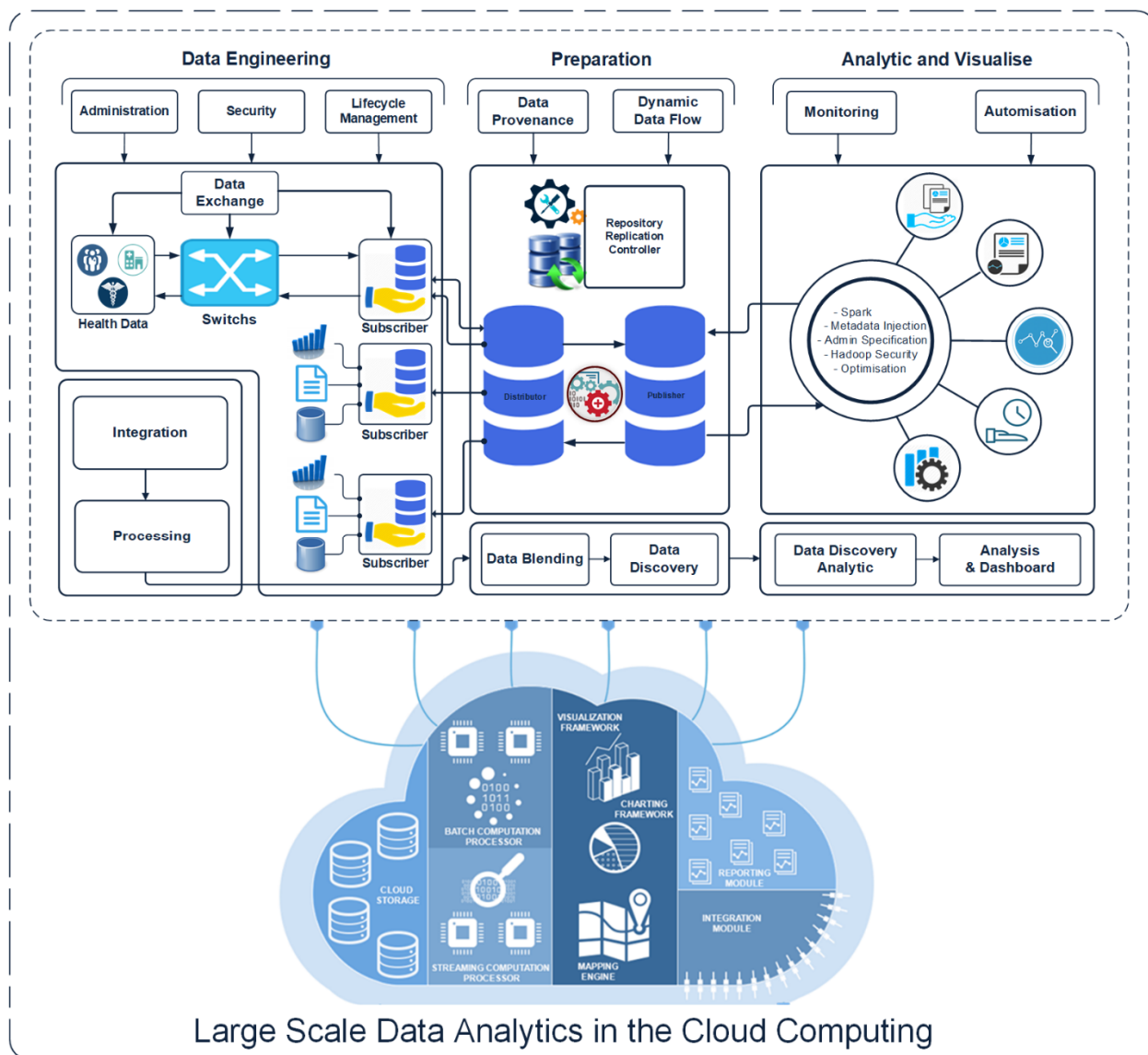


Figure 5.4: Data Materialisation in Large-Scale EHR Systems (researcher source)

Anxieties about codification imprecisions or preference commenced by the elimination of codes, conducted by billing inducement, rather than healthcare, may

be diminished when healthcare clinicians capture data directly into the EHR systems or when EHRs are used in all areas of the EHRs systems, but such systems have not yet been extensively accomplished. The potential for impediments in EHRs penetration is a significant consideration when EHRs are used in healthcare experiments. EHRs may comprise data initially collected as free text that was later captured for the EHR. Thus, coded data may not be in existence for consumer determination/recruitment during the admission. Likewise, data may generate weeks or months after the patient discharge. In nationally integrated systems, data availability may also be delayed. These delays may be critical depending on the intention of data extracted from the EHRs.

Consumers may be treated by several healthcare clinicians who manage independently from each other. Such patients may have more than one EHRs and these EHRs may not be linked. This heterogeneity adds to the complexity of using EHRs for clinical trials as data coordinating centres have to develop processes for interacting or extracting data from any number of different systems.

Differences in quality, non-standardised terminology, incomplete data capture, issues related to data sharing, data privacy, lack of common data fields and the inability of systems to be configured to communicate with each other, may also be problematic. Privacy issues and information governance are among the most complex aspects of implementing EHRs for clinical research, in part because attitudes and regulations related to data privacy vary markedly around the world.

Data security and appropriate use are high priorities, but access should not be restricted to the extent that the data are of limited usefulness. Access to EHRs by regulatory agencies will be necessary for auditing purposes in registration trials. Distributed analyses have the advantage of allowing data to remain with the individual site and under its control. EHRs are useful data sources to support comparative effectiveness research and new trial designs that may answer relevant clinical questions as well as improve efficiency and diminish the cost of cardiovascular clinical research.

Finally, the sustainability of EHRs in clinical research will largely depend on the materialisation of their promised efficiencies.

## 5.9 The query answering in Electronic Health Records

Query Answering (QA) is a computer science discipline within the fields of information retrieval and Natural Language Processing (NLP), which is concerned with building systems that automatically answer questions posed by humans in a natural language (Frank 2014). The analysis of search engine query logs is especially useful for understanding search behaviours of users (Kumar *et al.* 2009), categorising users' information need (Broder 2002), enhancing document ranking (Joachims 2002) and generating useful query suggestions (Mei *et al.* 2008; Lu *et al.* 2009).

EHRs search queries are in general much more sophisticated than Web search queries. It is also more challenging for the users to formulate their information need well with appropriate queries. This implies an urgent need to design intelligent and effective query reformulation mechanisms and social search mechanisms to help users improve their queries automatically and collaboratively. Data from EMR/EHR systems are increasingly disseminated, for purposes beyond primary care and this has been shown to be a promising avenue for improving research. This is because it allows data recipients to perform large-scale, low-cost analytic tasks, which require applying statistical tests (for example, to study correlations between BMI and diabetes), data mining tasks, such as classification (for example, to predict domestic violence) and clustering (for example, to control epidemics) or query answering.

To facilitate the dissemination and re-use of patient-specific data and help the advancement of research, a number of repositories have been established, such as the Database of Genotype and Phenotype (dbGaP), in the United States and the United Kingdom, and Biobank in the United Kingdom. Part of the reason is the lack of specialised EHRs search engines and longitudinal collections of query logs.

## 5.10 Different approaches to handle Large-Scale Data

Many researchers and software providers across the world have undertaken initiatives aimed at handling and addressing the big data issues in LSDB to meet the interoperability need of the organisations. It is important to examine the approaches that have been taken by these organisations that can be learned. Publicly available documentation on the approaches taken by these organisations that have integrated big data to support the interoperability of Information Systems (ISs), was reviewed. The number of publications reviewed was constrained by the limited number of documents available in the public domain. Enormous amounts of big data are becoming increasingly accessible through the LSDB adoption of ISs. DI and interoperability have started to play a key role in Cos to run their business process.

Much of this can be attributed to business data transitioning from a volume-based incentive model to a value-based model. With the way that HCOs are set up today, providing value-based business logic is largely unsustainable. HCOs have had to alter priorities from business expectation to the customer. Large-scale data are a significant resource for data and business workflows. However, because many big data domains are unable to talk to each other or talk to applications outside their own silos, data within the big data remain largely inoperable.

## 5.11 Summary

This chapter provided an in-depth exploration of the EHRs critical issues and DQ challenges. The chapter began with the detail analyses of different type of critical issues, including the source and reason for the critical issues and then the way that these impact in the DQ challenges for a successful EHRs implementation. This was followed by a review of diverse organisations and researchers involved in EHRs integration activities. The various ways in which EHRs critical issues and DQ challenges can be classified were explored. The general benefit of the EHRs integration and the DQ challenges that negatively impact on the EHRs

implementation were identified. An overview of important barriers and constraints was presented, including the most meaningful association among heterogeneous data sources. The integrity constraints and uncertainties were also discussed. The way to do the data materialisation and the final important issue in the context of data exchange, namely query answering, have been discussed thoroughly.

## 5.12 Conclusion

Investments in EHRs have the potential to transform the healthcare sector. However, the critical issues including the DQ challenges have a negative impact on the EHRs potential. DQ can reduce the complexity associated with the need for the EHRs integration from diverse heterogeneous sources. Figure 5.5 describes the combined outcome of Chapter Three, Chapter Four and Chapter Five, as follows:

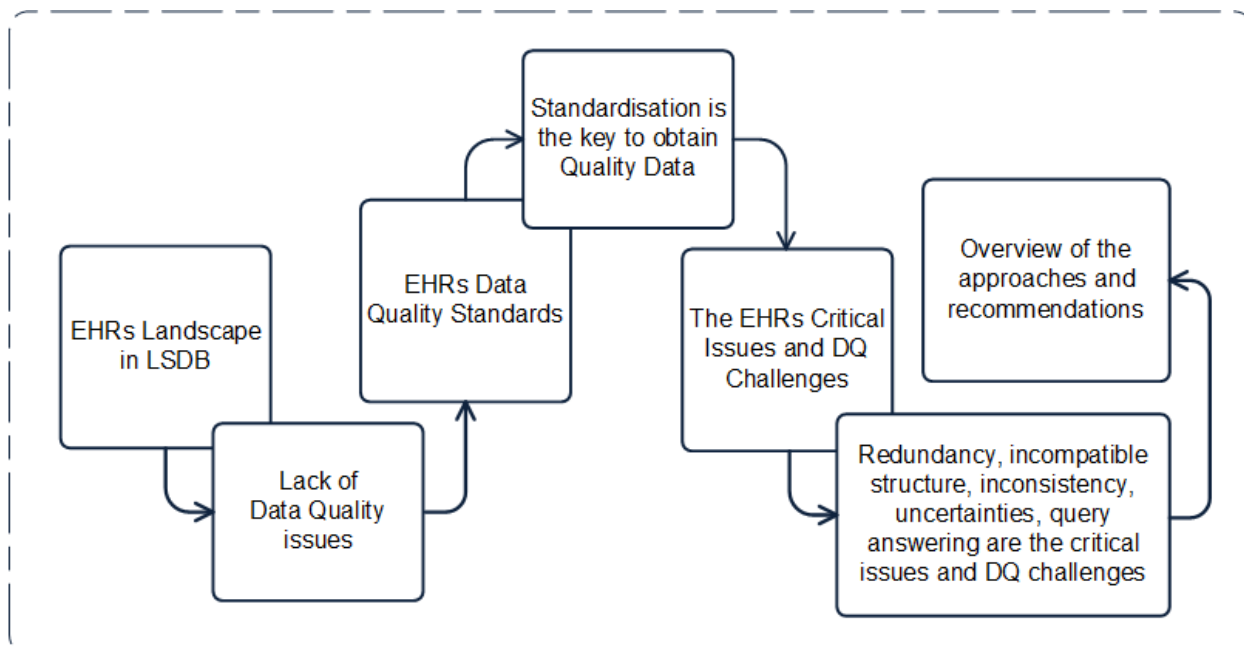


Figure 5.5: The combined outcome of Chapter Three, Chapter Four and Chapter Five

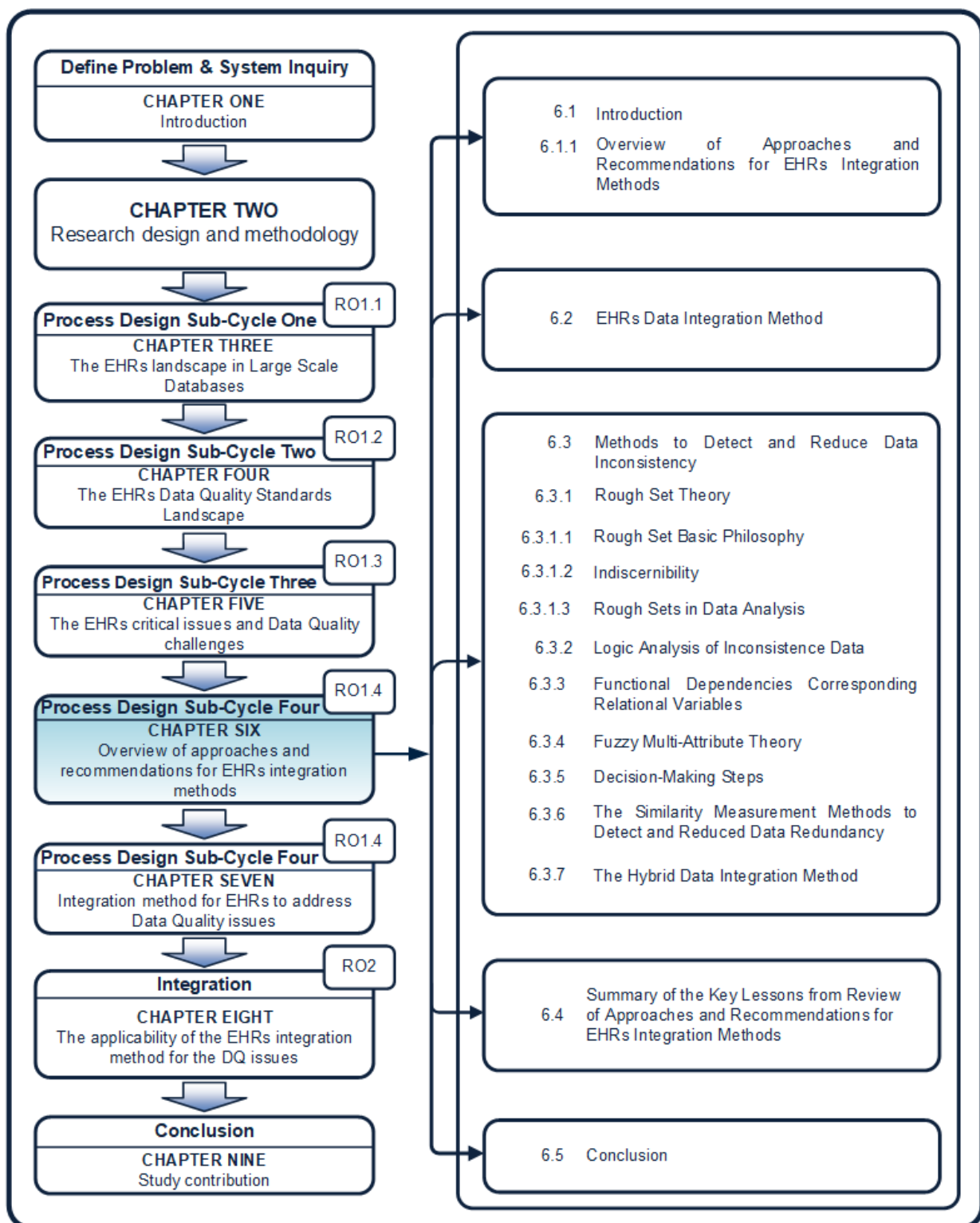
The DQ in EHR is one of the key components for seamless facilitation of the integration and successful implementation. The current literature showed, however, that the critical issues and DQ challenges can be difficult and sometimes impossible due to the different HCOs standards and incompatible data sources,



hence the need to implement new interoperable EHR systems. The prevalence of DQ standards integration activities is one of the critical issues, one has to contend with when integrating the appropriate set of integration standards to implementation.

Over the years, several International Standard Development Organisations (ISDO) have emerged to cater to the diverse needs of the healthcare domain. As illustrated in figure 5.5, the first step in EHRs integration systems of such a method is to review the approaches and recommendations for EHRs critical issues and DQ challenges. This is addressed in Chapter Six.

## CHAPTER SIX: Overview of approaches and recommendations for EHRs integration methods



Outline of the Chapter Six

## CHAPTER SIX

### 6.1 Introduction

The aim of this chapter is to review the initiatives and approaches for EHRs Integration Methods (IMs) that have been integrated into LSDB. This is to determine the lessons that can be learned from the processes they followed. The chapter maps to Sub-Cycle Four of the DSR process, described in section 2.3.2.2.4 and highlighted in figure 6.1.

Many researchers and Commercial Organisations (COs) across the world have undertaken initiatives aimed at determining and addressing the DQ issues in EHRs to meet the interoperability need of the HCOs. It is important to examine the approaches taken by these COs that can be learned. Publicly available documentation on the approaches taken by these COs to have integrated EHRs to support the interoperability of HISs, was reviewed. The number of publications reviewed was constrained by the limited number of documents available in the public domain.

The review is structured in order of importance. Section 6.2 provided an overview of EHRs integration methods. Section 6.3 reviewed the methods to detect and reduce data inconsistency whereas section 6.3.1 – Rough set theory, section 6.3.2 – Logic analysis of inconsistent data method, section 6.3.3 – Functional dependencies corresponding relational variables and section 6.3.4 – Fuzzy multi-attribute theory, is the overview of the different methods to detect and reduce the data inconsistency.

A summary of key lessons that can be learned from a review of approaches and recommendations for EHRs integration methods is provided in section 6.4 whereas section 6.5 concludes the chapter.

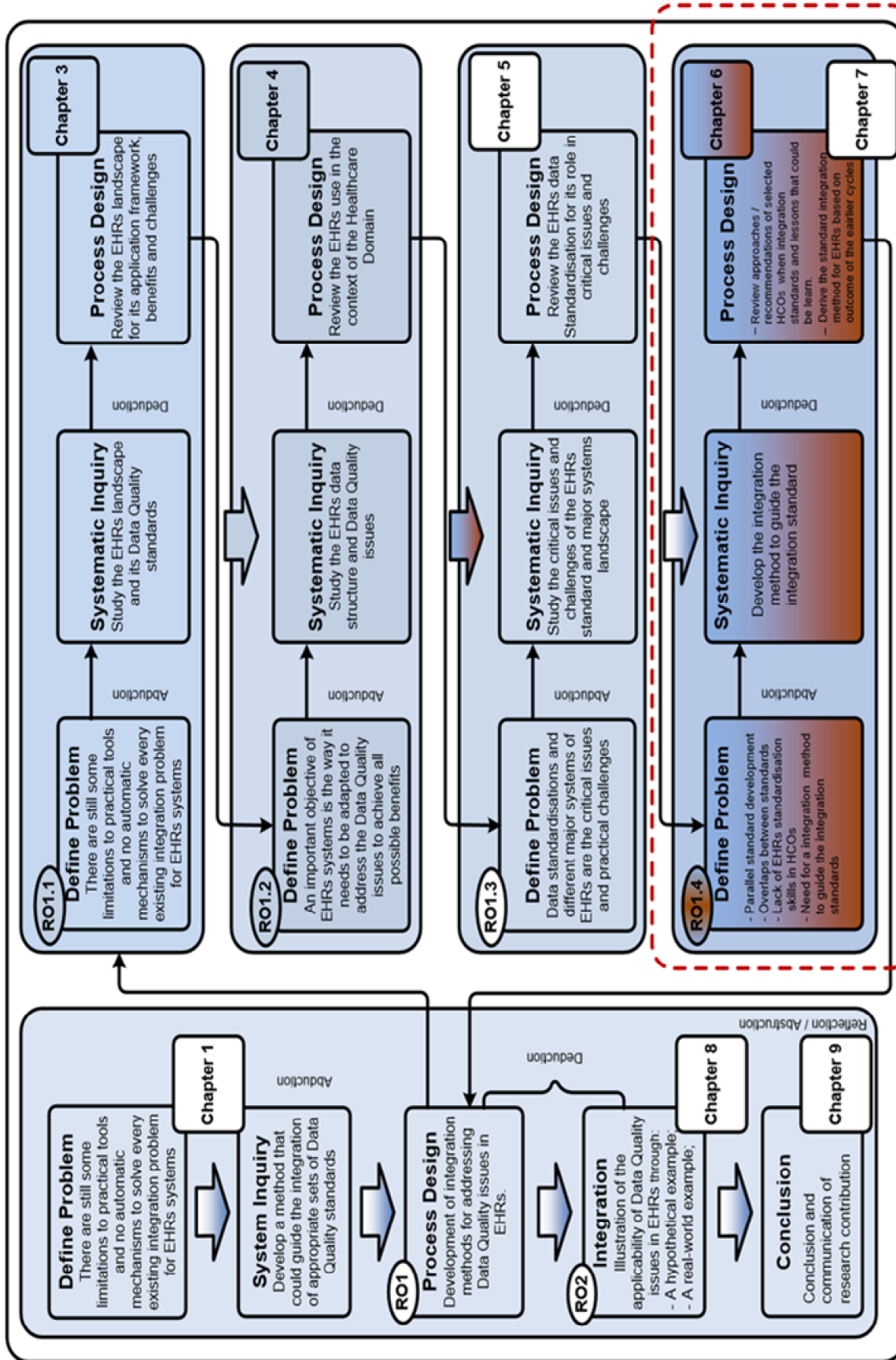


Figure 6.1: The position of Chapter 6 in the design science research process used in this study

## 6.1.1 Overview of approaches and recommendations for EHRs integration methods

In our daily life, numerous data collections composed of unstructured text with uncertain incidents may cause data inconsistency. In healthcare, EHRs are becoming widely accepted as the efficient standard of storing medical information (Hsiao 2012; Cimino 2013).

The information contained within these EHRs not only provides direct health information about patients but is also used to monitor hospital activities for medical billing and demographical health management. For instance, the web domain is largely composed of pages with text paragraphs. Other examples of document collections with this structure are text notes, provider comments, patient history and medical records. The majority of these document collections are accessed these days by using information retrieval systems devised to index text contents so that users can search by expressing their information need using textual keywords. However, the vast amount of non-text data available in several document collections may lead to the design of other effective ways of accessing and finding information.

## 6.2 EHRs data integration method

Enormous amounts of healthcare data are becoming increasingly accessible through the large-scale adoption of EHRs. Healthcare integration and interoperability have started to play a key role in healthcare. Much of this can be attributed to healthcare transitioning from a volume-based incentive model to a value-based model. With the way in which hospitals are set up today, providing value-based care, is largely unsustainable. Healthcare organisations have had to alter priorities from the provider to the patient. To better provide for patients and to maintain compliance with the HITECH Act and other mandates, EHRs have become an integral part of the majority of hospitals. EHRs are a significant resource of clinical data and clinical workflows. However, because many EHRs are unable

to talk to each other or talk to applications outside their own silos, data within the EHRs remain largely inoperable.

Integration is the exchange between providers, payers, vendors and other important players that bring data or function from one application to another. Integration is important because when one looks at the quantity and diversity of data involved in the healthcare system, it would be virtually impossible to process or analyse without breaking through the data silos. Traditional health IT systems, such as EHRs utilise completely different technical and semantic standards to depict and housing data. This makes it extremely difficult to correctly and simply integrate data from various conflicting systems.

An EHR stores a linear record of patient health info. It includes patient demographics, progress notes, problems and medications, vital signs, past medical history, immunisations, lab data and radiology reports. By offering a clinician the EHR record, it helps automate and streamline their workflow. Given that a large part of the picture exists within the EHR, one would have expected that to integrate into and out of EHR systems would have been easier for other sources. However, for the majority of hospitals, it requires a skilled team of experts to bring the idea to fruition. This typically includes eight-team groups, as follows:

- a)* Project manager;
- b)* Operational owner;
- c)* Systems administrator or network engineer;
- d)* Interface engine analyst;
- e)* EHRs application interface analyst;
- f)* EHR web service analyst;
- g)* EHR analyst;
- h)* Support staff;

In addition to orchestrating all the involved parties, it is also important to ensure that the hospital is meeting all the healthcare data standards interpretations.

Despite its complexity, seven linear lists have to be followed for the project initiative, as follows:

- a) Planning and Paperwork:* Preparing contracts, business associate agreements and kick-off meetings;
- b) Gather requirements:* Sampling HL7 messages, sample JSON, data dictionaries, APIs and associated documentation;
- c) Infrastructure/VPN:* Infrastructure should be spun up with VPN creation and verification;
- d) Setup interface:* Making sure that the front end of an organisation's endpoints is set up, including the interface;
- e) Testing:* Using the sample messages collected previously to review and acknowledge with inbound and outbound to resolve any potential issues;
- f) Go live:* Migrating to production and deploy. Data feeds are opened and the integration will allow other sources to populate the system;
- g) Ongoing support:* Unfortunately, this is not a set it or forget it situation. Making sure the IT support staff is alert, well-trained and ready to tackle all the questions that will inevitably arise from this implementation;

With all the hands that have to go into creating and maintaining this integration, this can serve as a major bottleneck for health systems. It could result in a lower quality of care for patients and minimal return on investment for providers. An EHR's complexity can fall into three main buckets, as follows:

- 1) Custom data mappings;
- 2) Security of data connectors;
- 3) Projects requiring different degrees of project management;

Many organisations feel forced to limit the data they integrate because of a lack of understanding of that which is possible. These issues become worse and more complicated with scale - the total opposite of which healthcare actually needs.

### 6.3 Methods to detect and reduce data inconsistency

Data inconsistency detection and reduction are the central issues for the DQ issues in EHRs for LSDB. However, it becomes far more challenging when data is centralised and distributed, in which inconsistency detection often necessarily requires integrating data from diverse heterogeneous sources into the LSDB. To provide quality care service, the EHRs integration systems must be able to prevent dirty data from the heterogeneous data source, while integrating. Inconsistent data is one of the dirty data that occurs because of the different data structures from various HCOs data sources and different standards.

Generally, factual dependencies are among the heterogeneous sources in the EHR value and characterise the same objects, which are defined as inconsistency in the data value level. If the information is no longer reliable in the healthcare domain and consumes more cost and efforts, then the issues rise up and identify themselves as data inconsistency. The rapid development of EHRs technology currently has transformed most data sources from paper and manual base to electronic-based health records. Basically, heterogeneous data conflict between each other in the source level at the schema, representation and value level. DQ is recognised as one of the most important issues for EHRs integration systems.

Detecting and identifying the inconsistency are the central technical issues for the DQ concerns to successful EHRs implementation. This challenge raised the need for a DQ tool that is effective in interoperability; the tool has to be automatic and efficient in inconsistency detection and addressing methods. The four latest methods to detect and reduce data inconsistency have been identified as follows:

- 1) Rough set theory (Jiawei *et al.* 2011; Krzysztof *et al.* 2012; Ewa 2013);
- 2) Logic analysis of inconsistent data method (Peck *et al.* 2011; Hervé *et al.* 2012; Cheng 2014);
- 3) Functional dependencies corresponding relational variables (Cataldo *et al.* 2009; Jácome 2009; Bernhard 2013);



- 4) Fuzzy multi-attribute theory (Fang *et al.* 2010; Meimei *et al.* 2011; Abdolhadi *et al.* 2012).
- 5) The similarity measurement methods to detect and reduce data redundancy;

### 6.3.1 Rough set theory

Rough set theory (PAWLAK 1982) is defined as, “mathematical methodology that partial consciousness (for example, to vagueness or imprecision) concerned with the analysis and modelling of classification and decision issues involving vague, imprecise, uncertain or incomplete information”. The rough set concept can be defined by means of topological operations, interior and closure, called approximations. In this approach, vagueness is expressed by a boundary region of a set. The concept of the rough set was originally proposed by Pawlak in 1982 as a mathematical tool to deal with vagueness and uncertainty in the classification of objects in a set. The rough set is defined as a formal approximation of a crisp set in terms of a pair of sets, which gives the lower and the upper approximation of the original set. Its philosophy assumes that from every object of the universal set, one associates some information. Objects characterised by some information are indiscernible in view of the available information about them.

In the standard version of the rough set theory (PAWLAK 1991), the lower- and upper-approximation sets are crisp sets, but in other variations, the approximating sets may be fuzzy sets. Rough sets have been proposed for a variety of applications, including artificial intelligence and cognitive sciences, especially machine learning, knowledge discovery, data mining, expert systems, approximate reasoning and pattern recognition. Its philosophy assumes that with every object of the universal set, one associates some information. Objects characterised by some information are indiscernible in view of the available information about them. The indiscernibility relation generated in this way is the mathematical basis of the rough set theory.

Rough set-based data analysis starts from a data set organised in a form of a data table. Rows of the table describe objects of interest by means of attribute values. Objects characterised by the same attribute values are indiscernible. The indiscernibility relation is the mathematical basis of the rough set theory. Any set of all indiscernible (similar) objects is called a componentry set or granule (atom) and can understand a basic knowledge about the universe. Any union of some componentry sets is referred to as a crisp (precise) set - otherwise, the set is rough (imprecise, vague). Each rough set has boundary-line cases (for example, objects which cannot be with certainty classified by employing the available knowledge) as members of the set or its complement. Thus, rough sets, in contrast to precise sets, cannot be characterised in terms of information about their components.

With any rough set a pair of precise sets - called the lower and the upper approximation of the rough set - is associated. The lower approximation consists of all objects which surely belong to the set and the upper approximation contains all objects which possibly belong to the set. The difference between the upper and the lower approximation constitutes the boundary region of the rough set. Approximations are basic operations in the rough set theory. The rough set approach to data analysis has many important advantages. The main advantages of the rough set approach (Zbigniew 2004) are as follows:

- a)* It does not need any preliminary or additional information about data-like probability in statistics, grade of membership in the fuzzy set theory;
- b)* It provides efficient methods, algorithms and tools for finding hidden patterns in data;
- c)* It allows to reduce original data (for example, to find minimal sets of data with the same knowledge as in the original data);
- d)* It allows to evaluate the significance of data;
- e)* It allows to generate automatically the sets of decision rules from data;
- f)* It is easy to understand;
- g)* It offers a straightforward interpretation of the obtained results;
- h)* It is suited for concurrent (parallel/distributed) processing;

- i)* It is easy Internet access to the rich literature about the rough set theory, its extensions as well as interesting applications;

According to Pawlak (2010), a set is defined by its components (for example, it is defined if all its components are uniquely determined). For example, the set of all HIV tests are in two groups, negative (positive) is determined as unique and every result can be classified, without any doubt, as negative or positive. These kinds of notions are usually referred to as crisp. Obviously, all mathematical notions are crisp; otherwise, it would be impossible to prove any mathematical theorem. But in many other fields, the situation is not so clear. It is impossible to uniquely define the notion of a healthy or sick person. These kinds of imprecise notions are called vague. The concepts of vagueness are defined by a boundary region, which consists of all components that cannot be linked to the concept or its complement. For example, the concept of HIV results, negative or positive, is precise, because every result is either negative or positive.

If the same information is associated with some components in view of this information, these components are indiscernible. For example, if some patients suffering from a certain disease display the same symptoms, they are indiscernible with respect to the information about them. It turns out that this indiscernibility leads to the boundary-line cases (for example, that some components cannot be linked to the concept or its complement in view of the available information). This is because the vague concept has a boundary line case, for example, the components which cannot be classified as being with certainty components of the concept; vagueness is strictly connected with the idea of certainty or uncertainty.

The rough set theory seems to be well suited as a mathematical model of vagueness and uncertainty. Vagueness is a property of sets (concepts) and is strictly related to the existence of the boundary region of a set, whereas uncertainty is a property of the components of sets. In the rough set approach, both concepts are closely related due to the indiscernibility caused by insufficient information about the world we are interested in. Furthermore, the rough set

theory is related to discriminate analysis (Krusinska *et al.* 1992), Boolean reasoning methods (Skowron *et al.* 1992) and others.

The rough set theory has many interesting applications. The rough set approach seems to be of fundamental importance to algorithmic logic and cognitive sciences, especially in the areas of machine learning, knowledge acquisition, and decision analysis and knowledge discovery from databases, expert systems, inductive reasoning and pattern recognition. It seems of particular importance to decision support systems and data mining.

The main advantage of the rough set theory is that it does not need any preliminary or additional information about data (for examples, such as probability in statistics, basic probability assignment in the Dempster-Shafer theory, grade of membership or the value of possibility in fuzzy set theory). The rough set theory has been successfully applied in many real-life problems in medicine, pharmacology, engineering, banking, financial, market analysis and others. Particularly, in pharmacology, the analysis of relationships between the chemical structure and the antimicrobial activity of drugs has been successfully investigated.

Very promising new areas of application of the rough set concept will emerge in the near future. These include rough control, rough databases, rough information retrieval, rough neural network and others.

### **6.3.1.1 Rough set basic philosophy**

The rough set philosophy assumes in contrast to the classical set theory. The rough set has some additional information (knowledge, data) about components of a set. Consider as an example, a group of patients suffering from a certain disease. In the hospital treating the patients, data files contain information about patients, such as, for example, body temperature, blood pressure, name, age, address and others. All patients revealing the same symptoms are indiscernible (similar) in view of the available information and form blocks, which can be understood as componentry granules of knowledge about patients (or types of patients). These

granules are called componentry sets or concepts and can be considered as componentry building blocks (atoms) of our knowledge about the reality we are interested in.

Componentry concepts can be combined into compound concepts (for example, concepts uniquely defined in terms of componentry concepts). Any union of componentry sets is called a crisp set and any other sets are referred to as rough (vague, imprecise). With every set  $X$ , we can associate two crisp sets called the lower and the upper approximation of  $X$ . The lower approximation of  $X$  is the union of all componentry sets, which are included in  $X$ , whereas the upper approximation of  $X$  is the union of all componentry sets, which have a nonempty intersection with  $X$ . In other words, the lower approximation of a set is the set of all components that surely belongs to  $X$ , whereas the upper approximation of  $X$  is the set of all components that possibly belong to  $X$ .

The difference between the upper and lower approximation of  $X$  is its boundary region. Obviously, a set is rough if it has no empty boundary regions whatsoever, otherwise, the set is crisp. Components of the boundary region cannot be classified, employing the available knowledge, either to the set or its complement. Approximations of sets are basic operation in rough set theory and are used as the main tools to deal with vague and uncertain data.

### **6.3.1.2 Indiscernibility**

As mentioned in section 6.3.1, the starting point of the rough set theory is the indiscernibility relation, generated by information about objects of interest. The indiscernibility relation implies that due to the lack of knowledge one is unable to discern some objects employing the available information. This means that, in general, one is unable to deal with single objects, but have to consider clusters of indiscernible objects as fundamental concepts of knowledge.

The rough set approach to data analysis has many important advantages. Some of them are listed below:

- 1) Provides efficient algorithms for finding hidden patterns in data;
- 2) Identifies relationships that would not be found using statistical methods;
- 3) Allows both qualitative and quantitative data;
- 4) Finds minimal sets of data's data reduction;
- 5) Evaluates the significance of data;
- 6) Generates sets of decision rules from data;
- 7) Is easy to understand;
- 8) Offers straightforward interpretation of obtained results;
- 9) Most algorithms based on the rough set theory are particularly suited for parallel processing, but to exploit this feature fully, a new computer organisation based on the rough set theory is necessary;

The starting point of the rough set theory is the indiscernibility relation, generated by information about objects of interest. The indiscernibility relation implies that due to the lack of knowledge one is unable to discern some objects employing the available information. This means that, in general, one is unable to deal with single objects, but has to consider clusters of indiscernible objects as fundamental concepts of knowledge.

### 6.3.1.3 Rough sets in data analysis

A data set is represented as a table, where each row represents a case, an event, a patient or simply an object. Every column represents an attribute (a variable, an observation and a property) that can be measured for each object; the attribute may be also supplied by a human expert or the user. Such a table is called an information system. Formally, an information system is a pair  $S = (U, A)$  where  $U$  is a non-empty finite set of objects called the universe and  $A$  is a non-empty finite set of attributes such that a  $a: U \rightarrow V_a$  for every  $a \in A$ . The set  $V$  is called the value set of  $a$ . The modelling problem of the approximate reasoning process on the base of knowledge is included into experimental decision tables with uncertain,

imprecise and vague information. The Rough Set Theory solution is obtained by realising the following three stages (Zbigniew 2004):

- *Stage 1:* Data representation;
- *Stage 2:* Knowledge representation;
- *Stage 3:* Transformation of decision tables into timed approximate Petri nets;

As discussed above, one can summarise that the rough set theory is meant mainly as a new mathematical approach to discover patterns in data (for example, one is given a data set as a result of observations of some real-life phenomena and one's task is to find out hidden patterns in the data). The patterns are usually represented in the form of a set of decision rules. The rules can be used to explain the data (for example, to explain the phenomena or processes underlying the data which are usually interpreted as a description of some cause-effect relation).

### 6.3.2 Logic analysis of inconsistency data

Quality data is appropriate for use and presupposed for analyses; big data is used and the value of the data guaranteed. Data consistency refers to whether the logical relationship between correlated data is correct and complete. In the field of databases (Silberschatz *et al.* 2006), it usually means that the same data located in different storage areas should be considered to be equivalent. Equivalency means that the data have equal value and the same meaning or are essentially the same. Data synchronisation is the process of making data equal.

HCOs rely on data analysts to model customer engagement, streamline operations, improve production, and inform business decisions and combat fraud. The analyst needs to define criteria to divide the logs into logical segments. Verifying that criteria accurately in split records can be difficult, especially with collections of log files containing terabytes of data.

Another common data format is the so-called *block data* that was difficult to parse. In a block format, logical records of data are spread across multiple lines of a file. Typically, one line (the *header*) contains metadata about the record, such as how many of the subsequent lines (the “payload”) belong to the record. Data from third-party services often required a level of processing before analysis could begin. Identifiers useful for joining records across datasets were often missing in one or more data sets, inconsistent across data sources or incorrect for certain records. One hospital analyst recounted issues integrating patient medical data.

The easiest patient identifier is the Medical Record Number (MRN). One should have consistent MRNs in any data source but five to ten percent of MRNs are mistyped or incorrect or blank. In emergencies, a patient may get assigned a temporary MRN. Later it is reassigned but sometimes one forgets to reassign. One has to identify patients by other means, namely first name, last name, birthdates and gender. Several data points together might identify a single patient. There may be slight inconsistencies. The three types of inconsistency in identifiers during integration, are as follows:

- 1) Identifiers used slight variations in spelling or formatting that make direct matches difficult. For instance, a patient’s first name might be stored as *Alex* in one record and *Alexander* in another. Some analysts defined ad hoc rules (*fuzzy matches*) to detect similar items. The analysts then inspected the matches to verify that the records referred to the same entity.
- 2) Data sources used two different encodings to represent the same identifier. For instance, a province might be identified by its full name (for example, Eastern Cape, South Africa) or by its Province Information Processing Standard (PIPS) code (for example, 4). In this case, an analyst must find or construct a mapping between identifiers.
- 3) Identifiers used inconsistent units of measurement or class definitions. Multiple analysts described attempts to consolidate their respective company’s industry codes. Others described difficulties integrating geographic data with varied regional definitions. Similarly, many data sets use overlapping conventions for financial quarters. The situation is



complicated when sets of regions overlap and one standardisation does not subsume the others.

Data sets may contain a number of quality issues that affect the validity of results, such as missing, erroneous or extreme values. In some cases, observations contained missing or null attributes. In other cases, entire observations were missing from a data set. Missing observations were much more difficult to detect. Another common problem is heterogeneous data in a column, such as **a column with an expected type which may contain values of another type**.

This might occur due to errors in automated processes, such as log file generation, errors in human data entry or because of an explicit decision to overload the use of a column. For example, a database table at one organisation contained a longitude field that was empty for many of the observations. Instead of creating a new column, some analysts decided to overload this field to store additional data unrelated to longitude. This type of error also occurred when IT teams introduced a new field into a log file, breaking existing scripts that expect the files in a certain format.

Common assumptions included the way in which values were distributed within an attribute (was an attribute normally distributed?), what values were unique (were there duplicates?) and the way the different attributes related to each other (was X always greater than Y?). Other assumptions required domain expertise to verify. It was particularly difficult to understand the relationships among features spread across multiple databases. Also, many fields needed to be transformed before useful patterns would emerge. The threshold for when data sets were too big was obviously different depending on the tool used.

Hackers were less limited by large amounts of data because they could typically run distributed jobs over multiple machines. However, hackers were often limited by the types of analysis they could run because useful models or algorithms did not have available parallelised implementations. It is difficult for hackers to **take**

powerful algorithms that work on medium data and make them pluggable in the big data stack.

Analysts typically performed a sequence of operations that can affect the interpretation of results, such as correcting outliers, imputing missing data or aggregating data. These operations are often context specific, with no standards for each analysis. The patient episodes correspond to all visits to a hospital to treat a given symptom. However, the database did not contain an episode identifier associated with each patient visit. The analysts had to use heuristics, such as the duration between visits, to group hospital visits into episodes. This heuristic was imprecise, as hospitals may treat a patient concurrently for two different symptoms or for the same symptom after a long period of time.

Analysts often lost track of all the operations they performed and their rationale for performing them. It was often difficult to assemble these products into a repeatable, reliable and scalable process. Reconstructing a repeatable workflow is difficult without a coherent linear history of the operations performed. Even with a coherent history, an existing workflow may break when applied to new or updated input data. This new input data may contain nuances not accounted for that would cause existing code to break.

Finally, analysts reported that they wrote experimental code that could not run on large datasets or at the necessary speed in real-time systems. They, therefore, required the IT team to operationalise many of their workflows. Over the last few years, we see three factors driving an increasing demand for *hacker-level* analysts.

- **Firstly:** Constrained IT departments are making it necessary for analysts to be self-serving. When discussing recruitment, one chief scientist said *analysts that cannot programme are disenfranchised here*. IT support was prioritised for shipping products, not helping analysts to experiment on the code.
- **Secondly:** the increasing scale of data requires many organisations to perform in-database analytics. Analysis software tools such as R and Matlab do not

currently scale. Instead, analytic routines are performed within the data warehouse, typically in a shared-nothing parallel database (such as those offered by Aster, Greenplum or Teradata) or via Map-Reduce or related higher-level languages such as Pig.

Analysts, therefore, need to be adept at both statistical reasoning and writing complex SQL or Map-Reduce code.

- *Thirdly:* Organisations are frequently relying on multiple processing frameworks and tools as requirements evolve. For instance, some organisations will use relational databases to support interactive queries and analysis rely on Hadoop for batch jobs and processing log files and also require analysts who can build *prototype* models in R.

One clear implication of one's studies is the need for visualisation methods that scale. Scaling visualisation requires addressing both perceptual and computational limitations. Visualisations that render raw data suffer from overplotting with even moderately large data sets and certainly when applied to datasets containing billions of observations. Visual analytic tools must consider using density or aggregation-based plots, such as histograms and binned scatter plots for large datasets. One approach to improved scalability is to leverage existing data processing engines for manipulating data. As the scale and diversity of data sources increase within enterprises, an opportunity arises for visual analytic tools to improve the quality of analysis and the speed at which it takes place.

### 6.3.3 Functional dependencies corresponding relational variables

Detecting inconsistencies in distributed data are recognised as one of the most important issues for DQ (Wenfei *et al.* 2010). Given a database  $D$  and a set  $\Sigma$  of dependencies as DQ rules, EHR systems want to identify tuples in  $D$  that violate some rules in  $\Sigma$ . When  $D$  is a centralised database, there have been effective SQL-based techniques for finding violations. It is, however, far more challenging when

data in D is distributed, in which inconsistency detection often necessarily requires exchanging data from one site to another. The detecting techniques of the violations of Functional Dependencies Corresponding Relational Variables (FDCRVs) that are fragmented and distributed across different sites, which are focused on, are detailed below:

- a)* Formulating the detection problem in various distributed settings as optimisation problems, measured by either network traffic or response time;
- b)* It is beyond reach in practice to find optimal detection methods: the detection problem is NP-complete when the data is partitioned either horizontally or vertically and when one aims to minimise either data exchange or response time;
- c)* Data that is horizontally partitioned, provided several algorithms to find violations of a set of FDCRV, leveraging the structure of FDCRV to reduce data exchange or increase parallelism;
- d)* Verifying experimentally that the algorithms are scalable on large relations and complex FDCRV;
- e)* Data that is vertically partitioned provides a characterisation for FDCRV to be checked locally without requiring data exchange, in terms of dependency preservation. The intractable minimally refines a partition and makes it dependency preserving;

The methods and techniques for detecting violations of Conditional Functional Dependencies (CFDs) in relation to those are fragmented and distributed across diverse heterogeneous sources, as follows:

- 1)* Formulating the detection problem in various distributed settings as optimisation problems, measured by either network traffic or response time;
- 2)* When data is partitioned horizontally or vertically and the aim is to minimise either data shipment or response time. It is beyond reach in practice to find optimal detection methods;

- 3) When data is horizontally partitioned, to find violations of a set of CFDs, leveraging the structure of CFDs to reduce data shipment or increase parallelism need to provide several algorithms;
- 4) The algorithms should perform the scalability on complex CFDs and LSDB;
- 5) When data is vertically partitioned, the characterisation for CFDs needs to be checked locally without requiring data shipment, in terms of dependency preservation. It is intractable to minimally refine a partition and make it dependency preserving;

A central technical issue for DQ concerns is inconsistency detection, to identify errors in the data. More specifically, given a database  $D$  and a set  $\Sigma$  of dependencies serving as DQ rules, the detection problem is to find all the violations of  $\Sigma$  in  $D$ , for example, all the tuples in  $D$  that violate some rules in  $\Sigma$ . For a DQ tool to be effective in practice, it is a must to support automated and efficient inconsistency detection methods. When  $D$  is a centralised database, the detection problem is not very hard, but it makes it very difficult with heterogeneous data sources. In practice, however, a relation is often fragmented and distributed across different sites.

**Conditional functional dependencies:** A functional dependency is defined in a single relation. Consider a relation schema  $R$  defined over a set of attributes, denoted by  $attr(R)$ . For each attribute  $A \in attr(R)$ , its domain is denoted by  $dom(A)$ . For a tuple  $t$  of  $R$ , one uses  $t[A]$  to denote the value of the  $A$  attribute of  $t$  and for a list  $X$  of attributes in  $attr(R)$ , one uses  $t[X]$  to denote the projection of  $t$  onto  $X$ .

**Syntax:** A FD CRVs  $\varphi$  defined on  $R$  is a pair  $R(X \rightarrow Y, T_p)$ , where (Gartner 2007)  $X, Y$  are sets of attributes from  $attr(R)$ , (Fan *et al.* 2008)  $X \rightarrow Y$  is a standard FD to as the FD embedded in  $\varphi$  and (O'zsu *et al.* 1999)  $T_p$  is a tableau with attributes in  $X$  and  $Y$ , referred to as the pattern tableau of  $\varphi$ , where for each  $A$  in  $X \cup Y$  and each pattern tuple  $T_p \in T_p$ ,  $t_p[A]$  is either a constant 'a' in  $dom(A)$  or an unnamed (yet marked) variable '\_' that draws values from  $dom(A)$ . One writes  $\varphi$  as  $X \rightarrow Y, T_p$  when  $R$  is clear from the context. If  $A$  occurs in both  $X$  and  $Y$ , one uses  $t[A_L]$  and

$t[A_R]$  to indicate the occurrence of  $A$  in  $X$  and  $Y$ , respectively. One separates the  $X$  and  $Y$  attributes in a pattern tuple with ‘||’. For a pattern tuple  $t_p$ , one refers  $t_p[X]$  to as the LHS of  $t_p$ .

**Semantics:** One defines an operator  $\simeq$  on constants and ‘\_’:  $\eta_1 \simeq \eta_2$  if either  $\eta_1 = \eta_2$  or one of  $\eta_1, \eta_2$  is ‘\_’. The operator naturally extends to tuples, e.g., (Mayfield, EDI)  $\simeq$  (\_\_, EDI) but (Mayfield, EDI)  $\not\simeq$  (\_\_, NYC). An instance  $D$  of schema  $R$  satisfies the CFD  $\varphi$ , denoted by  $D \models \varphi$ , if for each tuple  $t_p$  in the pattern tableau  $T_p$  of  $\varphi$  and for each pair of tuples  $t_1, t_2 \in D$ , if  $t_1[X] = t_2[X] \simeq t_p[X]$ , then  $t_1[Y] = t_2[Y] \simeq t_p[Y]$ . Intuitively, each tuple  $t_p$  in the pattern tableau  $T_p$  of  $\varphi$  is a constraint defined on a subset  $D_{t_p}$  of tuples rather than on the entire  $D$ , where  $D_{t_p} = \{t | t \in D, t[X] \simeq t_p[X]\}$  such that for any  $t_1, t_2 \in D_{t_p}$ , if  $t_1[X] = t_2[X]$ , then (a)  $t_1[Y] = t_2[Y]$  and (b)  $t_1[Y] = t_p[Y]$ . Here (a) enforces the semantics of the FD embedded in  $\varphi$  and (b) assures that the constants in  $t_p[Y]$  match their counterparts in  $t_1[Y]$ . The semantic verification purpose, highlighted in section 8.4, 9.1 and 9.2.

**Detection of algorithms for a set of FDCRVs:** Two algorithms are presented for detecting violations of multiple FDCRVs. Both algorithms invoke algorithms for detecting violations of single FDCRVs and details are given below:

**a) The first algorithm:** SEQDETECT, follows a naive approach. It processes FDCRVs one by one, by sequentially executing an algorithm for detecting violations of single FDCRVs (either PATDETECTS or PATDETECTRT). The algorithm is based on pipelined processing: as soon as a site is done with processing the current FDCRVs (for example, partitioning tuples or detecting violations), it starts checking the violations for the next FDCRV, such that no site is idle before it processes all the FDCRVs. Algorithm SEQDETECT, however, may incur unnecessary network traffic: the same tuple may be shipped multiple times, once for each matching FDCRV.

**b) The second algorithm:** CLUSTDETECT, aims to reduce unnecessary data shipment by leveraging common attributes of the input CFDs. To do this,

CLUSTDETECT *merge* two CFDs  $\varphi = (X \rightarrow A, T_p)$  and  $\varphi' = (X' \rightarrow B, T'_p)$  into one if either  $X \subseteq X'$  or  $X' = X$ . More specifically, it first partitions D based on the (sorted) projected pattern tableau  $T_p[X \cap X'] \cup T'_p[X \cap X']$  if the overlap condition above holds. It then assigns a coordinator for each of the pattern tuples in this projected tableau as described in PATDETECTS and PATDETECTRT.

- c) The final algorithm:* At each site, the violations of the corresponding CFDs are checked locally by executing the violation detection queries for each CFD.

The novelty of one's work consists in as follows:

- a)* A formulation of FDCRV violation detection as optimisation problems to minimise data shipment or response time;
- b)* The NP-completeness of these optimisation problems when the data is partitioned either vertically or horizontally;
- c)* Algorithms to detect FDCRV violations in horizontally partitioned data, aiming to minimise either data shipment or response time;
- d)* A characterisation of locally checkable FDCRV for vertically partitioned data in terms of dependency preservation and the intractability of minimally refining a vertical partition to make it dependency preserving;

As verified by one's experimental results, the algorithms scale well with regard to the size of data, the number of fragments, and the complexity of FDCRV and hence provide effective methods for catching inconsistencies in distributed data.

### 6.3.4 Fuzzy multi-attribute theory

The multi-attribute scoring methods are widely used while comparing the alternatives because of their simplicity. In the case of incomplete information and vagueness, these multi-attribute scoring methods have been extended to obtain the fuzzy versions. The Multi-Attribute Decision-Making (MADM) problem is one of

the key sectors in modern decision science. The theory and method have been widely applied in the fields of medical science, engineering design, social life, investment decision-making and project evaluation. Regarding the multi-attribute decision-making problem both the attribute value and attribute weight of a scheme are exponential fuzzy numbers.

In recent years, research on fuzzy numbers has attracted attention from scholars and experts and has been widely used in the field of MADM problems (Sha *et al.* 2016). The MADM problem is of profound theoretical significance and has a wide practical application background in various industries. Therefore, research on MADM problems has always been a key subject for people. In real life, the uncertainty of decision information is regularly caused due to complex object environments and the fuzziness of human thinking. In solving such a class of problems, fuzzy numbers, such as interval numbers, triangular fuzzy numbers and trapezoidal fuzzy numbers, are usually adopted to express such uncertainty of decision-making information.

Lakshmana *et al.* (2011) proposed and discussed a new method for the sequencing of an interval-valued intuitionistic fuzzy set and stated this method by the way of calculation example analysis and made a comparison with other methods as well. Park *et al.* (2011) extended the TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) method to solve problems of Multi-Attribute Group Decision-Making (MAGDM) in interval-valued intuitionistic fuzzy circumstances, in which all preferential information provided by decision-makers would be indicated by an interval-valued intuitionistic fuzzy decision matrix. Xu (2007) proposed an ideal method to solve the problem with interval-valued intuitionistic fuzzy MADM with the attribute weight not completely known or completely unknown. Wang (2006) and Wei (2008) established an objective programming model based on the distance measure and difference maximum respectively and proposed a multiple attribute decision-making method where the attribute weight information is incomplete and the attribute value is an interval-valued intuitionistic fuzzy number.



**Problem description:** Regarding a fuzzy multi-attribute decision-making problem, suppose  $A = \{A_1, A_2, \dots, A_m\}$  is the scheme set,  $C = \{C_1, C_2, \dots, C_n\}$  is the attribute set and  $R = (\tilde{a}_{ij})_{m \times n}$  the fuzzy decision-making matrix, where  $\tilde{a}_{ij} = (c_{ij}, \sigma_{ij}, \tau_{ij})$  represents the exponential fuzzy number attribute value of the  $j^{\text{th}}$  attribute  $C_j$  of the  $i^{\text{th}}$  scheme,  $1 \leq i \leq m, 1 \leq j \leq n$ . In addition, the attribute the weight is still given in the form of an exponential fuzzy number (for example, the attribute weight vector  $W = (\tilde{\omega}_1, \tilde{\omega}_2, \dots, \tilde{\omega}_n)$ , where  $\tilde{\omega}_j = (C_{\omega j}, \sigma_{\omega j}, \tau_{\omega j})$  represents the weight of the  $j^{\text{th}}$  attribute  $C_j$ ). Try to sort and prioritise schemes according to the fuzzy decision-making matrix and attribute weight information (Mendel 2016).

### 6.3.5 Decision-making steps

Five steps follow by such a decision-making method which are listed below:

**Step 1:** Constructing a fuzzy decision-making matrix  $R$  according to the attribute value of each scheme.

**Step 2:** Standardised processing of the decision-making matrix. Since the attribute values of schemes in the decision-making matrix  $R$  have different units of measurement they are subjected to different measurement criteria. To achieve unified processing, the following equations can be employed for standardised processing. Therefore, as to obtain the standardised decision-making matrix  $R' = (\tilde{R}_{ij})_{m \times n}$ .

Where,  $\tilde{R}_{ij} = (\tilde{c}_{ij}, \tilde{\sigma}_{ij}, \tilde{\tau}_{ij})$ .

In the multi-attribute decision-making problem, the benefit-oriented attribute and cost-oriented attribute are two major attribute types.

For the benefit-oriented attribute, there is:

$$\tilde{c}_{ij} = \frac{c_{ij}}{\sum_{i=1}^m c_{ij}}, \tilde{\sigma}_{ij} = \frac{\max c_{ij}}{\sum_{i=1}^m c_{ij}} \sigma_{ij}, \tilde{\tau}_{ij} = \frac{\max c_{ij}}{\sum_{i=1}^m c_{ij}} \tau_{ij} \quad (6.1)$$

For the cost-oriented attribute, there is:

$$\tilde{c}_{ij} = \frac{\frac{1}{c_{ij}}}{\sum_{i=1}^m \frac{1}{c_{ij}}}, \tilde{\sigma}_{ij} = \frac{\min \frac{1}{c_{ij}}}{\sum_{i=1}^m \frac{1}{c_{ij}}} \sigma_{ij}, \tilde{\tau}_{ij} = \frac{\min \frac{1}{c_{ij}}}{\sum_{i=1}^m \frac{1}{c_{ij}}} \tau_{ij} \quad (6.2)$$

**Step 3:** Determination of positive ideal scheme  $A^+$  and negative ideal scheme  $A^-$  of the fuzzy MADM problem.

Positive ideal scheme:

$$A^+ = [A^+_1, A^+_2, \dots, A^+_n], A^+_j = (C^+_j, \sigma^+_j, \tau^+_j) = (\min_i \tilde{c}_{ij}, \min_i \tilde{\sigma}_{ij}, \min_i \tilde{\tau}_{ij}) \quad (6.3)$$

Negative ideal scheme:

$$A^- = [A^-_1, A^-_2, \dots, A^-_n], A^-_j = [c^-_j, \sigma^-_j, \tau^-_j] = (\min_i \tilde{c}_{ij}, \min_i \tilde{\sigma}_{ij}, \min_i \tilde{\tau}_{ij}) \quad (6.4)$$

Where  $j = 1, 2, \dots, n$

**Step 4:** Determination of the weights of the attributes. The weight information of the attribute is given in the form of an exponential fuzzy number (for example, the attribute weight vector).

$$W = (\tilde{w}_1, \tilde{w}_2, \dots, \tilde{w}_n) = \{(c_{w1}, \sigma_{w1}, \tau_{w1}), (c_{w2}, \sigma_{w2}, \tau_{w2}), \dots, (c_{wn}, \sigma_{wn}, \tau_{wn})\} \quad (6.5)$$

Calculate the score value of each attribute using, suppose the exponential fuzzy number

$\tilde{a} = (c, \sigma, \tau)$ . The score function is:

$$p(\tilde{w}_j) = \frac{\phi E(\tilde{w}_j)}{(1-\phi)Var(\tilde{w}_j)} \quad (6.6)$$

Where  $\tilde{w}_j$  represents the weight of the  $j^{th}$  attribute  $C_j$  and  $\phi$  represents the decision-makers attitude preference on the expected value of the decision-making information and variance. Therefore, the accurate weight of each attribute can be obtained as:

$$W_j = \frac{P(\tilde{w}_j)}{\sum_{j=1}^n P(\tilde{w}_j)} \quad (6.7)$$

**Step 5:** Calculating the weighted distance between each scheme and the positive/negative ideal scheme. The weighted distance between scheme  $A_i$  and the positive ideal scheme:

$$D_i^+(A_i, A^+) = \sum_{j=1}^n w_j D(A_{ij}, A_i^+) \quad (6.8)$$

The weighted distance between scheme  $A_i$  and the negative ideal scheme:

$$D_i^-(A_i, A^-) = \sum_{j=1}^n w_j D(A_{ij}, A_i^-) \quad (6.9)$$

**Step 6:** Calculating the relative closeness  $\varepsilon_i$  of each scheme and sort schemes according to the value of relative closeness. A larger  $\varepsilon_i$  the value represents a more optimal scheme.

$$\varepsilon_i = \frac{D_i^-(A_i, A^-)}{D_i^+(A_i, A^+) + D_i^-(A_i, A^-)} \quad (6.10)$$

According to the definitions of expectation and variance in probability theory and by comprehensively considering the attitude preference of the decision-maker, this method can realise the accurate treatment of attribute weight. Subsequently, based on the distance measure between exponential fuzzy numbers, the distance between each scheme and the positive/negative ideal schemes is calculated, therefore, as to obtain the relative closeness of each scheme.

Finally, the feasibility and effectiveness of the proposed method were verified through case analysis. The proposed decision-making method is superior for clear logic, a simple decision-making process and ease of being understood. In addition, such a method has excellent application value and practical decision-making value, providing a scientific and practical decision-making reference for solving fuzzy MADM problems.

### 6.3.6 The similarity measurement methods to detect and reduced data redundancy

The importance of EHRs data integration is to find the correlation among entities manifested in diverse EHRs. Figure 6.2 (Saïod *et al.* 2017) shows the workflow and architecture of the similarity detection service, as follows:

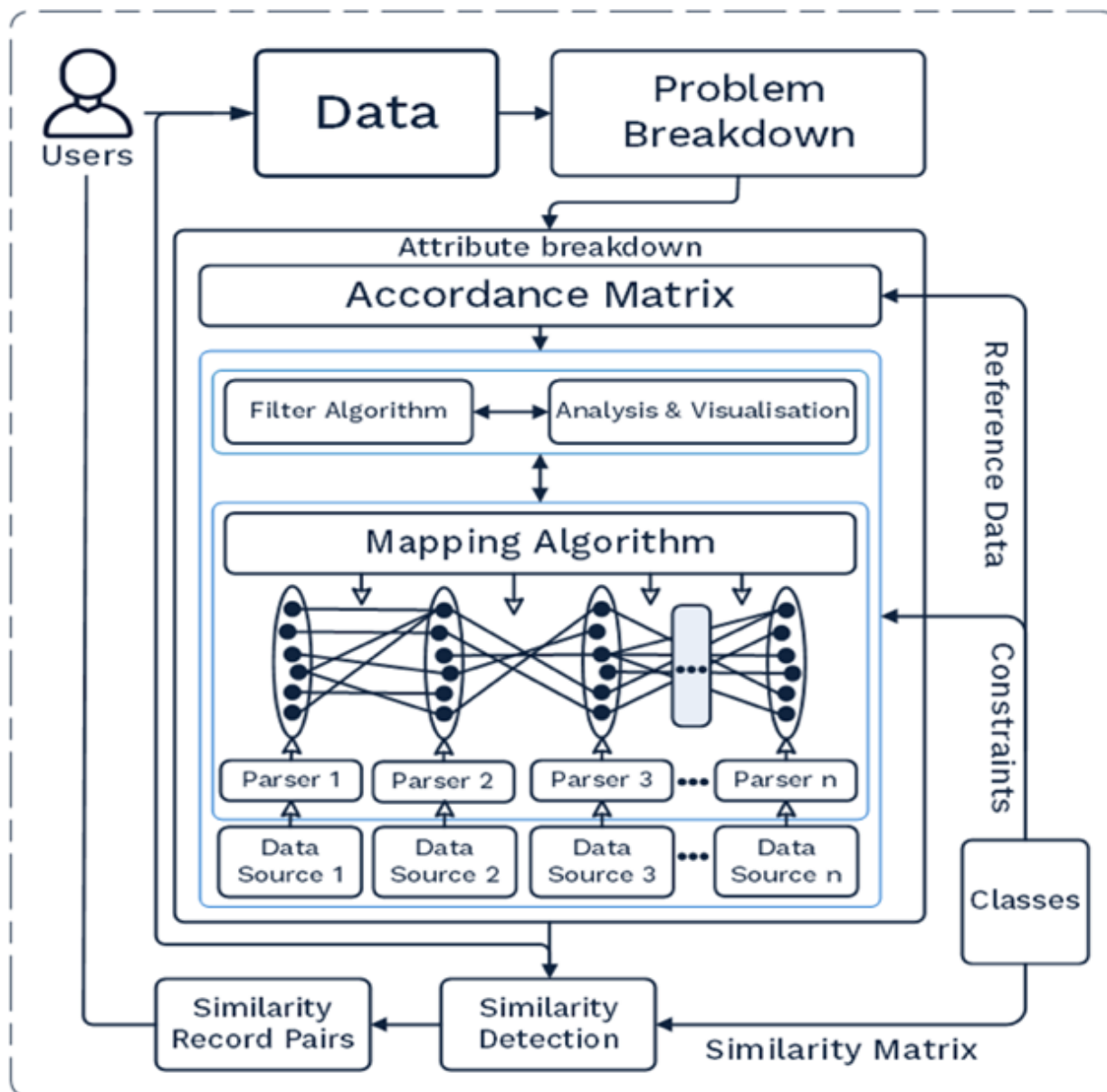


Figure 6.2: Workflow and architecture of the similarity detection service (Saïod *et al.* 2017)

The similarity measure technology may discover those conflicting entities are approximately identical among EHRs entities. With the tremendous growth in the adoption of EHRs, heterogeneous sources of patient health information are becoming available.

It is practically challenging to discover significant similarity entity and the way to measure and leverage providers' inputs. It is a very important aspect to identify the accurate subsidiary uses of EHRs data to achieve the goal (Jimeng *et al.* 2016). The objective of similarity becomes even more important when measuring the similarity among equivalent patient entity based on their EHRs data. Three effective similarity measurement types are appropriate in many applications, such as:

- a)* Case-based retrieval of similar;
- b)* Treating similarity between the batch similarity;
- c)* Cohort comparison and comparative effectiveness;

According to the aforementioned EHRs integration, the similarity measurement technology is assorted in four different groups:

***Instance-based similarity:*** The similarity between concepts is determined by common instances, as well as comparing new problem instances with instances and stores in memory, instead of performing explicit generalisation. Instance-based ontology mapping is a promising solution to a class of ontology alignment problems. The similarities between concepts are defined as ordinary instance and matching new entity issues and storing them, but not executing the exact generation. The common entity is the key value of similarity among the two concepts.

The promising solution is instance-based ontology mapping for a class of ontology classification. Measuring among similarity and annotated entity sets crucially depend on this. A set of abstractions evolved from significant entities does not maintain by Instance-based algorithms. If it has a large storage capability, then this approach reaches the nearest neighbour algorithm. The classification accuracy

significantly reduced the large storage requirements, but its performance degrades rapidly.

The instance-based similarity equation is:

$$\text{Similarity}(x, y) = -\sqrt{\sum_{a=1, n} (x_a - y_a)^2} \quad (6.11)$$

Where,  $x$  and  $y$  are instances in an  $n$  – dimensional instance space (David 1992).

**a) Lexical-based similarity:** The similarity between two concepts is based on the analysis of the linguistic interpretation of associated names. For example, let us find the most similar words for word  $W$ , to combine an estimate. Weight the evidence provided by word  $W'$  by a function of its similarity to  $W$ . Combining information among similar words is a similarity measuring function between words. Which word pairs require a similarity-based measurement is determined by a scheme method. If word  $W'_1$  is “alike” to word  $W_1$ , then  $W'_1$  can participating entity about the probability of invisible word pairs involving  $W_1$  (Ido *et al.* 1997). The Lexical-Based Similarity methods for language modelling of combining evidence evaluated as:

$$P_{sim}(W_2|W_1) = \sum_{w'_1 \in S(w_1)} \frac{W(w_1, w'_1)}{N(w_1)} P(w_2|w'_1) \quad (6.12)$$

$$N(w_1) = \sum_{w'_1 \in S(w_1)} W(w_1, w'_1) \quad (6.13)$$

$S(W_1)$  – The set of words most similar to  $W_1$ ;

$W(W_1, W'_1)$  – similar function;

**b) Scheme-based similarity:** The similarity among amalgamated characteristics is the analysis of similarity among two intentions. Two types of structure-based similarity are:

- (i) The internal structure-based similarity;
- (ii) The external structure-based similarity;

It can be the alteration of  $P_{sim}(W_2|W_1)$  in the back of equation 6.13, such as interpolating with the unigram probability  $P(w_2)$ :

$$P_r(w_2|w_1) = \gamma P(w_2) + (1 - \gamma)P_{SIM}(w_2|w_1) \quad (6.14)$$

The linear combination illustrates in yielding between the similarity estimate and the back-off estimate:

*if  $\gamma = 1$ , Then possibly to make  $\gamma$  depended on  $w_1$ .*

Therefore, the similarity allotment for achievement could vary between words (Karov *et al.* 1996).

*c) Taxonomy-based similarity:* The structural relationship breakdown is the base of similarity among two concepts in taxonomy-based similarity. It considers the relations as links connecting concepts. If two concepts are already matched, their neighbours (concepts are collected along with the links from the already matched concepts) may also be somehow similar.

*Let us consider two generic products  $W_1$  and  $W_2$ , and being represented by collections of terms  $W_1 = T_{11}, \dots, T_{1i}, \dots, T_{1n}$  and  $W_2 = T_{21}, \dots, T_{2i}, \dots, T_{2n}$ .*

Based on the two sets, the goal is to define a natural similarity between

*The main goal is to determine a natural similarity, based on two sets among  $W_1$  and  $W_2$ , denoted as  $S(W_1, W_2)$  (Raychaduri *et al.* 2003). The considered two principal approaches are:*

*(i) First approach:*

*The similarity is computed pair wise, say  $S_{ij}(T_{1i}, T_{2j})$  and then*

*the aggregation is performed using, for example, the average as:*

$$S_a(W_1, W_2) = \frac{\sum_{i=1}^n \sum_{j=1}^m S_{ij}}{mn} \quad (6.15)$$

*It is an interesting factor, when the objects  $T_{1i}, T_{2j}$ , belong to a given ontology.*

Here, the pairwise similarity can be determined as in (Lord *et al.* 2003) using the shortest paths and information theoretic constructs. The problem arises, only if the average is used with this approach.

*Even when the two sets are very similar,  $S_a(W_1, W_2)$  may not be 1.*

*When,  $W_1$  and  $W_2$  have only one common entity, then the similarity is 1 and it will ignore the other. Then the real trouble is to choose the maximum.*

### 6.3.7 The hybrid data integration method

The Hybrid Data Integration Method (HDIM) is based on combining multiple matrix factorisation methods, that can be used for in and out-of-matrix prediction of inconsistent data. The HDIM is very general and can be used to integrate many datasets across diverse heterogeneous sources, different entity types, including repeated experiments, similarity matrices and very sparse datasets, extensively comparing it to state-of-the-art machine learning and matrix factorisation models. The world of information processing must always be split into two parts as below:

- 1) Operations;
- 2) Information;

Fuzzy multi-attribute can obtain a higher average level of correct data of inconsistency solution and has an ideal performance compared to other methods (Semih *et al.* 2010; Yusuf *et al.* 2012; Evangelos *et al.* 2013). However, a method for reducing data inconsistency has to be combined with a method for data integration to coherently solve data inconsistency and data integration problems simultaneously. Ontology-based data integration involves the use of domain ontologies to effectively combine data and information from multiple heterogeneous sources (Longbing 2010; Dnyanesh *et al.* 2011; Matthias 2012). But the existing ontology integration methods are not sufficient to implement Fuzzy-Ontology (Nguyen *et al.* 2010; Duong *et al.* 2011; Hai *et al.* 2013).



The necessity arises, therefore, to develop a new method for DI using the efficient hybrid method of fuzzy-ontology, which motivates the current study. As a result, this study has focused on a novel approach based on fuzzy-ontology to tackle the DQ issues in EHRs for LSDB. The systems have applied to not only performing the function of receiving and displaying information but also to automatically and accurately extract information from heterogeneous data sources. The proposed hybrid method has equivalently matched two concepts across different data sources and automatically resolved any inconsistency arising from multiple data entities.

The important expected contribution of this study had realised the necessity of a hybrid method to improve on DI from heterogeneous and inconsistent data sources. The key outcome of this research has discovered a new merged concept by finding consensus among conflicting data entries. A series of experiments have performed therein to show that the proposed hybrid method is effective and accurate for data integration, which reduced the DQ issues in EHRs for LSDB.

## **6.4 Summary of the key lessons from the review of approaches and recommendations for EHRs integration methods**

The main objective of reviewing the approaches and recommendations for EHRs integration method was to determine, from publicly available publications, the way in which researchers and HCOs have tackled the DQ issues. This is despite the limited number of published methods and approaches to the DQ issues in EHRs. The publications reviewed provided insights into the factors that should be considered when embarking on an arguably complex process of determining an appropriate set of methods to be integrated by the HCOs. This section provides a summary of the lessons learned from the reviewed approaches and recommendations.

The process of determining the DQ issues should also involve consultation with the key stakeholders, who would be affected by the introduction of DQ standards.

Involving key stockholders, such as healthcare professionals, different interest groups, policymakers and suppliers of EHRs, facilitates active participation and reduces the risk of resistance to change that could accompany the introduction of standards. A mathematical modelling methodology based on hybrid fuzzy-ontology, which was initially explored in for DI (Hai *et al.* 2013; Jonas *et al.* 2013; Maio *et al.* 2014; Abdullah *et al.* 2015; Uthayan *et al.* 2015) is nominated.

Due to the diverse interests of many of the technical experts in the working group of HCOs and the consensus-based process of international standards development, ensuring standards frequently does not meet the specific requirements of some form of localisation to resolve the DQ issues to the LSDB. Such localisation would require an in-depth understanding of the specific standards. A lack of necessary skills could make the localisation of eHealth service much more difficult. The review of the process used in different countries to determine the key issues to be considered to resolve the issues, showed that different approaches were taken.

## 6.5 Conclusion

This chapter provided an overview of the approaches used and different methods that have tackled different DQ issues for a different incident in the LSDB at the national and global level to determine the lessons that could be learned from the processes they followed. The review found that different methods and approaches were taken by different organisations reviewed to determine and manage the DQ issues.

Given the diversity of publicly available methods and approaches to tackle the DQ issues in EHRs, a need was identified for a generic hybrid method that can solve every EHRs DQ integration problem to overcome these barriers and challenges. These include the provision of EHRs as it pertains to DQ, which will combine features to search, extract, filter, clean and integrate data, to ensure that users can

coherently create new consistent data sets. Figure 6.3 demonstrates the combined outcome of Chapter Three, Chapter Four, Chapter Five and Chapter Six, as follows:

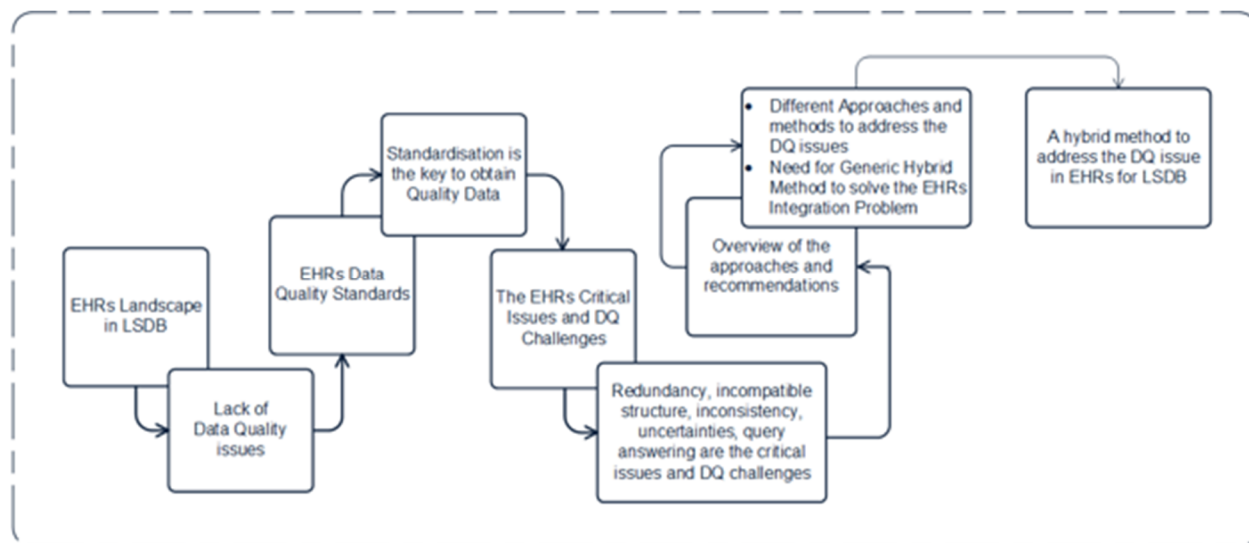
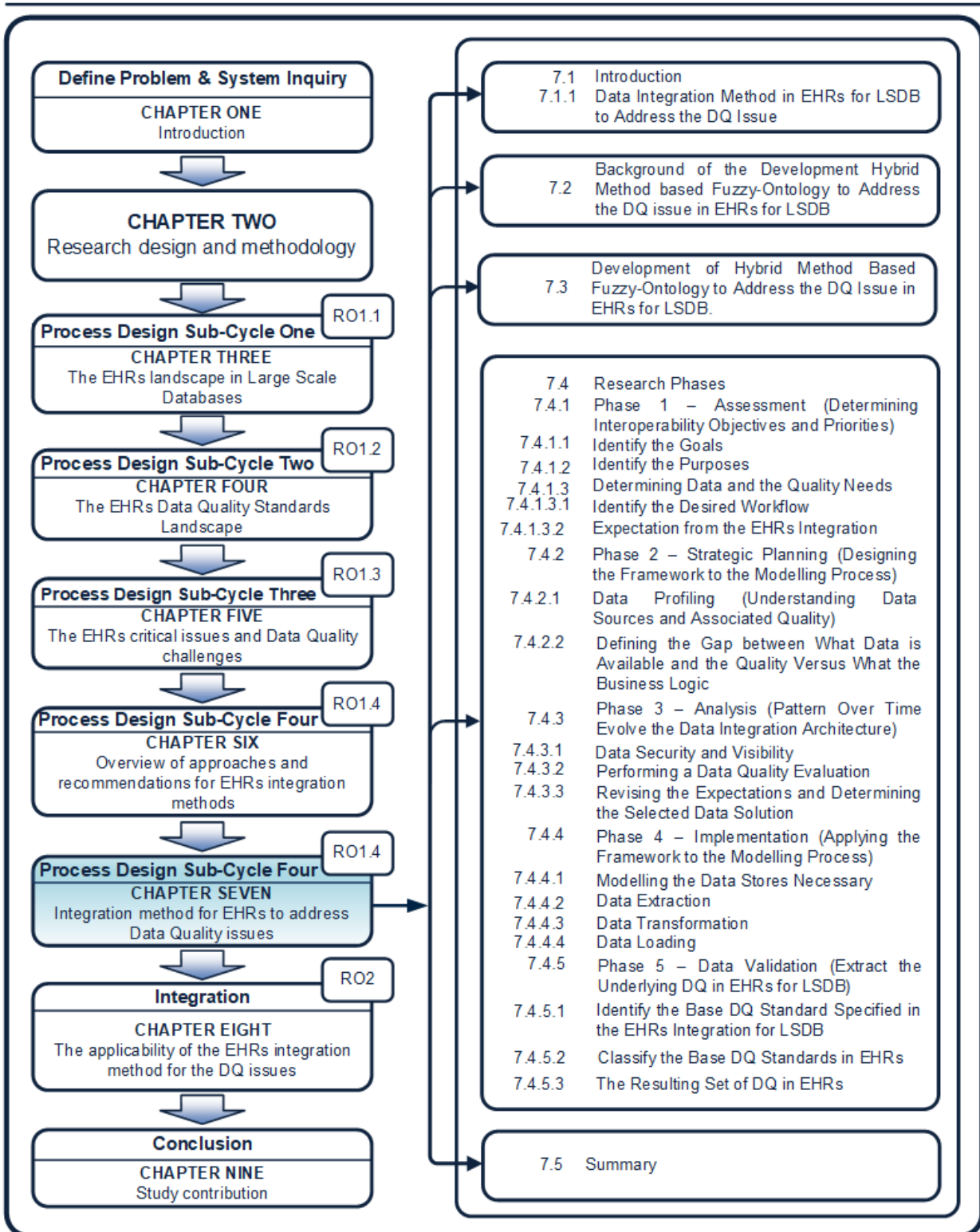


Figure 6.3: The combined outcome of Chapter Three, Chapter Four, Chapter Five and Chapter Six

As illustrated in figure 6.3, this thesis proposes a generic hybrid method based on Fuzzy-Ontology, a mathematical simulation to measure the probability risk and similarity measurement for EHRs integration to address the DQ issue. The development of Fuzzy-Ontology is discussed in Chapter Seven.

## CHAPTER SEVEN: Integration method for EHRs to address Data Quality issues



Outline of the Chapter Seven

## CHAPTER SEVEN

### 7.1 Introduction

The purpose of this chapter is to deal with the development phases for EHRs integration methods to address DQ issues in EHRs for the LSDB. It also maps to Sub-Cycle Four of the DSR process, described in section 2.3.2.2.4 and highlighted in figure 7.1.

Background of the development of HM is provided in section 7.2. Section 7.3 describes the discussion of the research phases. The motivation for the inclusion of each phase in the Hybrid Method (HM) based on Fuzzy-Ontology is provided in the detailed description of each phase, as presented in section 7.4, which provides details for research phases including assessments, goals, purposes, needs, workflows, expectations and planning. Section 7.5 summarises the chapter. The actual development, applicability and research result analysis of the EHRs integration method to deal with DQ issues is provided in Chapter Eight.

#### 7.1.1 Overview of EHRs Integration Methods for LSDB

Data integration (DI) is defined as the coalescence of technology and procedure to assemble data from diverse heterogeneous sources into significant and meaningful datasets. To transform diverse heterogeneous data into a unified single dataset of truth, the DI system integrates diverse trusted data from heterogeneous sources obeying the business requirements in the entire integration procedure.

Three primary components of DQ, are as follows:

- 1) Data profiling;
- 2) Data correction;
- 3) Data monitoring;

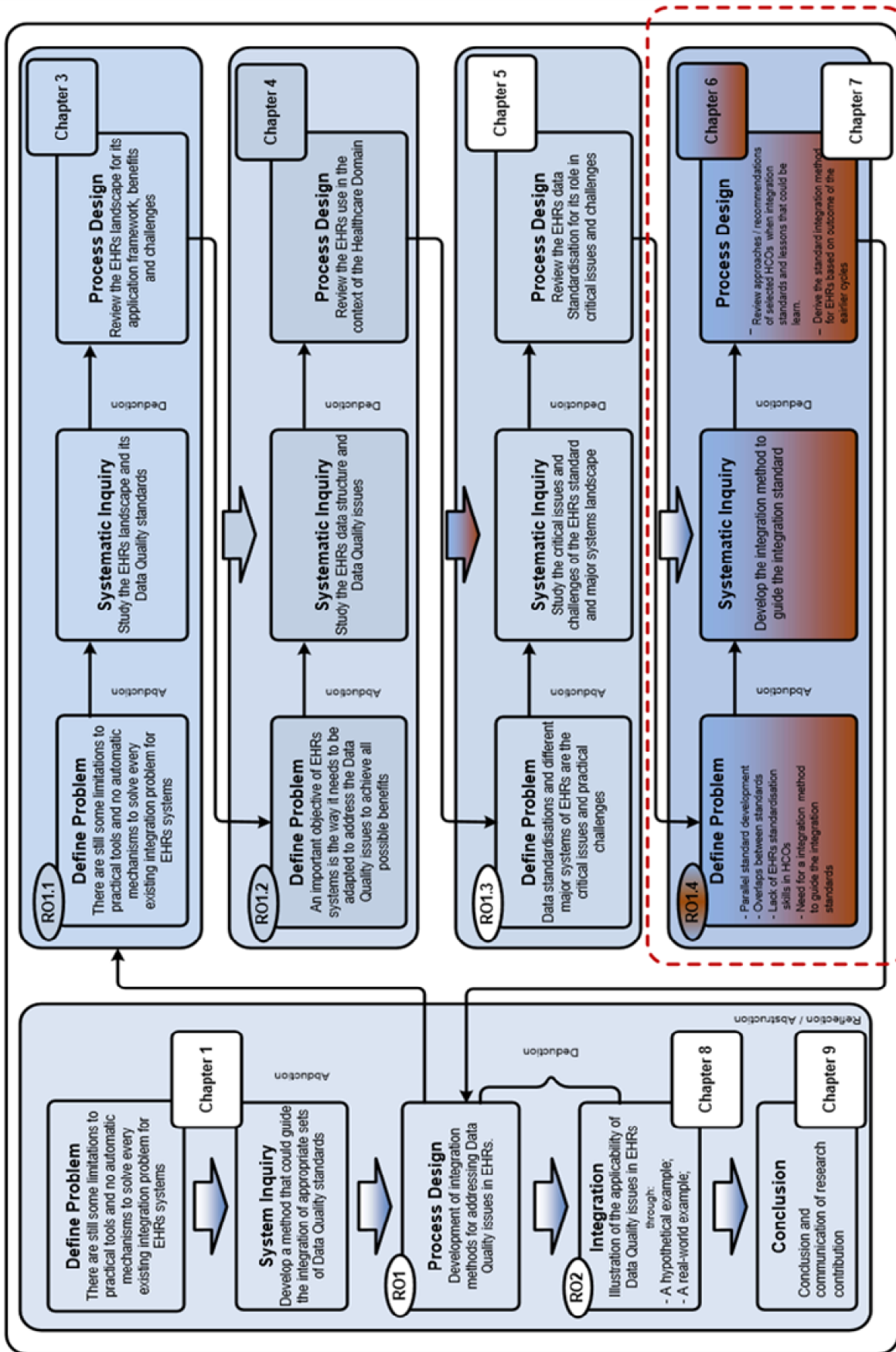


Figure 7.1: The position of Chapter 7 in the design science research process used in this study

In addition, the DI system in HCOs is the data retrieving procedure from diverse heterogeneous sources in a spontaneous way to yield a comprehensive, accurate, current and consistent health data set for reporting, decision-making and analysis. The objective of DI becomes even more important in the case of merging systems of different similar organisations (Saïod *et al.* 2017).

## 7.2 Background of the hybrid method based Fuzzy-Ontology for EHRs integration

The development of the purpose Hybrid Method (HM) based Fuzzy-Ontology to address the DQ issue in EHRs for the LSDB, is discussed in Chapter Six. One of the key lessons from Chapter Six is that the EHRs initiatives, especially at the integration level, should consider the HCOs health system needs and challenges. It should also consider the data inconsistency challenge that specific EHRs initiatives will play in addressing the DQ issues in the LSDB, the HCOs interoperability requirements and its health priorities.

As discussed in section 6.4 the mathematical model based on Fuzzy-Ontology has been nominated to address the DQ issue in EHRs for the LSDB, as this provides a strong theoretical and practical framework to work with heterogeneous, complex, conflicting and automatic consensus methods for DI. Representing fuzzy set concepts in ontologies are a feasible approach to express imprecise concepts and relationships (Zadeh 1965; Fernando *et al.* 2013; Pérez *et al.* 2013). In the semantic web, ontology is commonly used as an efficient conceptualisation solution.

Fuzzy logic can easily incorporate into ontology to integrate data efficiently and to reduce data inconsistency. Fuzzy logic is defined as a form of probabilistic logic, which deals with vague data, probably preferable to irreversible and precise values. Fuzzy logic uses the interval value range for the variable between 0 (zero or false) and 1 (one or true), where traditional logic uses an exact binary variable (where variables may take on true or false values). Fuzzy logic even extends to deal with

the variable of partial truth concept, where the variable range spreads between completely true and false.

The characteristics of ontology are to provide conditional and precision of a shared conceptualisation (Haibo *et al.* 2017). This is one of the most promising development technologies to represent data. The major advantages of ontology are:

- a)* Formal;
- b)* Machine-readable;
- c)* Sharable;

Ontology has the undeniable success in perfect, classical and crisp ontology, but therefore fails to deal with imprecise or vague meaning data. Imprecise or vague meaning data are, however, so common when integrating data from diverse heterogeneous sources and it has become an essential part to deal with vagueness in the knowledge representation field in the integration process. According to Zadeh (1965), “Since fuzzy set theory and fuzzy logic seem appropriate to manage the vagueness which is inherent to real-world information, fuzzy ontology, which introduces those two techniques into crisp ontology”, emerged in the early 2000s (Cross 2014). This concept concludes that fuzzy logic can be introduced with ontology as a hybrid method to deal with tradition and imprecise or vague meaning data with a world belief or truth degree in the integration process.

### **7.3 Development phases of the hybrid method based on Fuzzy-Ontology for EHRs integration**

This chapter deals with the application of the hybrid method based on fuzzy-ontology intelligent logic in diagnosing cruelty identical level and counselling apposite therapies for patients having hypertension.

This would be followed by the development of a series of business use cases using the hybrid method that describes healthcare processes requiring information



sharing. The hybrid fuzzy-ontology inherent in the business processes should then be identified so that the appropriate DQ that can support information exchange can be determined. Figure 7.2 describes the phases of the proposed HM based on Fuzzy-Ontology, as follows:

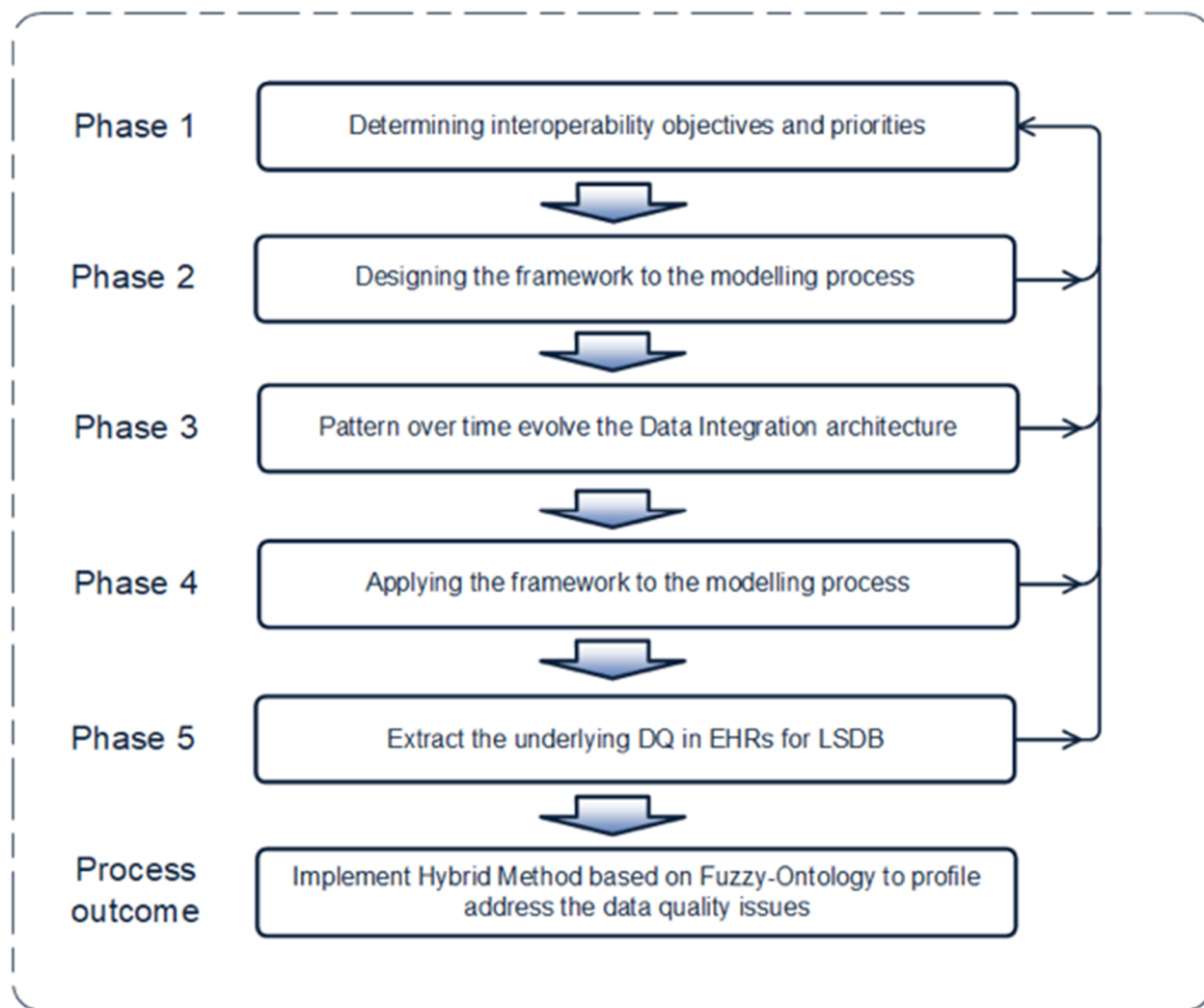


Figure 7.2: The phases of the proposed Hybrid Method based Fuzzy-Ontology

The HM based on fuzzy-ontology consists of five phases, as illustrated in figure 7.2, described below:

1. *Phase One:* Determining interoperability objectives and priorities;
2. *Phase Two:* Designing the framework for the modelling process;
3. *Phase Three:* Pattern over time evolving the data integration architecture;

4. *Phase Four*: Applying the framework to the modelling process;
5. *Phase Five*: Extracting the underlying DQ in EHRs for the LSDB;

The phases in HM based on fuzzy-ontology are implicitly iterative according to the dynamic character of the healthcare domain and the need to accommodate changes in health DQ priorities and their associated interoperability objectives. The right pointing arrows coming out of each phase to revise their outcome. The result of the entire HM based on Fuzzy-Ontology would be an amalgamated list of functions mapped to the applicable outlines and their foundation.

## 7.4 Research phases

The proposed HM based on Fuzzy-Ontology's five phases are described below:

### 7.4.1 Phase One – Assessment (determining interoperability objectives and priorities)

Drawing the different approaches and methods to address DQ issue in EHRs in the LSDB discussed in Chapter Six, the HM based on fuzzy-ontology begins with the identification of the interoperability objectives that should be addressed by DQ standards. Interoperability goals and/or objectives are aimed at articulating the desired state of healthcare systems that facilitate access to healthcare information by only authorised users when required, to improve the DQ as well as the quality healthcare service. This is to ensure that funding and investment in eHealth initiatives are aligned with state priorities.

Ideally, eHealth initiatives and DQ standards should be led by an independent stockholder governing body, vested with the necessary power and authority to set up the interoperability, priorities, goals and objectives, as discussed in section 6.2, namely the EHRs data integration method. The successful implementation of eHealth initiatives and the adaptation of e-health standards that can support

interoperability, also require collaboration with different types of stockholders to understand their interests. All these successes and objectives discussed above depend on the DQ.

#### 7.4.1.1 Identify the goals

The EHR systems' implementation is extremely critical and difficult to identify whether the implementation was successful. The commensurable goal and objective of this section are to establish this reality. The EHR implementation step should help the HCOs to identify their present situation to ascertain where they need to be improved. Some of the questions needed to identify during this phase are as follows:

- a)* Achievements versus where the needs are to be improved?
- b)* Are the HCOs providing the best possible quality healthcare service or simply trying to make it?
- c)* What would be done differently to improve the quality of healthcare service?
- d)* Which area needs more attention that performs far from the ideal HCO goals?

At this phase, the HCOs team leader should identify the business goal, exigent, economic and technological preparative as they require. The inclusion of relevant stakeholders in eHealth standards adaptation process provides valuable inputs, improves communication and ensures continued support for standardisation initiatives. This is important as a wide range of stakeholders may be affected by the HM based on fuzzy-ontology adaptation process. It is important to identify the stakeholder's groups that should be involved and develop a practical HM based on fuzzy-ontology to address the DQ as well as the various stakeholder groups that can potentially be involved in DQ in the EHRs standard. The HCOs can find the goals and objectives upon their different diminution aspects. Goals for DQ in EHRs improvement activities include, as below:

- Accurate data anticipation from the HCOs performance;

- Accurate data allowing effective aggregation and interexchange across HER systems for quality healthcare services and payment initiatives;

All goals need to be listed and implemented through the entire EHRs integration process. These guidelines will be needed when the EHR systems are re-appreciated for the efficient integration for the HCOs and all stakeholders.

#### 7.4.1.2 Identify the purposes

To determine the requirements to address the DQ issues using HM based on fuzzy-ontology the focus should be on identifying the data inconsistency that could be useful to quality care of the health service. DQ is one of the most essential requirements as healthcare professionals (nurses, doctors, pharmacists, and stakeholders) are using EHRs for the decision-making proposes. Poor data could have a serious effect on the healthcare service and even cause the death of the patient. EHRs should be exchanged to support the continuation of quality care and at the time that the information would be required. This chapter focuses on three main identifications, as follows:

- Identifying the most meaningful associations among heterogeneous data sources that could be explored to improve DQ in EHRs;
- Identifying the integrity constraints to be specified in the global schema of data mapping that can be explored to improve DQ in EHRs;
- Identifying uncertainties in the data integration that when minimised, would result in an improved DQ in EHRs;

The diverse heterogeneous data can often mismatch and cause inconsistency while integrating the standards and interoperability requirements for their implementation. Therefore, the DQ standard prioritisations are also important, to identify the most critical area that should be requiring minimal efforts to detect and address the DQ issues in EHRs for the LSDB. Examples of the DI objectives could be, as follows:

1) *Healthcare objectives:* Traditionally HCO professionals require EMRs that have the following functionality:

- To provide healthcare provision supporting the provider to present individual patients demographic and historical data to make an accurate diagnosis;
- Share the EMRs across HCOs so patient data is available to other providers caring for the same patient;
- Providing EMRs for introduction into other documents (for example, diagnosis result requests, references and clinical reports);
- Integrating EMRs received from other HCOs (for example, diagnosis result and hospital updates);
- To provide data exchange across HCOs where the patient thereafter registers;
- To provide EMRs to patients about their treatment;
- To monitor the progress of patient health promotion initiatives;
- To represent EMRs for medical audit purposes initiatives;

2) *Non-healthcare objectives:* HCOs also need a patient's EMRs that can be used to meet directorial and covenant constraint by, as follows:

- To provide clinical legal testimony (*for example, for medical aid claims*);
- To provide legal testimony when a patient claims against a third party (*for example, for injuries, occupational diseases, claims and in respect of product liability*);
- To provide reports and EMRs for third parties (*for example, medical aid and insurance companies*);
- To provide EMRs regards to claims for additional benefits and other social support;
- To meet the requirements of specific legislation on subject access to personal demographic and EMRs;
- To provide testimony of workload within an HCO;

3) *Complementary objectives:* Most of the HCOs would prefer to have EHR systems that can interact, as follows:

- To use for decision-making;
- To provide medical training for distance and non-distance purposes;
- To provide health administrative purposes;
- To provide data security and access control systems for patient confidentiality;
- To provide healthcare quality appreciation and revalidation;
- To provide epidemiological inspection and clinical research;

4) *Appearance objectives:* Share and integrate EHRs across other EHR systems, as follows:

- To provide read-only access according to the health act;
- To provide read-only access to an external provider when caring for the same patient;
- To provide integration access;
- To share EMRs to other providers or HCOs;
- To provide an application interface to medical apparatus and distance healthcare performances;
- To support a shared record dependent on messaging, such as with pathology or some electronic prescribing systems;
- To support an interface with medical devices, supporting TeleHealth and TeleCare activities;

The purposes of the EHRs are multiple and varied, depending on the context of use. New purposes are identified and required through changes in models of healthcare provision and these influence the underlying architecture of the EHRs, as well as the records systems that employ them. Although healthcare had many evaluations, the individual patient care has remained the same since the earliest general practice perspective.

Any new purposes must recognise the EHRs primary purpose and its use and either improve this or not detract from it.

### 7.4.1.3 Determining data and the quality needs

High-quality EHR systems support the HCO's performance and provide significant centralised patient care with the lower cost. The key driver for quality care requires high-quality data that provide accurate and efficient healthcare to patients. The high-quality data is defined as data stored and managed in such a way to represent accurate, feasible and usable as per the needs. According to other research, the DQ issues in EHRs have expressed immensely disparate results and the accurateness of the health data, range between 44% to 100% and the completeness from 1% to 100%.

The DQ issues often arise as many HCOs skip the attention on DQ measurement, whether the EHRs are correct, feasible and usable. Therefore, these EHRs are not providing enough transformation initiatives. The necessity of the DQ improvement realisation happens only after providing the audit for EHR systems achievement. The DQ issue often limits the providers' performance and flabbergasted the HCOs to conclude their attention for DQ improvement. Several specific terms need to be considered before the DQ improvement initiatives in EHRs for LSDB take place and are as follows:

- a) Measuring the current EHR systems performances over the society and individual HCO;
- b) Measuring the available resources to perform DQ improvement;
- c) Measuring the healthcare professional's obligation, quality, intellect, proper training;

#### *Determination and identification strategy scheme:*

- Determining the key of HCO's success and providing attention to the individual health team leader;
- The synchronisation between HCO and community;
- Providing details to all stakeholders for the current and ongoing process at the starting point;
- Identifying all effective impression and attention needed to improve the DQ;

The need to measure all DQ dimensions to identify a proper assessment strategy entails five major dimensions of DQ, as follows:

- 1) *Completeness*: Defined as the degree of the trust level of a patient's EMRs;
- 2) *Correctness*: Defined as the component degree of a patient's EMRs;
- 3) *Concordance*: Defined as an agreement between components of the patients' EMRs or between other HER systems;
- 4) *Currency*: Defined as accurate EMRs representation of a patient's present state;
- 5) *Plausibility*: Defined as EMRs that make clear sense;

#### 7.4.1.3.1 Identify the desired workflow

Measurement is the compulsory process for DQ improvement. Two different value-based compensation measurements are:

- 1) Compensation based quality;
- 2) Cost based quality;

Several factors such as data capture standards can limit the DQ measurement during the significant use of EHR systems. Different EHRs contain format, types, structure and integration procedures, which may not comply with the data standards. Different providers present information differently and captured information may not be recognised by the EHR systems. The EHRs need to be validated before capturing into the EHR systems to provide DQ measurement in a different dimension. The study has found a few incidents that need to be considered to improve the DQ in the workflow procedure, as follows:

- *Incident One*: The EHR systems are not centralised and capture and store data in different locations. Often one section does not have access to another section. The quality reports pull data from one single source only and probably miss the significant data that may present in the other source.



- *Incident Two:* The clinical staff does not have proper training and/or the EHR systems do not specify clearly to capture data in a correct location and information places are mixed.
- *Incident Three:* EHR systems do not validate data while capturing. For example, the hypertension measurement value or BMI calculation value is supposed to be captured in a numeric value, rather than a textual representation or it is captured elsewhere in notes.

*Sustainability and resources:* HCOs or the community may not have enough resources to provide support with the DQ improvement or redesign of the entire workflow. Therefore, this needs to be identified and placed at all necessary resources. It may depend on every individual stakeholder and they must evaluate themselves to commit in terms of the DQ improvement initiative. HCOs will have a complementary focus beyond the individual stakeholder, such as communicating with other HCOs and documenting different standards to implement in the aggregation of high DQ in the integration process. The DQ improvement initiative procedure is divided into six groups, as follows:

- 1) *Administrative leadership and staff:* To provide management workflow;
- 2) *Clinical leadership:* To provide proper training for the clinical staff;
- 3) *The EHR vendors:* To provide EHR systems structure;
- 4) *IT professionals:* To develop, maintain and support the infrastructure and EHR systems;
- 5) *Patients and patient families:* The privacy and security of patient EMRs;
- 6) *Payers:* To provide communication between the EHR systems and payer IT systems;

#### **7.4.1.3.2 Expectation from the EHRs integration**

The potential of EHRs is widely recognised and only the DQ can guarantee the actual picture of the patient health condition by EHRs integration from diverse HCOs. The patient EMRs would be available by access control systems to

authorised providers for reducing the extra diagnosis test and avoiding allergenic medication. The Health Service Executive of Ireland (HSE) has recognised central EHR systems for healthcare services. EHR systems are rapidly growing over the HCOs in respect of usability. The centralised EHRs will reduce the time limited performance as the improvement of the DQ remains as the incessant ongoing procedure. The HCOs must determine the statutory policy and complement the expectation with the healthcare professionals that DQ inspection is the ongoing procedure. The DQ inspection process is divided into three different principles for consideration, as follows:

- a)* Establishing DQ inspection procedure;
- b)* Establishing responsive terminology to HCOs;
- c)* Storing the DQ addressing procedure;

Due to circumstances over time the HCOs goals may change and switch to others, therefore, new DQ measurement dimensions need to be set and the old processes removed as the maintenance standard. The DQ monitoring process will help to determine and resolve future DQ issues based on that which was suitable according to the HCOs policy. This procedure can prescribe to a potential stakeholder that will be responsible for the DQ inspection procedure.

The social application also can be considered through centralised feedback. It is impossible to set a single definition for the meaningful use of EHR systems because the definition always depends on the final regulation of the needs. Five major expectations from the EHRs integration, are as follows:

- 1)* To improve healthcare service quality, reliability, competency and reduce health inequality;
- 2)* To conclude patients and families in healthcare;
- 3)* To develop healthcare synthesis;
- 4)* To improve public healthcare;
- 5)* To improve privacy and policy;

The EHRs integration system is a significant and complex procedure. The EHR system is one of the revaluation technologies that particularly has wide diverse convenience. Healthcare with EHR systems provides quality healthcare in all aspects. Figure 7.3 demonstrates the expectation and conceptual approach of EHRs integration, as follows:

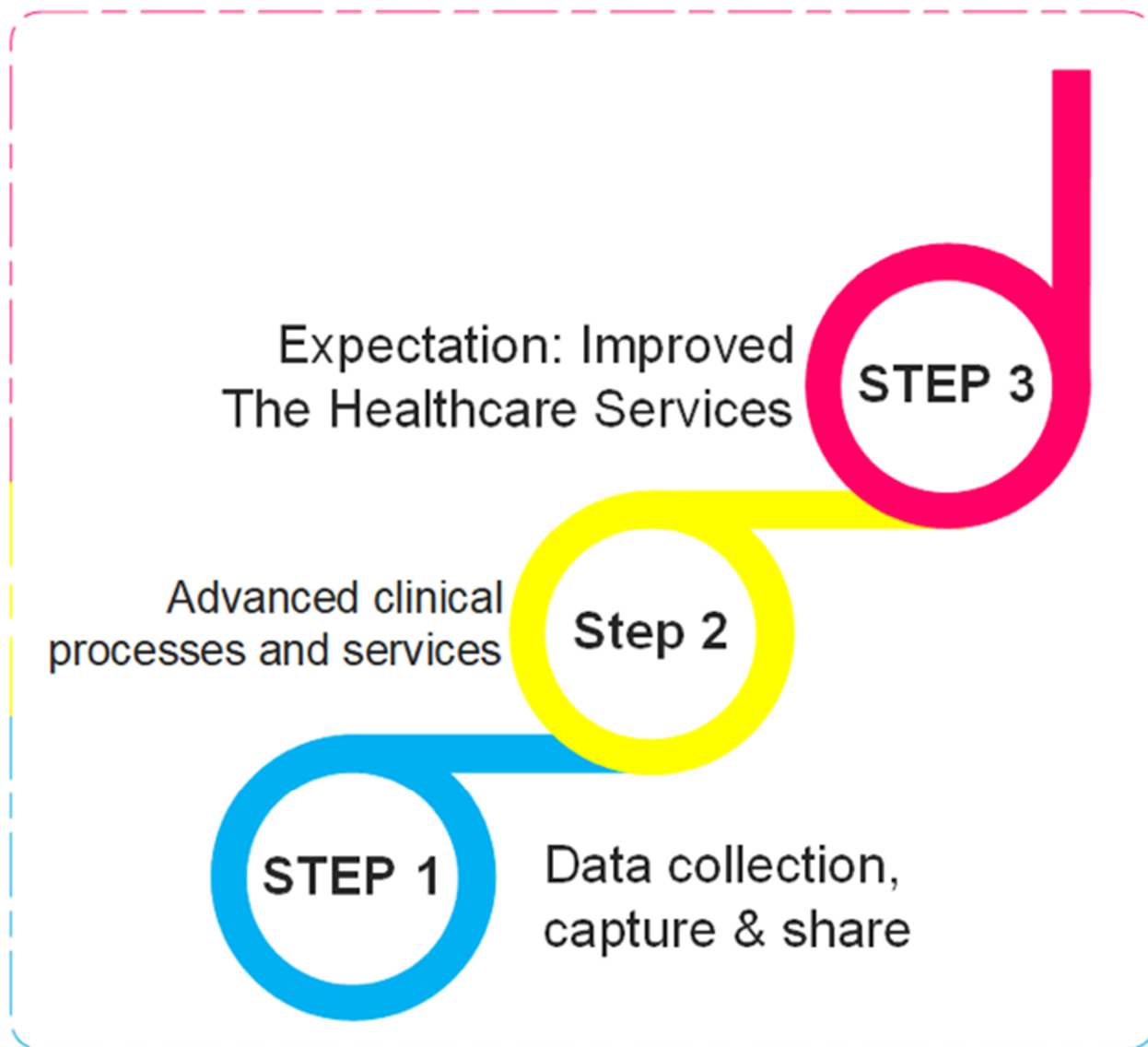


Figure 7.3: Expectation and conceptual approach of the EHRs Integration  
(researcher source)

*Patient expectation:* Population-based statistics recommend that online EMRs access and email communication are desired by the patients. Patients expect the

possibility of accessing their own data from HCOs when needed. The EHRs interface should be simple and easy to handle for the consumer. The security management should be in place according to privacy and policy of the HCOs.

*Provider expectations:* The EHR system and integration includes the process of retrieving health data and manipulating them to use decision-making purposes. Insufficient information is unable to provide accurate diagnosis and would not be able to give the right decisions. Therefore, complete and accurate data is able to offer better healthcare diagnostics and treatments. Complete and accurate data prevents wrong treatment as this can cause even death.

The patient privacy and policy issues need to be in place according to each country's cyber health law after produced at a board meeting of the consumers.

#### **7.4.2 Phase Two – Strategic planning (Designing the framework to the modelling process)**

At tacit intellects, this study focuses on understanding the health-care services processes and the integration requirements to address the DQ issues in EHRs for the LSDB. Tacit intellect implements processes not only to assist in changes but capitalises on acquired knowledge to ensure that EHRs have a high quality of comprehensive DQ requirements, such as a comprehensive EHRs integration method to improve DQ for quality healthcare services. In addition, healthcare service processes are improved, a new product and new systems added or existing ones redesigned.

This study used an HM based on fuzzy-ontology methodological logic to obtain a complete integration of the EHRs and through close health services/IT collaboration, to then determine a cost-effective solution to delivering those needs. The study at tacit intellect understands that methodologies alone do not make integration successful. The IT professionals apply an HM to make HCOs success repeatable, but ultimately the success depends more on clinicians than

on the process. Figure 7.4 describes the strategic planning of the EHRs integration of the framework, as follows:



Figure 7.4: Strategic planning of the EHRs integration of the framework (researcher source)

A hybrid method based on fuzzy-ontology deals with the challenges of multi-scale models in which coupling is very high within and among scales. IT approaches together with the hybrid method will help to deal with these challenges. Moreover, the fuzzy-ontology based methods will improve the modelling process by the mechanisms to deal with the DQ issues. The ultimate aim of the proposed approaches is to enhance the EHRs model development and integration processes by providing the tools and mechanisms in EHRs.

#### **7.4.2.1 Data Profiling (Understanding data sources and associated quality)**

Data profiling is defined as a complicated initial procedure in assessing the EHRs that populate this data profiling, also called data archaeology, which empowers IT, teams, to evaluate the quality of information before using it in any data integration processes. Data profiling is the systematic analysis and calculation of data contained within the given information, which is obtained from a data set. It aims to look for consistent, unique and logical information. Data profiling collects data from a database or a file to determine whether the data can be used for other functions and to improve the searching system using appropriate keywords, descriptions and categories. It also provides metrics from the data gathered to check whether these correspond with specific patterns or standards applications, but also a demand to control, handle and inspect their existence value.

To manage the healthcare services using EHR systems, it is an essential part of the development process that data structure and the relationship to other components need to be clear enough for efficient integration and data accuracy as the key of business control. Determining, decomposing and clarifying the status of the enterprise information in advance, are as follows:

- a) Identifying the present status and determining all missing, mismatched and corrupted data;

b) Analysing and preparing about all risks in advance in case of data migration or integration failure.

Figure 7.5 describes the data profiling modelling process in EHRs, as follows:

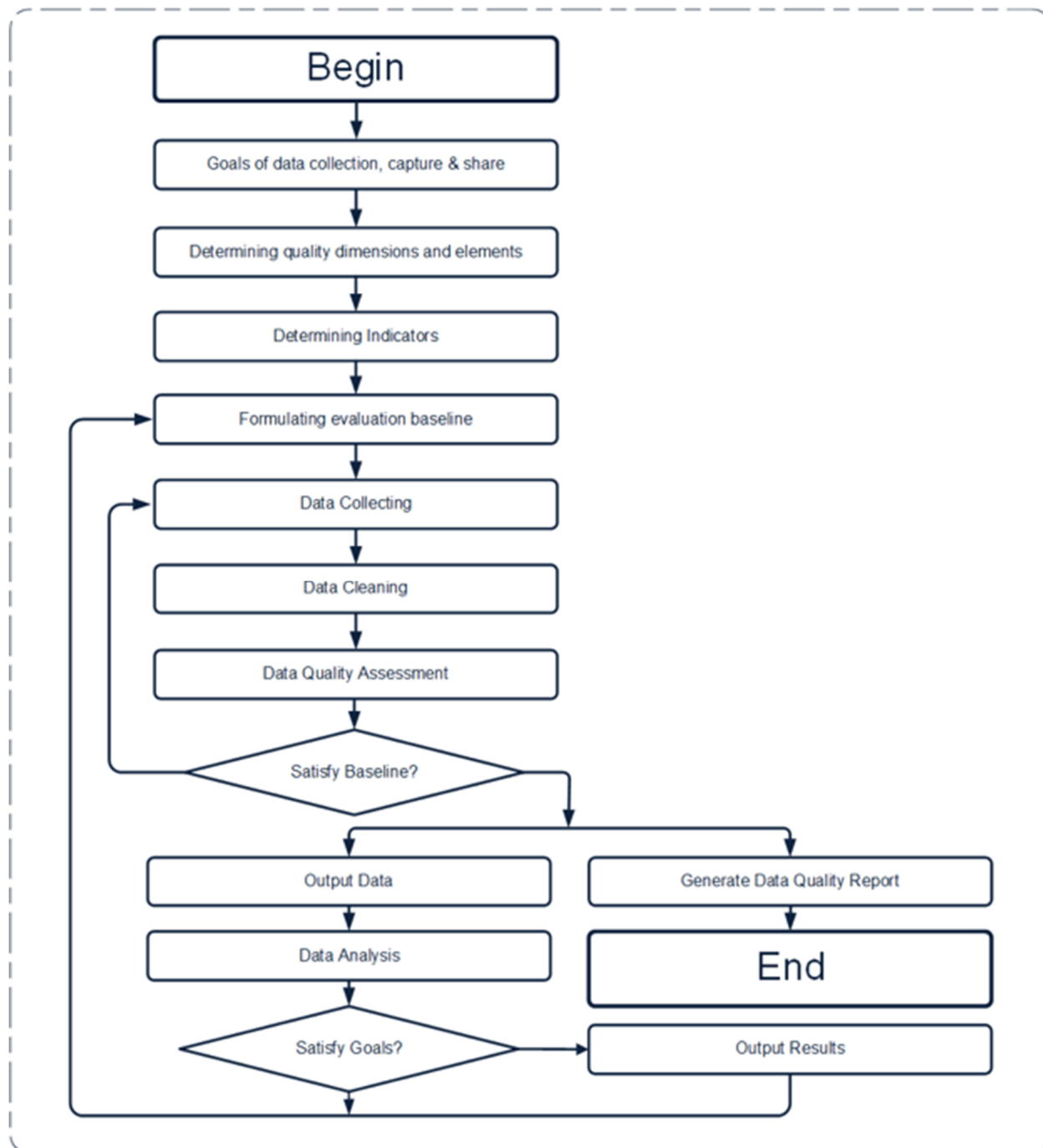


Figure 7.5 Data profiling modelling process in EHRs (researcher source)

The upfront data profiling process reduces the migration or integration failure risk to the minimum failure. Data profiling must, however, be embedded into insights in the EHRs procedure across data sources throughout the HCOs.

The data profiling modelling process applies for a pre-crafted automated data profiling process to expose unforeseen data correlations in the HCOs business logic across disparate EHR systems. This automated process expresses the relationship when integrating from diverse heterogeneous sources. This can be applied in statistical data auditing and analysing of exorcism to determine the DQ issues, as below:

- a)** Inaccurate data;
- b)** Data absence;
- c)** Data exception;
- d)** Mismatched data;
- e)** Empty or Null;

The data profiling process breaks down into the following six groups, are detailed below:

- a) *Column or attribute profiling:*** Column profile is defined as analysing value through each column to explore the actual metadata and discover DQ issues.
- b) *Dependency profiling:*** Dependency profiling is defined as the data relationship comparison between other attributes across a table. The functional detection dependencies are focusing on two aspects, as follows:
  - 1)** Primary keys;
  - 2)** DQ issues associated with data formation;
- c) *Redundancy or abundance profiling:*** Redundancy profiling is defined as the identical dataset comparison across a table. This procedure is to determine and address duplication across a table or EHR systems. The functional detection redundancy is focusing on three aspects, as follows:



- 1) Foreign keys;
- 2) Synonyms;
- 3) Homonyms;
- 4) Corrupted data;

*d) Transaction profiling:* Transaction profiling is defined as analysing the transaction process (business logic) and the target domain of the integration process. Functional transaction profiling is focusing on two aspects, as follows:

- 1) The purpose of the transaction;
- 2) Discover the target domain;

*e) Security and safety profiling:* Security and safety profiling is defined as the identification process of the user roles to the EHRs and that which they are authorised to access the data (view, insert, edit, delete, etc.).

*f) Custom profiling:* Custom profiling is defined as the data analysing, the process that is meaningful to the HCOs. This procedure is to identify how the EHR system is used by the organisation and consumers and to improve the system according to the findings.

Data profiling basically follows two methods, detailed below:

- 1) *Sample-based methodology:* The sample-based method is defined as the performing data analysis according to a specific data sample. This methodology requires representative data as a profile. For example, one might want to profile a 100 million row table. In EHR systems for the effort to be efficient, the sample data might be 30% of the rows where the EHR systems select every third row. Sample base profiling requires EHR systems to store the integration data sample in some temporary medium. Also, sample-based profiling requires ensuring a representative sample of the data. From a statistical standpoint, if the integration data is too small, the EHR integration systems can easily miss data patterns or not properly identify the column's domain.

2) *Profile-based methodology*: The profile-based method is defined as an additional query to compare the accuracy probability with the latest successful integration.

A key to successful DQ endeavours is to highlight the importance of data profiling and auditing as an important and necessary precursor to the success of the DQ endeavour. Without data profiling and auditing, a DQ endeavour will remain a confused activity and will miss addressing actual specific problems in EHRs. This study focuses on the practical problems encountered in a DQ endeavour with an emphasis on data profiling and auditing as well as a complete DQ process flow.

#### **7.4.2.2 Defining the gap between what data is available and the quality versus what the business logic**

The definition of DQ involves whether the data is correct, consistent and incidental, reliable and complete for further processing. One can also alternately define DQ as the availability of the right data in the right place at the right time. DQ is an indication of foresight and planning about data, which has been observed in designing and implementing any data capturing application or system. DQ is not just a preventive measure but also a corrective course of action in view of legacy systems and their data problems. DQ has shown strong benefits by improving customer names and addresses, which is one of the most cherished data in HCOs.

A gap analysis is essential to all HCOs conducted processes to measure and analyse captured data over the EHR system as a clinical workflow procedure. The initial expectation of EHR systems is to integrate data from diverse heterogeneous sources. The gap analysis is based on the data requirements set.

This an iterative and high-level step for the EHR systems implementation process and depends on:

- Data capturing and workflow redesigning;
- Data extracting;

➤ Data validation;

This gap analysis process introduces the HCOs to determine gaps in discrete documentation, particularly between the data structure measurement and documentation as a narrative text. Figure 7.6 describes the EHRs gap analysis process diagram, as follows:

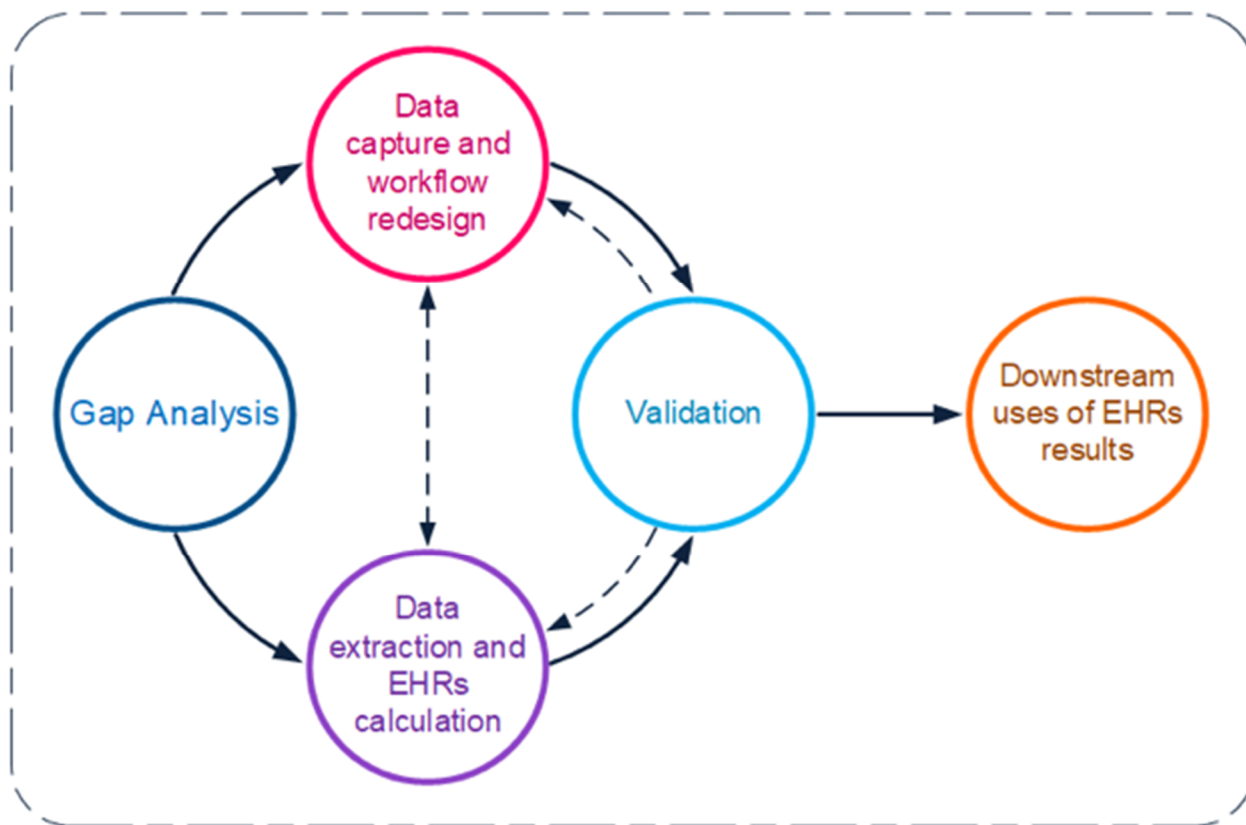


Figure 7.6: The EHRs gap analysis process diagram (researcher source)

### 7.4.3 Phase Three – Analysis (*Pattern over time evolve the data integration architecture*)

Most EHRs are integrated and stored in an extremely unprecedented way. Profiling and integrating of huge data sets are time-consuming, expensive and an obstacle to the investigation without a proper integration system. Therefore, the big challenges are associated with meaningful EHRs integration to store, extract, handle and analyse. Decision Support System (DSS) methodology such as a hybrid

method based on fuzzy-ontology is very effective to decide on the support system regarding data extracts, groups and analysis. Figure 7.7 (Saïod *et al.* 2019a) describes the DQ process model in EHRs, as follows:

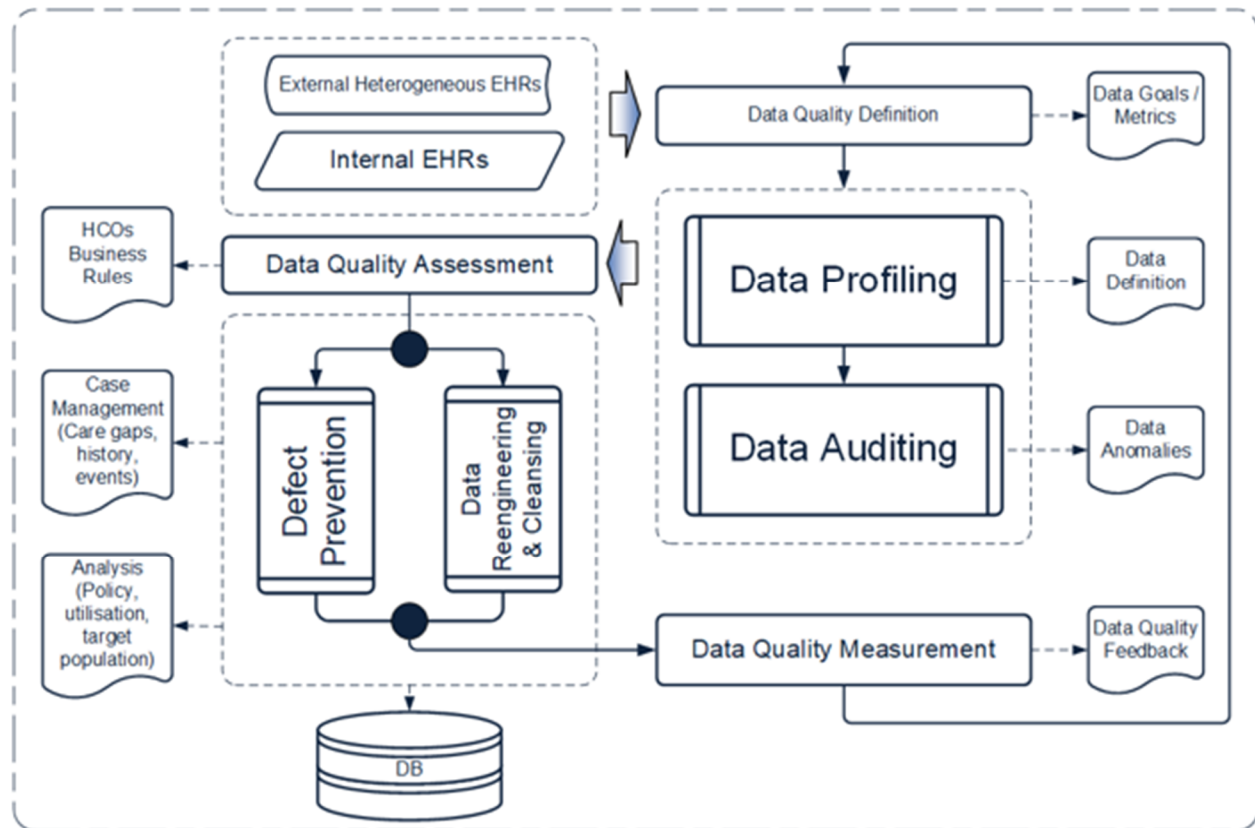


Figure 7.7: The DQ process model in EHRs (Saïod *et al.* 2019a)

Three types of EHRs analysis, are as follows:

- **Policy assessment:** Policy assessment is defined as the identification of new policy according to the patient's data and healthcare specimen initiatives;
- **Determination and stratification:** This process model uses data from cross-agency to analyse and determine new targeting applications, redesign the intervention strategies and ameliorate differential healthcare services;
- **Utilisation analysis:** Utilisation analysis is defined as the way to analyse cross-agency data to identify feasible resources to allocate scarce application;

The EHR integration implementations require complete structural design and infrastructure such as the location where the integrated data would be stored and the way in which the data would be transferred from the data sources to the other sources. The methodology to be used for data profiling, mismatch and inconsistency detection also needs to be ascertained. A successful DI should require the following features:

- Completing data structure;
- Data matching and mapping tools;
- Data retrieval and delivery functionality;
- Providing data security;
- Data management process;
- Proper infrastructure tools;

The hybrid architectures combine the architectural approaches to the third-party data transaction for incident governance purposes, whereas statistics data is used to meet the real-time requirements.

*Combine two figures at work:* Hybrid Approach Integration (HAI) architecture is used for policy and population analytics. The performance for the DQ in EHR systems determines the individuals and procedure for governing, assessing and synchronising the data for correctness, accurateness, timeliness and conciseness of data. Some values need to be set as the DQ standard for effective EHR management, such as priorities, expectation and metrics for EHRs integration. The cross-data measurement dimension of DQ constitutes multi-level disciplinary for governance.

These days heterogeneity constitutes common cases for both target and data sources, due to several data standards, application tools and data formats. As these destinations and target data domains have different data structures, the integration should transform through the sagging process for efficiency. Therefore, the integration system can handle these complex and diverse data to achieve certain goals. Integration should fulfil in such certain orders, as follows:

- *Structural model as development standards:* An integration solution combines three category solutions, as below:
  - 1) Data server;
  - 2) Integration interfaces;
  - 3) Data transfer;

The data integration structure is a general model developed when the data server is introduced across interfaces. The integration structural hub is to provide an overview of the implementation infrastructure; therefore, all stockholders will have a clear overview of the integration process collaboration. The hub can provide the guideline for the feature developer (someone else) to follow the development standard and implementation model. Well-organised HCOs have development standards and implementation models and should be developed within these.

- *Ingenuousness for reapplication and consistency:* As design standards and implementation models are introduced to the diverse data integration process, the outcome of ingenuousness, is to easy reapply for other similar cases, which enhances consistency in the processing of data.
- *The coherence among common and custom application:* Integration systems should be a well-organised solution that enables targeted architecture.

#### 7.4.3.1 Data security and visibility

Managing EHRs security is a complex and difficult procedure. The EHRs network and business requirements are rapidly growing and require new well-defined security management for the delivery services. The traditional security management works with a single point of view but fails with multiple points of procedure. Efficient security management prevents any anticipated threats. The definition of visibility is defined in the Oxford dictionary as “*the state of being able to see or be seen*”. According to this concept, visibility is the ability to provide an

unhindered view of the security management system and could apply to cyber-security and security management. This feature will provide easy access to EMRs and managing them. Security data management is all about visibility. HCOs need total visibility, without blind spots, to effectively manage, as below:

- Network security, both inline and out-of-band;
- Network and application performance monitoring;

Figure 7.8 describes EHRs data security and visibility, as follows:

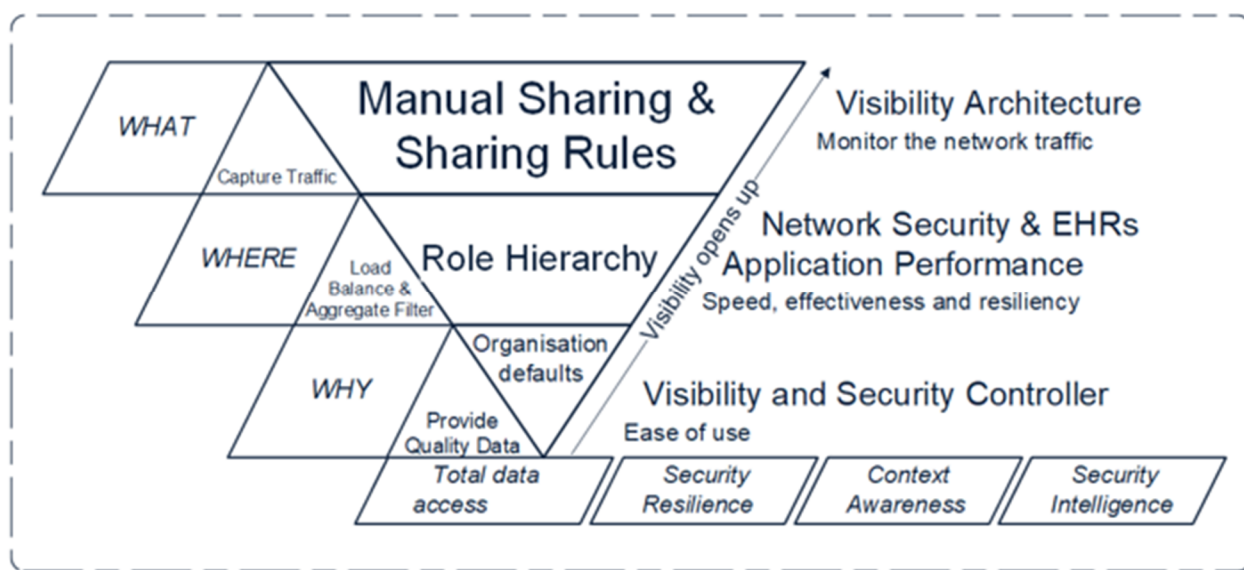


Figure 7.8: The EHRs data security and visibility (researcher source)

**Record-level security:** To control data access precisely, particular users may be allowed viewing specific fields in a specific record. But they might not be allowed to view the other individual objects. Record access determines which individual records users can view and edit in each object they have access to in their profile. The following two questions below first need to be identified:

- What row-level security needs to be implemented (open access or just a subset)?
- What rules need to be implemented to access the portion of data, if it is a subset?

Let us say a new profile called the recruiter to give recruiters the object-level permissions they need. They restricted the privileges to modify permission recruiting-related data, therefore, recruiters cannot perform the updating or delete any records. However, the security rules must obey the recruiter's granted permission to create and read recruiting data. This does not mean that recruiters may have viewing permission to every record in the recruiting records. This is a consequence of two essential apprehensions, listed below:

- 1) The authorisation level must combine the object-level and row-level security;
- 2) In case of conflict between the object-level security and row-level security, the most restrictive settings should have the privilege.

This concept means that even if the administrator grants a user to create, read and edit permissions on the recruiting objects, but the row-level security for an individual recruiting data is more restrictive, these are the permissions that identify that which the user can access. The control record-level is accessed in four ways, listed in increasing access. Org-wide defaults specify the default level of access users have to each other's records. Role hierarchies ensure managers have access to the same records as their subordinates. Each role in the theocracy explores a level of record access that a user or group of users need. Sharing rules are the automatic elimination to org-wide defaults for specific user groups, to give them access to records, they do not own or cannot normally see. Custom sharing lets record owners give read and edit permissions to users who might not have access to the record in any other way.

***Increasing levels of visibility:*** The visibility and access for any type of data are determined by the interaction of the above security controls, based on these key principles. Users' baseline permissions on an object are determined by their profile. If the user has any permission sets assigned, these also set the baseline permissions in conjunction with the profile. Access to records that a user does not own is set first by the HCOs defaults. If the HCOs defaults are anything less than public read/write, one can open access back up for certain roles using the role



hierarchy. One can use the sharing of rules to expand access to additional groups of users. Each record owner can manually share individual records with other users by using the *share button* on the record. The way in which to configure object-level and record-level permission using profiles and access sets have already been viewed and details of the various record-level security controls are shared.

### 7.4.3.2 Performing a data quality evaluation

Data Quality Evaluation (DQE) is the scientifically and statistical assessment process defining whether data meet the quality including the data type and amount to be capable of the actual support to the business process as determined by the HCOs. The DQE deliberated the quality assessment guidelines and methodology to evaluate the assessment for the considered application components to improve the DQ in EHRs for the LSDB.

The DQE process quantifies the DQ issues with real and empirical EHRs, which accommodate the HCOs to evaluate an accurate business plan to address the DQ and prosperity strategies. This process generally applies in the integration process to meet the DQ standard and accordance prospective. Often the DEQ process easily identifies and addresses the issues regarding the general data structure, standards, missing, mismatching and inconsistent data, but the difficulty arises when the data structure is more complex. Therefore, the more defined DQE process methodology needs to be applied in the identification and addressing process.

Generally, DQE provides the subjective correction associated with healthcare administrative processes, such as providing immaculate reports to assure that the EHR systems are performing as anticipated regarding the data collaboration and dependencies.

Figure 7.9 describes the dimensions between DQ and DQE, as follows:

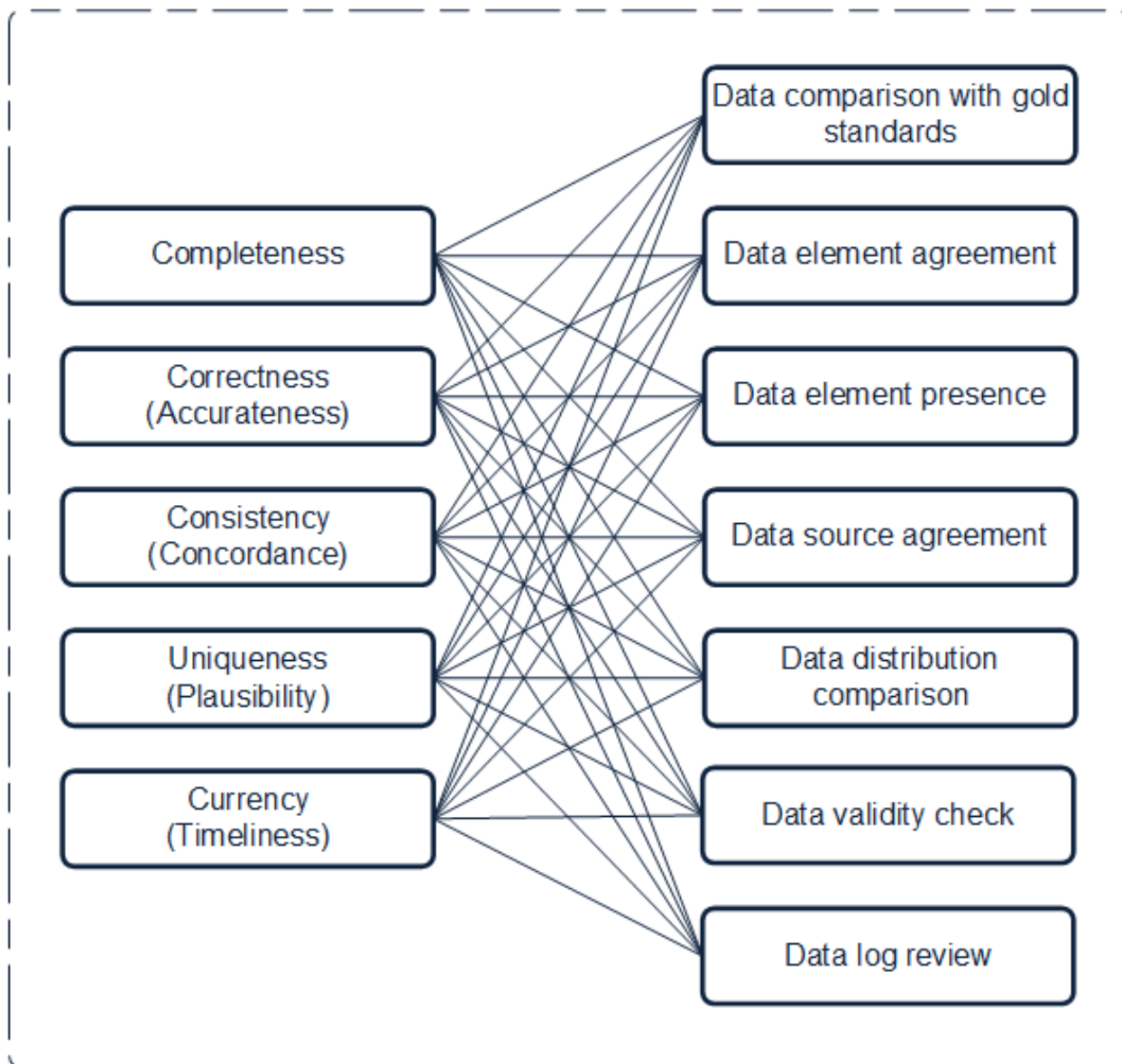


Figure 7.9: Dimensions between DQ and data quality assessment (researcher source)

According to Nicole *et al.* (2013), DQE processes are aligned with best practices and a set of prerequisites as well as with the five dimensions of DQ”, as follows:

- a) *Completeness*:** Have all patient data sets and components been stored in the EHR systems?
- b) *Correctness (accurateness)*:** Does the data reflect the EHRs set?
- c) *Consistency (concordance)*:** Does that set match across the EHR systems or among other HCOs’ other EHRs sources?

- d) Uniqueness (credibility):* Is there a single view of the EHR that gives a clear view of the concept regarding which the components are assessing?
- e) Currencies (timeliness):* Does the EHR match the rules and according to the patient conditions at a certain point in time?

Seven magnificent definitive denomination methodologies for the DQE, are as follows:

- a) Data comparison according to the HCOs' standard:* Comparing the integrated dataset with a specific dataset pattern that has been set as the HCOs' standards during the integration process;
- b) Data components compliance:* Several data components are compared across EHR systems to identify the data component compliances;
- c) Data component subsistence:* A denomination process to identify whether intentional EHRs exist;
- d) Data origin arrangement:* A denomination process to identify if any data origin agreement exists in EHRs;
- e) Data allocation assimilation:* A denomination process of data allocation to provide a summary of aggregation according to the HCOs' concept of interest.
- f) Data actuality verification:* A denomination process of data actuality verification on whether the EHRs make sense;
- g) Datalog reconsideration:* A denomination process to determine whether the substantive EHRs admission practices are verified;

*Other DQ enumeration:* The seven DQ definitive dimensions management is a complex task and difficult to comprehend. A few other complementary factors exist to efficiently manage EHRs. The DQ achievement intention still may remain below satisfactory, whereas the seven dimensions are considered as appeasement. However, EHRs may seem accurate, complete, timely and valid, but may fail and might be useless when data will be shared globally due to a different language, time zone and currency. To make them meaningful in the global share some additional three factors need to be considered, as follows:

- a) *The EHRs suppleness*: Are the EHRs flexible regarding compatibility and comparability with another EHR system?
- b) *The EHRs dependency*: Is the EHRs' security functional regarding data administration, manipulation, verification and protection?
- c) *The EHR standards*: Do the EHR systems operate meaningfully according to the HCOs' DQ standards?

The difficulties arise in this conduction when the DQ analysis is provided from the inconsistent EHR systems terminology. DQ dimension may overlap and occur inconsistently when integrating data into a single point based on different languages. Therefore, this is one of the reasons when health research denomination fails regarding interoperability. According to Wang *et al.* (2015), "conceptual framework of DQ (*for example, contains 15 dimensions*) grouped into four categories", as below:

- a) Indigenous;
- b) Contextual;
- c) Correspondence;
- d) Approachable;

The study focused on idiosyncratic (indigenous to the EHRs) and contextual (obligation related) DQ issues. The study has determined two overlap dimensions which are identical to accurateness and thoroughness of the idiosyncratic properties, as follows:

- 1) Appropriateness;
- 2) Dependability;

### 7.4.3.3 Revising the expectations and determining the selected data solution

Today, the EHR systems have begotten and suppurated in HCOs, therefore, the expectations have expanded to the full range of healthcare management systems.

The EHR definition has obligated in scope interoperability, which became a challenge for small HCOs. It is essential to do a realistic expectation analysis before embarking on the EHR systems implementation. To acknowledge all EHR possible benefits and challenges is a very troublesome task. Therefore, these considerations should draw meaningful attention to these issues profusely pertinent to the smaller HCOs, before providing the accomplishment guidelines. This intentional section is not to provide a magnificent analysis regarding the EHRs' benefits and challenges, but the section subjected the significant issues associated with the small HCOs.

Most HCOs prefer to provide more healthcare services and other administrative works more efficiently and securely within the same amount of time. The EHR systems assure more diaphanous, accurate, seasonable and manifestly to extend the HCOs' team collaboration performances. Baron *et al.* (2005), stated that "the HCOs expected to free their file room space and make it clinically productive". The egregious consideration of EHR systems is that it propagates luxation between healthcare staff while transforming paper-based healthcare systems to complete EHR systems. This is extremely realistic, more complex, critical and expensive than the majority HCOs anticipated. Therefore, the principal strategy of first discretion is to introduce all stakeholders to effective communication in the EHRs' implementation.

HCOs must motivate all clinical staff to provide input into the EHR systems implementation, to set the goal, expectation, determine meaningful capacity and debilitation within the HCOs. The early preparation of stockholders will assist comprehensively during the EHR systems realisation. This early preparation will help stockholders to understand and realise the development process and impact of the EHR systems. The EHR systems performance will be convicted due to the vindicated failure when a clinician does not have a clear understanding of the entire new systems feature. It does not matter how efficient the system is and how prepared all are for its realisation, most of the HCOs will initially contradict during the realisation of an EHR system. Four apprehensions should be considered and addressed during the realisation process, as follows:

- a) Introducing the patient;
- b) Managing and manipulating the changes;
- c) The fast realisation and efficient support;
- d) Motivating the HCO;

Several coefficients, such as captaincy, technology, preparation, administrative action and HCOs conditions, are reflecting on the EHR systems realisation perception. These coefficient conjunctions lead the differentiation of the EHR systems realisation perception.

According to Cynthia *et al.* (2016), “the majority of providers and other healthcare providers readily learn, collaborate and transform their daily work”. After all these complex, critical and difficult transformation, once the workflow and the EHR system introductions are in place, the EHRs’ benefit will be realised by the HCOs.

#### **7.4.4 Phase Four – Implementation (Applying the framework to the modelling process)**

The EHR systems implementations in HCOs are very diverse. Phase six represented the conduction of a well-concerned literature review of an exploratory study on the EHR systems realisation. The general EHR systems realisation commencement stoop to be the process of EHRs integration and local health information. The EHR system realisation becomes more critical when introducing the LSDB for HCO networks concerning a wide range of technological coefficients (infrastructure), health professional skills, organisational composition, culture and pecuniary resources combination.

Although plenty of positivity exists and healthcare performance is anticipated by an EHR system, its realisation is a critical commitment. This well-concerned literature review expresses the motive for this critical scope of the phase four interference, which will assist to defeat the emblematical issues in the EHR systems realisation. This phase can provide a functional guideline for IT

professionals to simulate efficient EHR systems for strategic realisation. Numerous challenges and barriers are present in the EHR systems realisation for HCOs and the procedure is determined in three echelons, as follows:

- a)* Pre-realisation;
- b)* Realisation;
- c)* Post-realisation;

According to the system simulation, each echelon coefficient introduces life-halo, formal and fortuitousness architectural theory. These coefficient doctrines are accumulating the architectural model and the cardinal apparatus based on maximised anticipated reimbursement used to evaluate the model and represent the EHR systems' meaningful uses.

#### **7.4.4.1 Modelling the data stores necessary**

The implementation of EHR systems is a continuous ongoing process task, including the expansion to new features. EHR systems envisage massive amounts of patients' stored data. The DQ investigation and amelioration are the key considerations for the EHRs adaptation that reassure the quality care services and reduce the potential healthcare service costs. The EHR systems' flourishing interoperable features make it feasible to manage a large amount of health data globally.

The EHR systems are capable to manage and manipulate large health data in the integration and application process of stored data. Generally, EMRs cover more articles than EHRs and large-scale data is the core concept in this term. Cloud computing is the combination of a distributed computing system that is hosted in the cloud. This is one of the predictive fast-growing development technologies and widely used. Data visualisation is a complex and pointed methodology of which the interpretation assists the provider in the decision-making process. Monitoring is defined as a comprehensive apprehension, which includes patient health condition monitoring, medication monitoring and administrative monitoring.

Regarding patient privacy and confidentiality, access control technology remains an important consideration for EHR systems. According to the Health Insurance Portability and Accountability (HIPAA) Act of 1996, “de-identification in biomedical informatics usually means removing sensitive personal information to protect privacy” and this technology covers differential privacy rules. Health specification, such as distemper phenol-typing is a recent evaluation regarding impersonate medication in EHR systems. Data modelling is defined as the data component designation process in the integration process, including the data relationship establishments and their constraint. Four indications regarding data modelling are detailed below:

- a) *Data types and formats:*** NUMERIC, DECIMAL, INT, VARCHAR, CHAR, DATETIME, TIME, etc.;
- b) *Obligation and constraints:*** Determining unique, missing, miss-field value;
- c) *Row data relationship:*** Row data may have one to one or one to many or a one to none relationship. Therefore, a row theocracy can ascertain a dataset concept;
- d) *Metadata specifications, sequences and considerations:*** Defining data process flow control, including the meaning, use, data collection method, validation and relationship of each component;

The data model describes data structure and the metadata concept of the process control and defines the data interpretation, subset description and data extraction. DQ check, such as adult patients cannot have pediatric treatment, requires experience that does not subsist in the data structure, but this concept is applied in the DQ process control and analysis. This confirms that the EHR system can create a DQ analysis procedure with a knowledge base of the data formation including the integration and storage.

#### **7.4.4.2 Data extraction**

The EHRs evaluation with new scope and benefits for conducting health data as a diffusive healthcare information terminology for quality assessment emerges with



respect to DQ, compatibility and availability. The data extraction objective is defined as the comparable DQ test procedure on whether the integrated EHRs satisfy the DQ standard initiative.

It is a technological challenge to measure Data Quality Indicators (DQIs) while extracting the accurate pieces of data from EHRs. Although the analysis on assimilating EHRs is problematic, the DQIs of EHRs is capable to apply for the measurement. The data standardisation process simultaneously compares data collection methodology and standards according to the different HCOs. Generally, EHR systems using Natural Language Processing Tools (NLPTs) explore narrative data, although EHR systems may have the functionality to handle them. This type of procedure reduces the similitude void according to DQIs on survey-based EHRs.

EHR systems are capable to ensure real-time DQ measurement to a healthcare professional, but some critical indication can make it difficult, such as vague and imprecise data including text notes. The routinely integrating diverse heterogeneous health data for business analysis, DQ enhancement and investigation, vindicate this extended procedure for the EHRs adaptability and inoperability purposes.

#### **7.4.4.3 Data transformation**

Data transformation refers to the conversion of the dataset into a unified form suitable for data mining. Data transformation methods include smoothing noise, data aggregation and data normalisation. According to the direction and target of data mining, data transformation methods filter and summarise EHRs. Data analysis can be more efficient by having a directional, purposeful data aggregation. The EHRs should be normalised to make the data fall into smaller common spaces, to avoid the subsection of the data attributes on the DQI units. Three forms of normalisation are as follows:

- a)* Min-max normalisation;
- b)* Zero-mean normalisation;

**c)** Fractional scale normalisation;

For neural network algorithms or classification algorithms based on distance measures (*such as nearest neighbour classification*), the normalisation method works better. The data amount can be reduced and deteriorated during the process; in this way, EHRs can fail to meet the business requirements. The EHRs lose data for a different reason during the collection or process (*for example, data diminishing, mismatching or error etc.*). The matter of EHRs secondary uses is the subject of a different analytical procedure and healthcare efficiencies (*for example, data conversation, formation, coding and transaction*). Each data visualisation considers data analysis according to the real-time ambience to allow data modifications to avoid possible mistakes, data reduction and/or data decay. This process can alternate particular data and even the distributional concept of the EHRs. Sometimes, this process can produce extrinsically and extremely inconsistent datasets.

#### **7.4.4.4 Data loading**

EHR is defined as an electronic version of a patient health record, which have the potential to improve quality healthcare services. Due to the incredulity associated with the EHR systems development and implementation cost, including the HCOs productivity impact and loss of revenue, the EHR systems adaptation and interoperability remain still below expectation. To control interoperability standards, EHRs are adapted to extract data from source systems, analyse and transform the data as necessary. It has to be prepared for loading into a repository accessed by end users or the population health analytics. Three main features must be present in EHRs loading processes detailed below:

- a)** Extracting data from outside sources;
- b)** Transforming (*cleanse, normalise, translate*) data to fit operational needs;
- c)** Loading data into the target database;

### 7.4.5 Phase Five – Data validation (*Extract the underlying DQ in EHRs for LSDB*)

Data validation is defined as an agility membership validation process for an admissible alliance between specific datasets. Specific individual datasets prescribed the verification activity without any specific indication of consistency between several datasets to the validation process. When the data validation determines any interpretation that the specific dataset belongs to the certain predefined dataset, it identifies the strictness and acceptable current data validation process. In addition, when a validation process is unable to decide whether to accept or reject the datasets due to some of the complex error allocations, the antecedent concept verification will be too comprehensive as this correction procedure will not strictly rely on the validation procedure. Generally, the initial help of the data validation ensures the correctness, completeness, secureness and conciseness according to the DQ standards to the associated applications. This is an ongoing process which is acquired during the continuous validation process. These logics are built-in into the EHR systems and identified in the EHRs dictionary and accomplished in the development process. Five different types of data validation processes are accomplished in EHR systems:

- a)** EHRs encoding validation;
- b)** EHRs complexion validation;
- c)** EHRs extent validation;
- d)** EHRs coercion validation;
- e)** EHRs framework validation;

According to the DQ standard, data validation processes ensure the specific DQ level of the destination data. DQ has certain dimensions in HCOs statistics, as follows:

- a)** Appropriateness;
- b)** Correctness;
- c)** Timeliness;

- d)* Preciseness;
- e)* Availability;
- f)* Limpidity;
- g)* Comparability;
- h)* Compatibility;
- i)* Completeness;

The next significant aspect is to corroborate which dataset needs to be validated as the business concerned. The first data validation phase is defined as the **modelling phase** with more exceptions to the development and exploration process. The statistical functionality is implemented while the **development and exploration** process are done which includes the EHRs encoding, correcting, capturing, integrating and validating declaration of datasets. Two principal integrity constraints of data validation are as follows:

- 1) System reconciliations;
- 2) Finding missing documentation and the reason for it being missing.

Data validation includes four different components, as follows:

- a) Credibility:* Trustworthy results?
- b) Completeness:* All valid codes entered?
- c) Reasonability:* Unexpected spikes/changes?
- d) Consistency:* An example – an adult person cannot have pediatric values;

#### **7.4.5.1 Identify the base DQ standard specified in the EHRs integration for LSDB**

The DQ identification is a critical process, which ensures patient safety to guarantee quality healthcare services, governing healthcare flow control systems and healthcare reputation. DQ standards require defining the data consistency and assessing the DQ. The procedure clearly identifies between the HCOs requirements

and healthcare professionals to promote the DQ standards for the EHR systems inoperability. This evaluation process combines the HCOs data standard level and EHR systems process to ensure EHRs appropriateness, correctness, timeliness, preciseness, availability, limpidity, comparability, compatibility and completeness.

Data standards are *“documented agreements on representations, formats and definitions of common data. Data standards provide a method to codify invalid, meaningful, comprehensive and actionable ways, information captured in the course of doing business”*.

These rules provide the way to capture data to ensure the DQ dimension across diverse heterogeneous sources, as another definition of DQ standards. Only the DQ standards prevent the bleakness of the future EHRs inoperability, where data fields and the components of these fields need to be standardised. DQ standards address diverse issues associated with the EHRs' procedure.

The EHR experts need to take the initiative to lead the HCOs process control, including collaborating with all internal and external stakeholders to meet DQ standards to ensure that EHRs contents are accurately identified, understandable, perfectly accomplished and maintained. The list below identified some leadership actions for EHR professionals, which include but are not limited to:

- Enhancing the HCOs process flow control and intellect of the DQ standards;
- Analysing the HCOs formulation;
- Introducing the assessment for the DQ standards necessity;
- Implementing a native EHRs dictionary to patronage interoperability over similar networks;
- Advancing the evolution of DQ standards;
- Eliminating integrated EHR systems which accept the EHRs necessity;
- Introducing the knowledge base to consequential stakeholders;  
Collaborating with development subscriptions, HCOs and clinical staff standards;

### 7.4.5.2 Classify the base DQ standards in EHRs

After extracting the underlying DQ base standards from all the applicable profiles and methods, the standards should be integrated into a list. Improving the EHR systems competency and abatement costs for the HCOs the EHRs exchange will remain as the essential part, as the world-class standard will change the healthcare service and use of EHRs universally. According to the ASTM E1384 Standard Guide for data and formation the EHR serves the base DQ standards, as below:

- Combined EHRs storage;
- Multiple accessibilities;
- Feasible EHRs exchange between HCOs;
- EHRs integration from diverse heterogeneous sources;
- Better healthcare process flow controls;
- Longitudinal EHRs for effective and efficient use;

However, as advantageous as it may be to develop an EHR, certain standards on the content of the health record are necessary to meet this goal. As Mary Brandt *et al.* 2011 states, *“until healthcare providers collect and maintain data in a standard format according to widely accepted definitions, it is nearly impossible to link data from one site to another. The lack of health informatics standards is one barrier to broad implementation of computer-based patient records”*. The base DQ standards include details below:

- Including a brief statement of each component in the EHRs;
- Determining indispensable elements in the EHRs (for example, age, gender, blood pressure and BMI);
- Specifying each field type (for example, field type and length);
- Accomplishing the EHRs granularity range;
- Implementing various EHRs granularity range within similar clinical data;
- Introducing standardisation for both structured and narrative EHRs;

- Anticipating the principal subsistence of the EHRs to determine the data segments;
- Introducing all existence EHRs in the healthcare analysis control to improve healthcare services and longitudinal RHRs improvement;

### 7.4.5.3 The resulting set of DQ in EHRs

As stated in section 7.4.5.2, it is an essential procedure to exchange EHRs between different HCOs regarding development standards, EHR systems adaptation and inoperability. During the development process, these forefront standards should be considered and implemented when integrating EHRs from diverse heterogeneous sources. Skipping these standards, the EHR system will be unable to provide efficient EHRs integration. The cost measurement systems also need to include clinical outcomes as development standards with billing data to identify the EHRs value, as the cost is determined as to quality over healthcare value. Figure 7.10 demonstrates the DQ management functions and characteristics, as below:

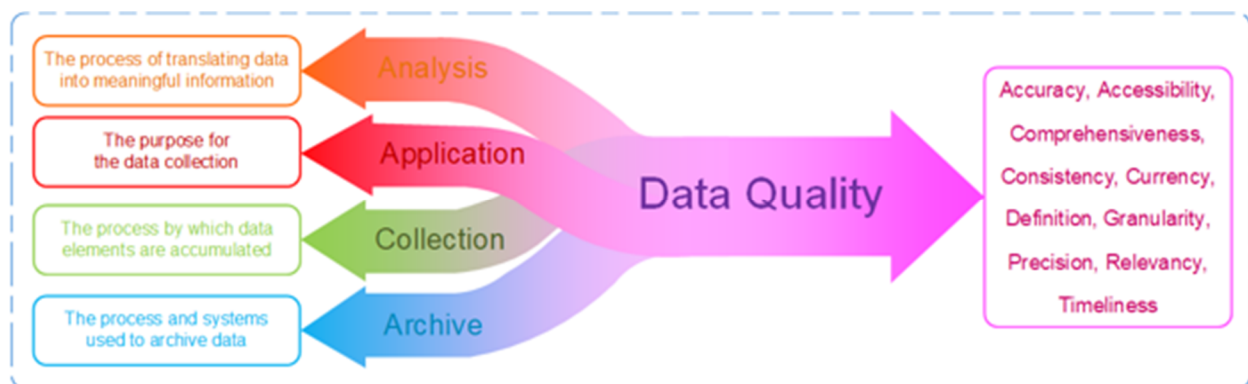


Figure 7.10: Data quality functions and characteristics (researcher source)

Each EHR system has its own standard to represent the clinical data. The final process is to map the functions that should be supported by their applicable profiles and method underlying the integration. The DQ management model domains and characteristics functions include, as follows:

- *EHRs appropriateness*: EHRs appropriateness is defined as free of ascertainable errors;
- *EHRs approachability*: EHRs approachability is defined as data efficiency and ease of use, which includes some other properties, such as legally procurable, secured by access management systems;
- *EHRs comprehensiveness*: EHRs comprehensiveness is defined as extended EHRs that comply with entire scope requirements and documented intentionally;
- *EHRs compatibility*: EHRs consistency is defined as feasible EHRs, which is identical and reusable across HCOs;
- *EHRs timeliness*: EHRs timeliness is defined as a data extended feature that contents are up-to-date. It is also called data concurrency for a specific period within the necessary, efficient or certain time.
- *EHRs assignment*: EHRs assignment is defined as a particular definition of clinical data elements;
- *EHRs granularity*: The EHRs granularity is defined as details of the DQ level that identified the symptom and distinguish clinical data components.
- *EHRs obviousness*: The EHRs obviousness is defined as data range measurements between the nearest components;
- *EHRs appropriateness*: The EHRs appropriateness is defined as the extension of clinical data components to identify the proposed and necessary reasons for which they were collected;

## 7.5 Summary

This chapter discussed the proposed HIDM based on fuzzy-ontology in EHRs for the LSDB to address the DQ issues. The development of the phases of HM took into account lessons learned from approaches in different HCOs. The ability of DQ in EHRs' advantage in HCOs to share and exchange health information can maximise the benefits that can provide quality healthcare service and accrue from



investment in eHealth. This interoperability capability requires the quality data of appropriate eHealth standards.

Different HCOs have, however, different data standards and quality assessments on the way to integrate EHR systems. Therefore, it is not easy to coherently manage data from the diverse heterogeneous source, especially for incompatible and inconsistent data. Figure 7.11 demonstrates the combined outcome of Chapter Three, Chapter Four, Chapter Five, Chapter Six and Chapter Seven, as follows:

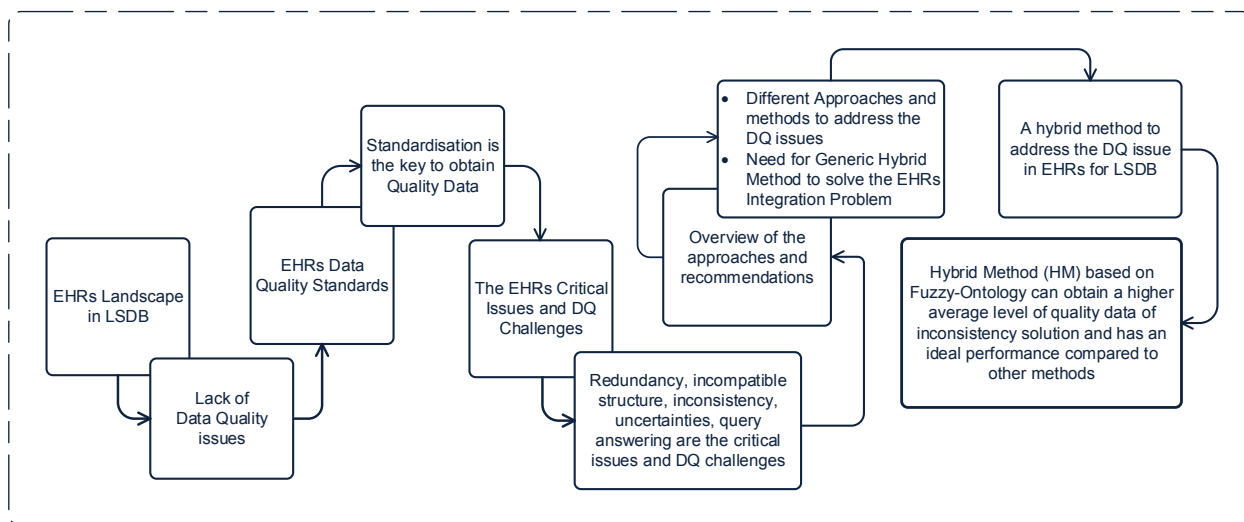
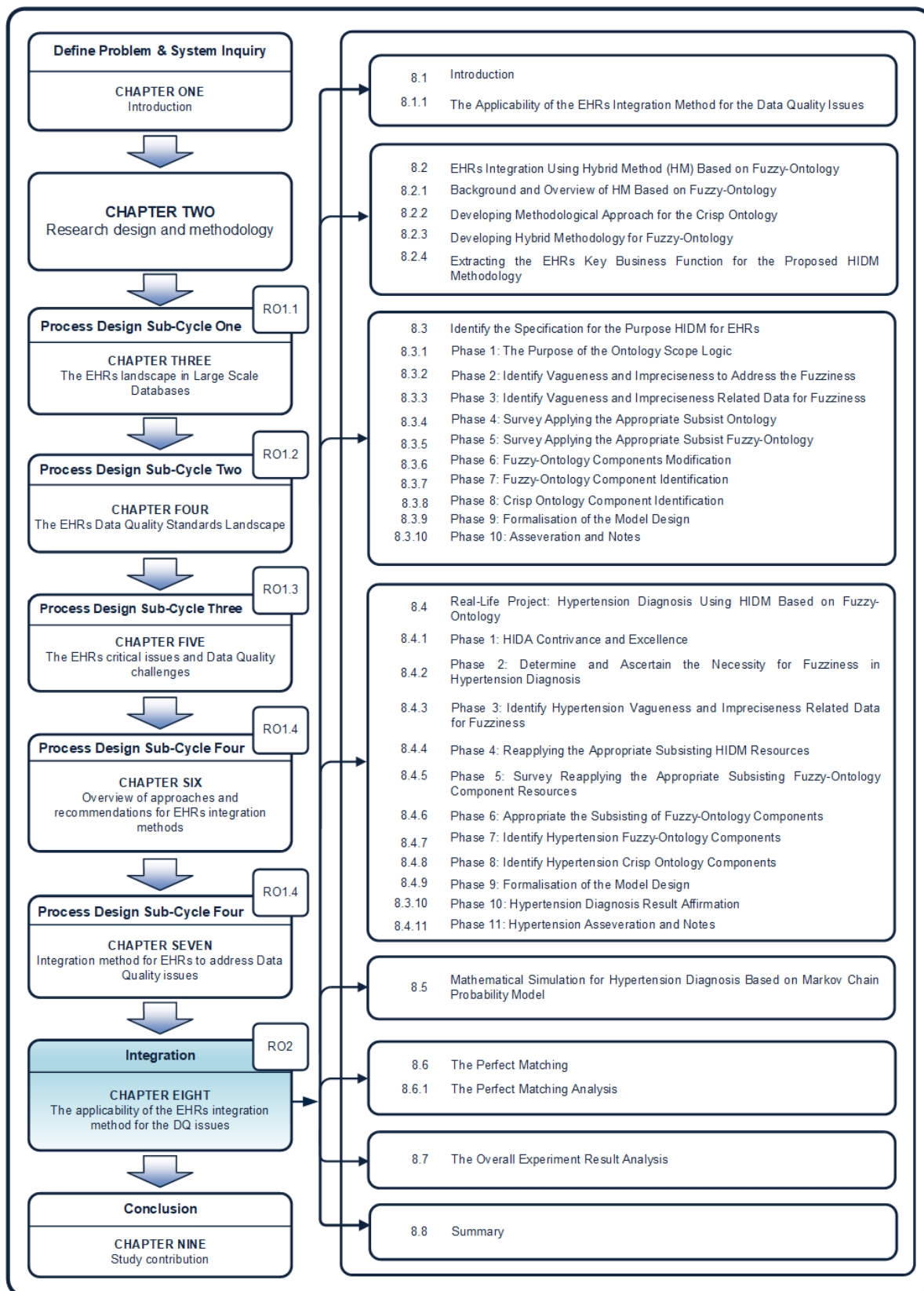


Figure 7.11: The combination outcome of Chapter Three, Chapter Four, Chapter Five, Chapter Six and Chapter Seven

The availability of a structured method to guide the data standard process could be highly valuable to developed countries. The proposed hybrid method based on fuzzy-ontology provides a solution aimed at addressing the DQ issues being faced in the healthcare domain. The method would be practically valuable to HCOs, where a lack of DQ is common. The applicability of the hybrid method based on fuzzy-ontology will be discussed in Chapter Eight.

## CHAPTER EIGHT: The applicability of the EHRs integration method for the DQ issues



Outline of the Chapter Eight

## CHAPTER EIGHT

### 8.1 Introduction

The first purpose of this chapter is to illustrate the applicability of HM based on fuzzy-ontology to address DQ issues in the LSDB, described in section 2.3.2.1 and highlighted in figure 8.1. This chapter focuses on the application of HM based on fuzzy-ontology intelligent systems to diagnosing hypertension. The applicability of HM is done in three stages, firstly to determine the usability of the proposed HM based on fuzzy-ontology (see 8.2). In section 8.3, the specification is identified for the proposed Hybrid Integration Development Methodology (HIDM) for EHRs integration.

This is followed by an overview in section 8.4 of the Real-World project: A Fuzzy Hypertension Specific Ontology. A mathematical simulation is performed using the Markov chain Probability model structure of hypertension progression risk in section 8.5. Section 8.6 performs a perfect matching modelling process using the dynamic Hungarian algorithm whereas, section 8.6.1 provides the perfect matching analysis. Section 8.7 provides the result of the analysis. A summary of Chapter Eight is provided in section 8.8. The result of the analysis is discussed in section 8.5.

HAs has attracted exciting research and commercial interest in the big data management fields. The term HAs refers to applications that offer complete full trust solutions through various combination approaches. HAs is considered an innovative heuristic solution package and essential mechanism that can tackle heterogeneous tasks deployed with multiple solutions in big DBMS. Today, the LSDB experience the massive voluminous datasets from every possible domain, such as Information Communication Terminology (ICT), EHRs, cloud computing, social media, chain management with the continuous applications of LAN, WAN, wireless, sensors, mobile and cloud computing technologies transmission.

Big data is defined as a so voluminous and complex, unstructured, semi-structured and structured data set that traditional data processing software applications are inadequate to deal with them. However, one of the main challenges in big data is the inherent difficulty to coherently manage incompatible and sometimes inconsistent data structures from DHS (John *et al.* 2012). Traditional approaches integrate well for comparatively small data domains but fail when they are being applied to unstructured, semi-structured and massive datasets. The efficient logical combination and existing approaches preserve the traditional approaches for the interoperability.

The HAs with its logical functions, including efficient sensitivity, can easily handle big data integration processes and with the efficient algorithmic logic addresses the DQ, data accuracy and inconsistency. Most of the existing approaches, such as peer-to-peer, data warehouses, middleware, data grid, data mining, semantic and ontology, actually establish semantic connections between heterogeneous data sources. Although these approaches offer advantages in some aspects, they do not provide a coherent mechanism to solve every DI problem. In addition, none of them pays strong attention to data inconsistency, which has been a long-standing challenge in database environments as stated in section 1.1. This implies the big data domain can show inconsistency because different DBMS have several standards and different major systems, which have emerged as critical issues and practical challenges (see section 1.1).

The challenges become even more important when implementing any big data domain, including an exponential increase of data, particular infrastructure need, a need for a skilled workforce, a need interoperable data standards, privacy and security and the need to include people, processes and policies, to ensure their adoption.

This implies the efficient approach to address this task is to format the original algorithm with respect to the big data domain, so as to integrate big data. Figure 8.1 describes the position of Chapter Eight in the design science research process used in this study, as follows:

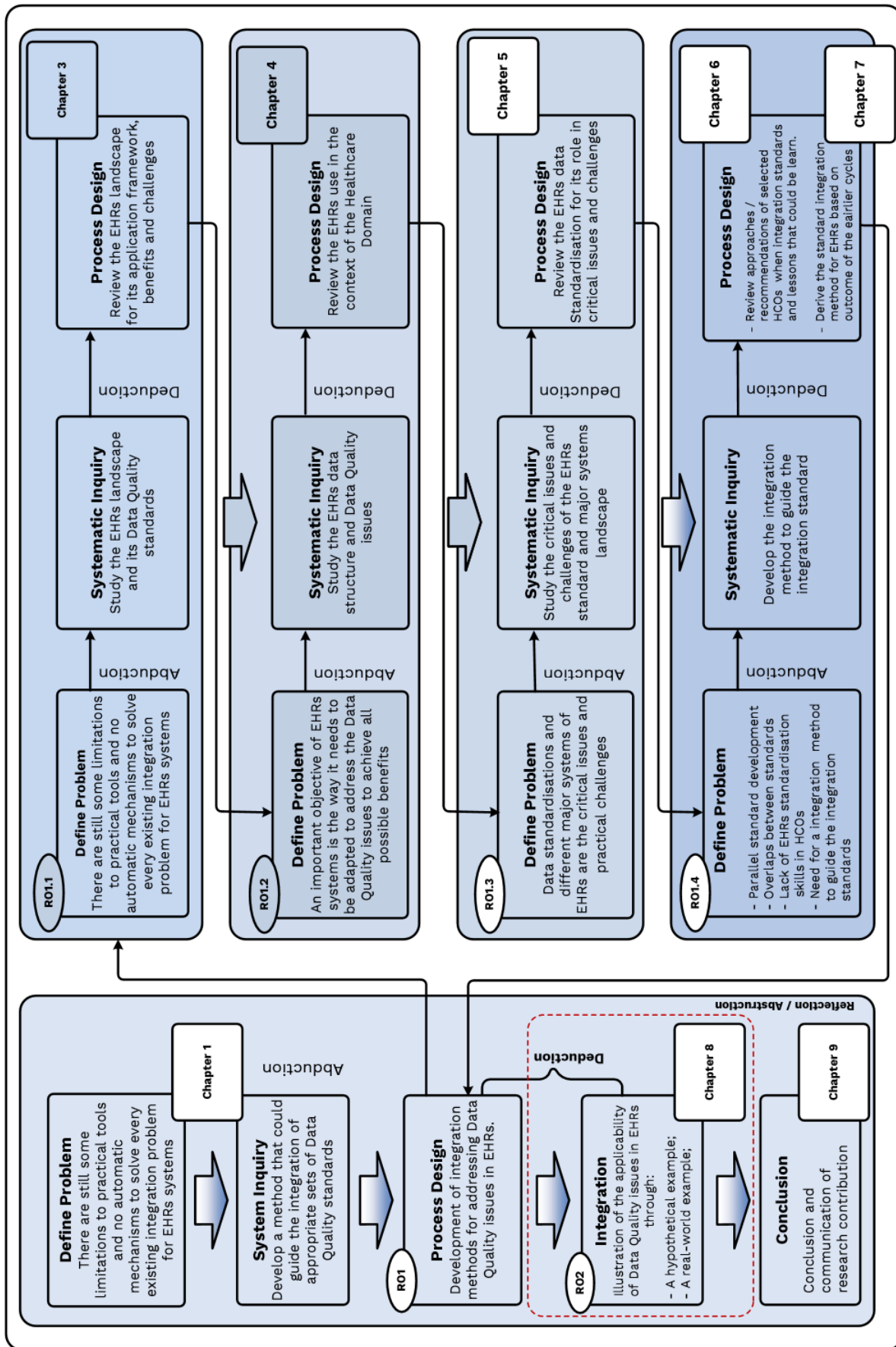


Figure 8.1: The position of Chapter 8 in the design science research process used in this study

As such, HAs that league dexterous performance along with the big data domain to preserve the generic data accuracy for the DI approach, is proposed in this chapter. In particular, this chapter focuses on developing the HAs based on fuzzy-ontology and performing a mathematical simulation and similarity measurement modelling process.

### **8.1.1 The applicability of the EHRs integration method for the Data Quality issues**

Information on EHRs defines it as the electronic version of health records that include patients' demographic data, registration details, measurements, translational research and decision-making support to provide quality healthcare services. To implement and obtain all these possible benefits, the EHRs are being promoted to drive health data, which must include an efficient knowledge base integration methodology for diverse health data sources and contexts.

Hybrid Approaches (HAs) based on fuzzy-ontology is one of the core functions to efficiently handle and process massive datasets from Diverse Heterogeneous Sources (DHS). HAs are becoming a noticeable trend in recent times due to its wide range of functionality to tackle all types of problem spaces. Therefore, big data communications are challenging the traditional approaches to satisfy the needs of the consumer, as data are often not capturing into the DBMS in a seasonably enough fashion to enable their use subsequently. In addition, big data plays a vital role in containing numerous treasures for all the fields in the DBMS. However, one of the main HAs challenges for the Big Data Integration (DI) systems is the inherent difficulty to coherently manage data from DHS, as different data sources have several standards and different major systems (see section 1.1). It is practically challenging to integrate diverse data into a global schema to attain what is looked forward to.

The challenges of HAs raise the need to find a better way to efficiently integrate voluminous data from DHS. To handle and align massive dataset efficiently, the

HAs algorithm with the logical combination of Fuzzy-Ontology along with the big data analysis platform has shown the results in term of improved accuracy. The proposed novel HAs will combine the promising features of Fuzzy-Ontology to search, extract, filter, clean and integrate data to ensure that users can coherently create new consistent datasets.

The development of the purposed HAs-based Fuzzy-Ontology for big data communication as proposed above takes into account the data inconsistency challenge that specific EHRs initiatives will play in addressing the DQ issues in the LSDB, the business interoperability requirements and its priorities. A mathematical modelling approach based on HAs-based Fuzzy-Ontology, which was initially explored in DI, is nominated.

This provides a strong theoretical and practical framework to work with heterogeneous, complex, conflicting and automatic consensus approaches for health DI. Fuzzy logic and the fuzzy set theory concept can be represented in the classic ontology to imprecise the expression concept and data relationship. Ontology is one of the effective feasible concepts and commonly used approach to efficiently conceptualise the semantic online applications. The hybrid combination of Fuzzy-Ontology can easily address the data uncertainty and reduce the data inconsistency.

Fuzzy logic is defined as a form of probabilistic approximate methodological logic to deal and define between the fixed and exact values. The interval of fuzzy logic variables ranges always between zero and one when the traditional approaches binary variable is always either zero or one. It is one of the biggest advantages of fuzzy logic that it can handle the concept of partial truth to extend information to deal with vagueness, where the concept range could be exactly false or exactly true. The concept of ontology is defined as to provide a generic and explicit requirement to represent information which becomes the most adapted domain today.

One of the major advantages of HA is the incremental academic interest to demonstrate the knowledge base applications, such as data scalability,

shareability, machine reliability, including readability and data stability. The classical ontology also refers to a crisp ontology that effectively deals with vague and imprecise information, which initially was undeniable on traditional ontology. To deal with vague data, data needs to be standardised to knowledge base representation value to structurally quantify and represent. The efficient combination of fuzzy logic and theory appropriate to deal with vagueness addresses the DQ in big data integration. The combination of two approaches, namely Fuzzy-Ontology introducing that functionality into crisp ontology was invented in the early 2000s. This concept summarised that encasing the fuzzy sets theory, Fuzzy-Ontology can associate the architect data, which has a vague and imprecise concept of truth degree, in other words, a belief of the world.

This chapter sought to design a HIDM based on Fuzzy-Ontology using the DSR methodology and performed a mathematical simulation and similarity measurement that will increase the quality of service over HCOs in an adaptive framework.

## **8.2 EHRs integration using Hybrid Method based on Fuzzy-Ontology**

This section describes the usability of the proposed HM based on Fuzzy-Ontology. This is a hypothetical project aimed at exploring the extent to which this method is more convenient to detect and address DQ issues in EHRs for the LSDB.

### **8.2.1 Background and overview of HM based on Fuzzy-Ontology**

The HIDM process is based on the use of a generic linguistic variable, which is used for fuzzifying and determining the ontologies. Generally, the selecting of this linguistic variable is problem-dependent. The proposed HM based on Fuzzy-Ontology alignment framework uses three core features, as follows:

- 1) The background knowledge;



- 2) The capacity to manage vagueness and impreciseness in the matching process;
- 3) Classifications in the resulting apprehension to improve the DQ in EHRs;

In the hybrid approach the first step, of each health domain apprehension is illustrated as a fuzzy set of reference apprehensions. In the next step, the fuzzified health domain apprehensions are assimilated to each other, resulting in fuzzy descriptions of the matches of the root apprehensions. Based on these apprehension matches, the HM propounds an algorithm that develops an absorbed Fuzzy-Ontology that captures that which is universal to the source ontologies. The considered HIDM methodology has been compared to a crisp ontology purpose and scope as well as Fuzzy-Ontology.

## 8.2.2 Developing a methodological approach for the crisp ontology

It is extensively recognised that no common single methodological approach exists that can solve every data integration problem (stated in section 1.1). According to the DQ goal to provide efficient guidelines to develop a crisp ontology, many different ontologies are presented as a guideline to develop these methodologies. While performing the design process, a formalisation for tasks scheduling has been provided, which was followed by ontological methodology. Different methodologies offer different flow processes, which could be worse or extremely worse according to the interoperability, rationality, competency or ease of use. The ontological methodology for efficient task scheduler can provide sufficient methodological support to the IT professional.

According to the literature review of this study, METHONTOLOGY is one of the most common renowned ontological methodology (Fernández *et al.* 1997) and others are: NeOn (Suarez-Figueroa 2010), DILIGENT (Vrandecic *et al.* 2005), On-To-Knowledge (Sure *et al.* 2004), HCOME (Kotis *et al.* 2006) and DOGMA (Jarrar *et al.*

2009). The study has found a very simple descriptive, yet guideline, to develop crisp ontology presented by Noy *et al.* (2001). Other guidelines are provided by Jarrar *et al.* (2009), the way to reapply the existing ontologies, the different development methodologies and describing their features.

To summarise a significant number of development methodologies, only a handy amount of methodologies is to be found. As there is a principal difference between Fuzzy-Ontologies and crisp ontology, it will not be possible to implement to develop Fuzzy-Ontology although it is a dedicated methodology for the development. This will mean that to implement Fuzzy-Ontologies, the additional fuzzy logic development process will consider an approximation for vagueness and conceptualisation for the fuzzified vagueness.

### 8.2.3 Developing hybrid methodology for Fuzzy-Ontology

The principal aims of the HIDM based on Fuzzy-Ontologies are to tackle the conceptualisation for the fuzzified vagueness and formalisms. In other words, the principal task is, the way to illustrate the fuzzified vagueness and formalisms into linguistic logics in HIDM. Using the Fuzzy-Ontologies in a standard and effective way is the main contribution to address the DQ issues in EHRs. An example, the IKARUS-Onto methodology (Alexopoulos *et al.* 2012), is a methodological logic for HIDM development. This methodology provides the guidelines to focus on the dispensation transformation from crisp ontology to fuzzy ones. This methodology consists of five generic steps, as follows:

- a)** Including acquiring crisp ontology;
- b)** Establishing the need for fuzziness;
- c)** Defining Fuzzy-Ontology components;
- d)** Formalising fuzzy components;
- e)** Validating Fuzzy-Ontology;

Another comprehensive guidelines presence for dispensation transformation from crisp ontology to fuzzy ones is IKARUS-Onto methodology. This methodology is

suitable even when crisp ontologies are present in the existence domain. The Fuzzy Ontomethodology as proposed by Ghorbel *et al.* (2010), has similar features to construct fuzzy augmentation based on the existing ontology to imprecise vagueness. The fuzzy Ontomethodology consists of three formal steps as follows:

- a)** Including the conceptualisation;
- b)** Providing the Ontologisation;
- c)** Providing the operationalisation;

In the real world, it is too difficult to understand as the processes grouped in each phase become too ambiguous. Initially, the fuzzy Ontomethodology has been developed to provide devote guidelines for ontology semantic web search. During the development and implementation process, the numerous amounts of Fuzzy-Ontology components will be defined and stored in the logic bank to reapply, which can enhance the interoperability and share ontology ability in the health domain to reduce the workload. To make the EHRs integration development process efficient is the main consideration and to reapply the stored Fuzzy-Ontology components from the logic bank.

The EHRs integration process should behave and guide in a formal way even if there were no existing Fuzzy-Ontology available to reapply from the logic bank while attempting to model the knowledge base in health domains. The presenting HIDM methodology will rely on the available crisp ontology. To best model and efficiently handle the imprecise and vague information, the ontology construct must follow methodological guidelines. To finalise this concept, this study chapter presents a HIDM development methodology, which could be helpful to improve existing Fuzzy-Ontology systems or develop a complete Fuzzy-Ontology integration system. This proposed HIDM methodology could well enable the interoperation of EHRs integration systems and resolve issues of inconsistent EHRs resolution in terms of generality, completeness, accuracy, reusability, efficiency and shareability.

## 8.2.4 Extracting the EHRs key business function for the proposed HIDM methodology

Based on ontology development methodologies, this section presents a formal Fuzzy-Ontology development paradigm. Its emphasis lies on introducing new changes brought by Fuzzy-Ontologies into the development process. The principal emphasis is to develop a different way with the new addition to Fuzzy-Ontologies for efficient EHRs integration. All new additions are the prior potential knowledge from real-life experiences and principles of crisp and Fuzzy-Ontology. Figure 8.2 (Saiod *et al.* 2019a) demonstrates the inputs inspiring to conceive the HIDM, as follows:

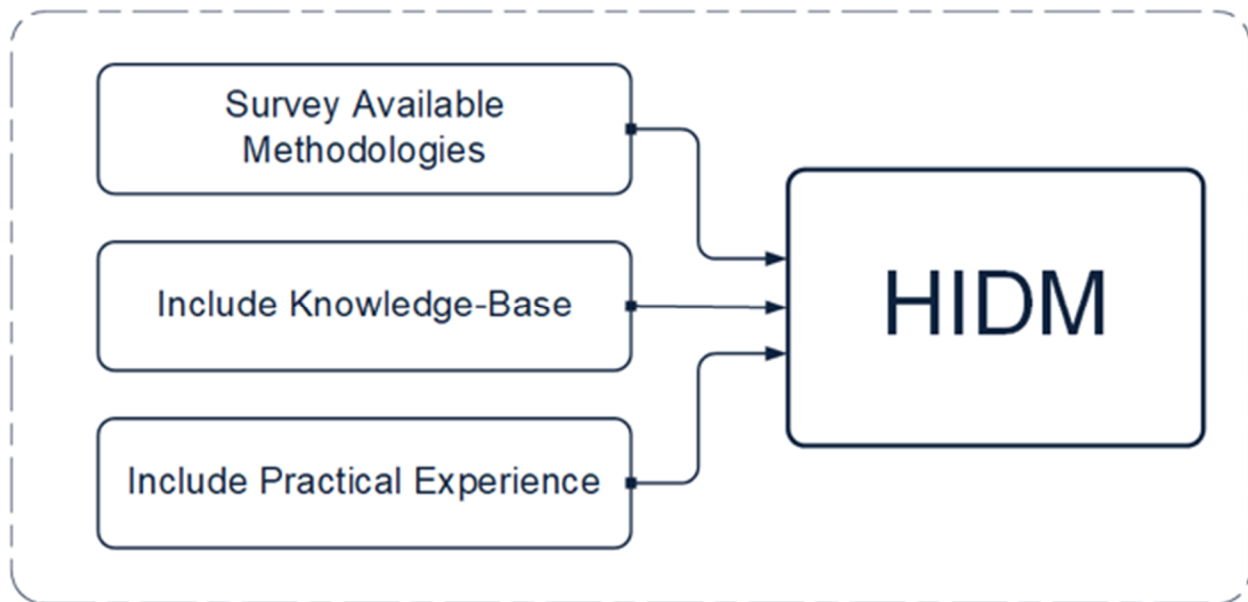


Figure 8.2: Inputs inspiring to conceive the HIDM (Saiod *et al.* 2019a)

The HIDM does not aim to complete the reform of existing development methodologies; it focuses on partial changes with new additions which were found as the weak considerations of the existing methodologies. As shown in figure 8.2, the proposed HIDM is based on the three basic principal resources for EHRs to address DQ issue for LSDB, as follows:

- 1) Surveying available methodologies for the ontology construction as an addition;

- 2) Including the knowledge base on exploring and developing HIDM development;
- 3) Including practical experience from subsisting methodologies;

Figure 8.3 describes the complete HIDM structure based on Fuzzy-Ontology for EHRs integration systems, as follows:

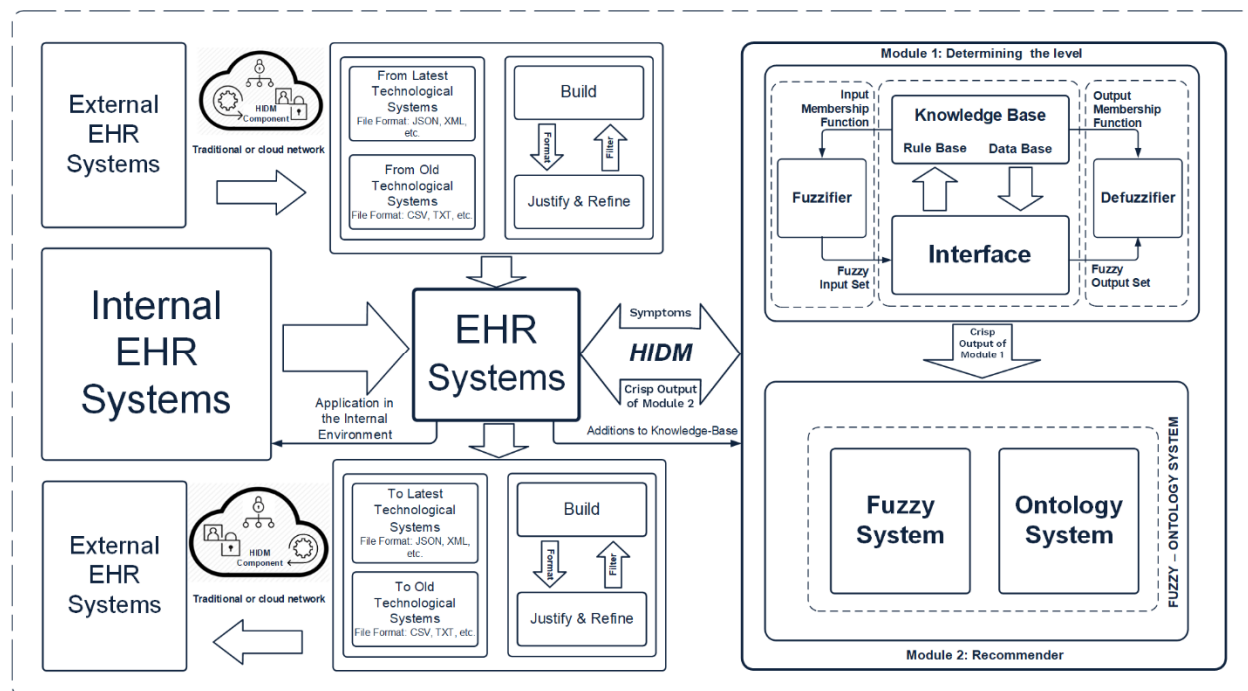


Figure 8.3: The complete HIDM structure based on Fuzzy-Ontology for EHRs integration systems (Saiod *et al.* 2019a and 2019b)

This complete HIDM architecture is based on the knowledge from real-life experiences and aspires to address DQ issues in EHRs for the LSDb, as follows:

- a) *Surveying available methodologies for the ontology construction as an addition:* Generally, because of the methodological nature of Fuzzy-Ontology, it is impossible to reconstruct the entire development process. The partial changes in the existing crisp Fuzzy-Ontology development flow control should be conventional, instead of a complete redevelopment. The partial changes will be strictly introduced by conventional Fuzzy-Ontology methodologies, only with additional convention considerations. These partial

additions have been elected as the initial goal of the proposed HIDM development. Each evaluation consideration, which adds newer ways fares better or worse regarding the development methodology in terms of progress evaluation (for example, such consideration for reapplying existing ontologies from the stores logic bank). This study and research comprehensively studied and analysed several methodologies, such as Methontology and NeON. This consideration has added the strengths in the new addition that have been applied in the HIDM methodologies and guaranteed that the addition is implemented accurately. In addition, some other methodologies have been also considered in the proposed HIDM, such as IKARUS-Onto methodology and the fuzzy Ontomethodology.

***b) Including the knowledge base on exploring and developing HIDM***

***development:*** The practical experience and investigation showed that the Fuzzy-Ontology model design has different preferences and views according to the different professionals based on practical incidents. This study has abstracted this real-life knowledge as an initial phase. To materialise the formal flow process, this initial phase has been considered as the base of the methodological foundation.

***c) Including practical experience from subsisting methodologies:***

Most of the Fuzzy-Ontology software and tools particularly refer to the editor from the existing components or from the archive. Tho *et al.* (2006), have referred to automatic Fuzzy-Ontology generation in the Fuzzy-Ontology Generation Framework (FOGA). Bobillo *et al.* (2011) have mentioned the easy and visualised way to plug into Fuzzy OWL2, which defines fuzzy related knowledgebase with OWL2 annotation. While abstracting the different Fuzzy-Ontology development tools, it has been identified that the Fuzzy OWL2 obtains conceptual design implemented by another developer. This means that the Fuzzy-Ontology development tools could be applied in an informal way in the default specified development process.

### 8.3 Identify the specification for the purpose of HIDM for EHRs

The goal of this section is to provide a general abstraction of the HIDM development process, which have been followed throughout the construction and implementation process. The proposed HIDM is the completely new methodological approach for hypertension diagnosis and has not just been converted into fuzzy from the available crisp ontology. Figure 8.4 (Saïod *et al.* 2019a) describes the structure of the proposed HM for EHR systems based on Fuzzy-Ontology, as follows:

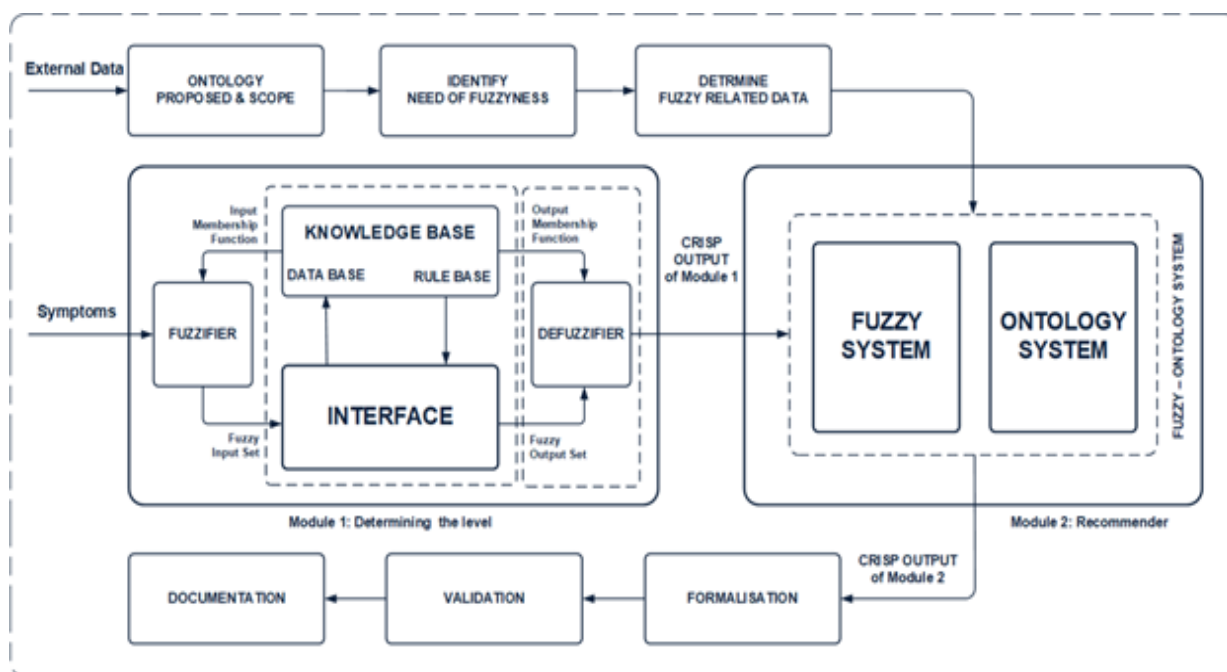


Figure 8.4: The structure of the proposed HM for EHR systems based on Fuzzy-Ontology (Saïod *et al.* 2019a)

The entire workflow of the proposed HIDM has been presented in figure 8.3. These entire HIDM process flow controls have 11 grouped phases in the Fuzzy-Ontology development. Each grouped phase has its own associated reason and process, which have been described in the subsections below, as follows:

### 8.3.1 Phase One: The purpose of the ontology scope logic

Methodology's primary aim, such as in any other majority development methodologies, is to identify the motivation for Fuzzy-Ontology development. This means that each and every Fuzzy-Ontology proposed methodological logic and scope must be clearly identified at the initial stage. All the motivational questions should be answered to define the motivation clear enough. Five questions have been raised at the initial stage as the motivation for the HIDM development that needs to be addressed:

- 1) What is the scope of the proposed HIDM development that needs to be constructed?
- 2) Which methodology has the best logic to tackle for given tasks across the available solutions?
- 3) What is the domain type that the process will be developed for?
- 4) How to compile the stakeholder list and distribute the play-roles across them?
- 5) How to guarantee the inoperable development with tight collaboration across all stakeholders?

All five questions described above, have been accurately addressed during the HIDM proposal and scope process to establish the methodology motivation. During the development process, this motivation might not be 100% identical to the motivation is not stated as an identical goal. But, it is clear that a crisp ontology will be developed for the real-life project called hypertension diagnosis ontology for the healthcare domain.

### 8.3.2 Phase Two: Identify vagueness and impreciseness to address the fuzziness

The HM based on Fuzzy-Ontologies has been evaluated to efficiently handle vague and imprecise data. This second phase is to determine and identify whether



fuzziness should be introduced in the development process. If so, then the ultimate task is to define the kinds of crisp ontology or Fuzzy-Ontology that need to be developed. In this phase, the IT professional with all the stakeholders will cooperate with each other to determine the kind of fuzzified incidents. This will help IT professionals to consider in the development process the type of fuzziness and where they need to be implemented. Three processes have been considered to have accurate answers:

- *Firstly:* A proper ontological scope analysis needs to be performed. The first determination should be analysed deeply to understand the entire process and to model the process design as well as where exactly crisp ontology and Fuzzy-Ontology need to be used. All possible process nodes where vagueness and impreciseness could be present, have to be identified.
- *Secondly:* The IT professional must investigate the entire development process to identify and determine if the Fuzzy-Ontology needs to be a consideration in the modelling process. Crisp ontology is widely used in diverse healthcare domains where vague and imprecise data is common. Therefore, this is important to be identified before introducing the Fuzzy-Ontology, as all data in health domains are assumed to be accurate and uncertainty inherent to information is neglected. The Fuzzy-Ontology is feasible to tackle and manage the vague and imprecise data whereas the ontology is not capable to deal with them. It is, however, very complicated to create a comprehensive Fuzzy-Ontology as it is very difficult to set the degree to balance between them. Before the degree of the schematic ontology will be set, the Fuzzy-ontology needs to be justified by all stakeholders.
- *Thirdly:* Due to the Fuzzy-Ontology definition and the needs of different ontology components the fuzziness could be presented. Different types of fuzzy components could be present in fuzzy individual instantiating concepts and fuzzy blurry relations.

It should be sufficient to obtain a rough grasp and the determination of fuzzy specific components does not always need to be exhaustive. Finally, the necessity of fuzziness would be identified including generic specific fuzzy components underlying in the schematic ontology.

### **8.3.3 Phase Three: Identifying vagueness and impreciseness related data for fuzziness**

Since this study is to investigate and address DQ issues in EHRs for the LSDB, the identification of vagueness and impreciseness in related data for fuzziness falls in the study consideration. The initial consideration of the HIDM developments in phase two is also true, which forces to introduce the fuzziness in the development process. Hence, the identification of the fuzzy related component and development of the logic to tackle them is the main goal of phase three.

During the development, the process will approve logic components and will be stored in the logic bank DB to reapply them into a similar case to process with the best performance and reduce the process time. Time is indeed one of the key factors in the decision-making process to save a patient's life. Phase two provides a detailed discussion to achieve the clearing of the vague component existing in the healthcare domain.

The following phase will determine to a greater extent which of these components have real vague meaning. The precise and vague components need to be grouped into established valuable inputs for further identification. According to the knowledge base that has been obtained in this section, the health domain could be divided into two groups, as follows:

- 1) Fuzzy appropriate components;
- 2) Fuzzy associated components;

Only with a clear understanding of fuzzy differentiation, the IT specialist can provide appropriate logic to handle fuzzy appropriate and fuzzy associated components in a well-defined convention.

#### **8.3.4 Phase Four: Survey applying the appropriate subsist ontology**

The main task of this phase is to identify and reapply appropriate ontology from the stored logic bank. Reapplying the available and verified resources provided lots of advantages in ontology development. Two main advantages exist when reapplying the appropriate existing ontologies, which are divided into two nodes:

- 1) Decreasing the domain workload and reducing the process time;
- 2) Enabling interoperability and providing a comprehensive process with other similar EHR systems;

The existing ontology components can refer to the crisp ontology and Fuzzy-Ontology at the same time. For the best performance, the fuzzy-Ontology will be the first consideration when reapplying takes place. Only a few existing Fuzzy-Ontologies exist and it is very hard to native each other. According to our knowledge, no available published Fuzzy-Ontology DB exists. Before the deployment to the destination domain, the existing Fuzzy-Ontology must be verified by the IT professionals and stakeholders.

The approximate and vague data could be inherited in Fuzzy-Ontology components. Crisp ontology components are considered as destination relevant modelling structures and could be useful when these are defined in existing Fuzzy-Ontologies. This concept increases the probability to reapply ontological components in the extension of Fuzzy-Ontology introduction. This way maximises the re-using of the existing ontology components in the development.

Reapplying the ontology components will totally depend on the fuzziness of the existing ontology that has been elected in the methodologies. To integrate the

existing ontology components into the intended ontology, different complications should be predefined for efficient integration.

### **8.3.5 Phase Five: Survey applying the appropriate subsist Fuzzy-Ontology**

The goal of phase five is to investigate whether recherché ontology components from subsisting ontology are fuzzy, as phase four already have clarified the survey deployment of the subsist ontology. If so, the survey ontology must be analysed for the introduction to the development model and this potential component stored for further use. The analyses have found three different kinds of ontology components:

- 1) Components are only Ontology;
- 2) Components are only Fuzzy-Ontology;
- 3) Components are both ontology and Fuzzy-Ontology;

The schematic ontology is capable to handle the combination components all three together when reapplying them. Phase seven will only be considered if the subsistence ontology will be determined in a crisp ontology component. In the meantime, impreciseness and vagueness might already have determined similar characteristics in the survey Fuzzy-Ontology components. Therefore, the potential survey Fuzzy-Ontology component will be selected and reapplied in the schematic ontology. If so, then phase six will be considered for detail investigations to make all possible modifications of these Fuzzy-Ontology components. If the scenario will not be clear enough, then both phase six and seven should be considered in the development process.

### **8.3.6 Phase Six: Fuzzy-Ontology components modification**

As discussed in phase five, phase six will be in development consideration only if similar characteristics are determined in the survey Fuzzy-Ontology components

or if they are not clear enough. Therefore, phase six will take into account the modification of the Fuzzy-Ontology modification that has been inherited from the survey Fuzzy-Ontology. The survey Fuzzy-Ontology components might not always determine the perfect match in the integration process for the target health domain QA requirements. Therefore, the modification remains the essential process to refine to accommodate them for interoperability. For example, in schematic ontology, the optimal blood pressure will be considered and reapplied over and over as survey Fuzzy-Ontology components for a hypertension diagnosis.

However, the association to a mismatch in the fuzzy components may happen when the left shoulder membership function (0,300,180,110) range for optimal blood pressure for hypertension diagnosis is determined by the IT professional as, **blood pressure level Systolic  $\leq$  120 and Diastolic  $\leq$  80 pressure range**. According to the IT professional, however, the accurate fuzzy set value (0,300,120,80) could reapply as optimal blood pressure. However, different Fuzzy-Ontology components can be stored and reapply optimal value in a different solution according to the needs. Therefore, it will be using full to select the most suitable components for the vague health data (for example, **blood pressure Systolic  $\leq$  120 and Diastolic  $\leq$  80 as optimal blood pressure**).

All vague hypertension diagnoses of the optimal blood pressure could be captured and diagnosed using the available Fuzzy-Ontology components. This is followed as optimal Blood Pressure (BP) as described above. The optimal BP (0,300,120,80) could be used as a modifier function and apply the modifier to limit the property (isRangeDefineAs) across hypertension concept and BP. Therefore, the IT professional can easily express the vague BP for hypertension (isRangeDefineAs, blood pressure) diagnosis. This section investigated the identification of the suitable Fuzzy-Ontology components and their availability in the logic store and setting the perfect approximation to the data to be accurate in the existing health domain or applications.

### 8.3.7 Phase Seven: Fuzzy-Ontology Component Identification

A comprehensive description of the difference between crisp and fuzzy data have been provided in phase three, which could be used as a valuable input in this phase. This phase aims to identify different Fuzzy-Ontology components that will provide an accurate approximation for the vague and imprecise data. This phase will require an exact collaboration among the IT and health specialist. The health specialist will provide and set accurate quantification for the vague and imprecise data according to the experience and statistics, whereas the IT professional will set the degree and model this information to convert for the membership functions.

This procedure is needed to set the degree of vagueness as a development model to precise data associated with the particular health domain or scope. Therefore, the IT specialist is playing a vital role in this phase. Although the vague and imprecise data might be a very small amount in the health domain according to the scope, it will play a key role in the entire development process to efficiently integrate the health data. All precise and vague data will be investigated and accurately addressed according to fuzzy components within the entire Fuzzy-Ontology development, in this phase.

### 8.3.8 Phase Eight: Crisp ontology component identification

Dealing with certain comprehensive knowledge is the main focus of this phase. The knowledge base is defined to identify different Fuzzy related components related to their specific attributes. This phase describes the crisp model development process of the ontology conventional methodology. According to Natalya *et al.* (2001), important historical data must be organised in the same way as the first step in this phase. Three knowledge-based approaches to crisp ontology development of the model process, are as follows:

- 1) *Top to bottom*: Using the most comprehensive adumbration at the beginning and enunciating the concept during the development;
- 2) *Bottom to top*: Identifying the most comprehensive adumbration and enunciating them to the top level;
- 3) *Association*: Associating both the “Top to Bottom” and “Bottom to Top” models together;

The relationship could be identified in several links in the development model concept. This phase also considers complementary crisp ontology components (for example, data properties, fundamental truth and data pattern). This phase must have identified all Fuzzy-Ontology components to complete the design model concept.

### 8.3.9 Phase Nine: Formalisation of the model design

To implement the HIDM some specific programming language needs to be selected to formalise the design model to machine-readable language. The traditional language is not capable to represent vagueness and imprecision defined in the Fuzzy-Ontology concept. Several languages are available to express the Fuzzy-Ontology concept. An example is the extended RDF that has been designed to express real number intervals  $([0,1])$ , to represent the specific degree (for example, subject, object and term).

The formalisation of the Fuzzy-Ontology components to standard formalisation should be capable of assimilating in the DLs of fuzzy set extensions. A well-defined Fuzzy-Ontologies formalisation has been presented by Bobillo *et al.* (2011) using the OWL2 methodological concept. A Fuzzy-Ontology rule formalisation has been developed by Fudholi *et al.* (2009) in SWRL. The advantage of SWRL is that it is easy to implement and increase the number of rules, but it considerably limits the scalability.

Different languages have different characteristics and capabilities to Fuzzy-Ontology formalisation and the concept may vary from one another. However,

there is no common methodology to formalise the Fuzzy-Ontology in a single development language, because the different languages have different properties and use different concepts to formalise the Fuzzy-Ontology components. According to Stoilos *et al.* (2006), the fuzzy data type cannot be formalised using Fuzzy Description Logic (FDL) f-SHIN but can be easily formalised by OWL2. Therefore, a certain language should be selected according to the needs, so that the formalism language should be capable to handle the requirements.

### **8.3.10 Phase Ten: Asseveration and notes**

This phase is to introduce the asseveration of the development principle. The asseveration includes the comprehensive guidelines for the maintenance, model architecture details, methodology details and notes and including different concept details. The comprehensive open access guideline should provide an easy and understandable process to follow so that non-professionals can use it into their own scenario by using the guideline. The guideline should be useful in terms of inoperability in similar tasks. In addition, the guideline could be revised and updated according to the beneficial feedback from users.

## **8.4 Real-life project: Hypertension diagnosis using HIDM based on Fuzzy-Ontology**

This section presents a real-life project to diagnosis a simple use case for hypertension to demonstrate the applicability and interoperability of the proposed HIDM based on Fuzzy-Ontology. The hypertension diagnosis can obtain information about optimal, normal, high normal, stage one (mid), stage two (moderate) and stage three (severe) blood pressure ranges as a description of the destination health domain. This real-life project has considered several input parameters to diagnosis hypertension, such as systolic blood pressure, diastolic blood pressure, age and BMI into the HIDM and *hypertension risk* is the contribution parameter.



The aim of generating hypertension diagnosis is to characterise different types of risk levels so that the providers can obtain a better diagnosis of their patient to make a better decision. The hypertension risk level, which is going to be the diagnosis, should be clearly identified as a certain risk level. Also, the stage of the hypertension risk level is the aim. Therefore, a generic linguistic variable has been introduced to measure the risk level associated with the application and hypertension diagnosis to obtain a generic comprehension.

*N.B: The modelling domain is limited to only hypertension classifications instead of all types of EHRs.*

Hypertension is defined as blood pressure enhanced to unhealthy and dangerous levels. The hypertension measurement process considers how rapidly the blood flows across the veins and how many resistances are received while the heart is pumping. Hypertension depends on many different factors and even varies between males and females. Cois *et al.* (2014) define that, “Although hypertension has been associated with factors such as alcohol consumption, smoking, high body mass index and inadequate exercise, research has suggested that the degree of association between socioeconomic status and hypertension varies between males and females”.

According to Steyn (2006), “Data on hypertension prevalence for the country are available from the 1998 South African DHS, which show a prevalence of 21% for both males and females using the 140/90 mmHg threshold”. Amy *et al.* (2016) state that, “the higher prevalence rates were predominately in the North West and Northern Cape provinces, as well as in the Eastern Cape. Only one district in KZN had a hypertension prevalence rate of  $\geq 30.8$  for example,  $\geq$  the highest quantile). Four of the six districts in the Western Cape had prevalence rates of  $\geq 37.4$  for example,  $\geq$  the highest quantile). Natal (KZN) district had a prevalence rate  $\geq$  the highest quantile”. These patterns with the socioeconomic disadvantage spatial pattern.

**Data collection:** The following sub-phase formalised in the proposed HIDM, a Fuzzy-Ontology hypertension risk level, is going to be developed. This study has

collected the data of 3000 patients with hypertension provided by different hospitals, clinics and a medical aid company in South Africa and measured the results that were in the hypertension optimal range level, which have been predefined by the health professionals under NDA. Table 8.1 (Saiod *et al.* 2019a) describes details according to various BMI category ranges about adult South African males' (18-60 years) distribution blood pressure (systolic and diastolic), as follows:

Table: 8.1 (Saiod *et al.* 2019a): According to various BMI category ranges about adult South African males' (18-60 years) distribution blood pressure (systolic and diastolic)

Blood Pressure	Male				Female			
	SBP (mmHg)		DBP (mmHg)		SBP (mmHg)		DBP (mmHg)	
	Percentage	Number	Percentage	Number	Percentage	Number	Percentage	Number
Normal	17.25	250	23.12	272	60.25	750	45.05	584
Pre-hypertension	81.2	950	49.45	584	38.25	256	45.24	560
Hypertension	2.56	300	30.01	352	2.49	320	8.97	120

Where, SBP and Gender chi-square = 49.44\*\*\* ( $P < 0.001$ ), DBP and Gender chi-square = 23.23\*\*\* ( $p < 0.001$ )

The subset of data collection been divided into two groups, namely male and female. All patients are aged 18 to 60 years. Additional information such as BMI, heart rate, working background, medical history and lifestyle have been considered in the apportioned questionnaire to achieve abundant knowledge on the diagnosis. To collect the BMI information traditional scales and tools were used to measure the patient weight and height. BMI is defined as each patient's individual body weight divided by the patient's height squared. This is the universal method and widely used globally. The unit of the measure considered as an international standard is  $\frac{Kg}{m^2}$ . Table 8.2 (Saiod *et al.* 2019a) describes details according to various

BMI category ranges about adult South African females' (18-60 years) distribution blood pressure (systolic and diastolic), as follows:

Table: 8.2 (Saïod *et al.* 2019a): According to various BMI category ranges about adult South African females' (18-60 years) distribution blood pressure (systolic and diastolic)

BMI	SBP				DBP			
	Normal	Pre-hypertension	Stage 1	Stage 2	Normal	Pre-hypertension	Stage 1	Stage 2
Underweight	73.1	26.9	0	0	65.4	26.9	7.7	0
Normal	54.4	43.4	0	2.2	45.6	41.2	13.2	0
Overweight	23.6	73.6	0.9	1.8	17.3	52.7	27.3	2.7

Where, SBP Chi-square = 43.24\*\*\* ( $P < 0.001$ ) and DBP Chi-square = 45.44\*\*\* ( $P < 0.001$ )

Table 8.3 (Saïod *et al.* 2019a) describes details about adult South African males' (18-60 years) distribution blood pressure (systolic and diastolic) where BMI is the risk factor, as follows:

Table: 8.3: Adult South African males' (18-60 years) distribution blood pressure "systolic and diastolic" where BMI is the risk factor (Saïod *et al.* 2019a)

BMI Classification	Odds ratio SBP			Odds ratio DBP		
	Pre-hypertension	Stage 1	Stage 2	Pre-hypertension	Stage 1	Stage 2
Underweight	0.45	0.59	1.15	0.53	0.42	0.53
Normal	NC	NC	NC	NC	NC	NC
Overweight	1.77		1.48	2.82	2.65	8.98

Where, NC = Reference Category/Normal Category

Three different BMI category ranges were considered, as follows:

- 1) *First category range*: From 19 to 25 are considered a healthy weight;
- 2) *Second category range*: From 16 to 18 are considered as underweight;
- 3) *Third category range*: From 25 to 30 are considered as overweight;

Table 8.4 (Saiod *et al.* 2019a) describes details about adult South African females' (18-60 years) distribution blood pressure (systolic and diastolic) where BMI is the risk factor, as follows:

Table: 8.4: Adult South African females' (18-60 years) distribution blood pressure “systolic and diastolic” where BMI is the risk factor (Saiod *et al.* 2019a)

BMI Classification	Odds ratio SBP			Odds ratio DBP		
	Pre-hypertension	Stage 1	Stage 2	Pre-hypertension	Stage 1	Stage 2
Underweight	0.56	0.61		0.54	0.63	1.1
Normal	NC	NC	NC	NC	NC	NC
Overweight	0.56	2.28	0.98	2.75	3.74	8.8

Where, NC = Reference Category/Normal Category

### 8.4.1 Phase One: HIDA contrivance and excellence

The aim of this section is to identify the objective description of the hypertension diagnosis. As the section's motivation, the different characteristics of hypertension risk levels need to provide the diagnosis semantically, so that the provider can obtain the hypertension risk level, including the alert performance. To ontology description as the target domain, the dictation is to formularise into different properties of a hypertension risk level to demonstrate a semantically efficient diagnosis for clinicians and providers. This approach will also enable its use as a hypertension alert execution treatment service process and can be carried out into a machine learning language database to be re-used in knowledge-based

approaches. For an efficient indication, a list of indicators has been delineated. Answers of these indications have been used to delaminate the proposed HIDM contrivance and excellence in the close-bodied process, as follows:

*1) Question:* What types of diagnosis contribution is expected from the HIDM?

*Answer:* The risk level for the hypertension diagnosis need to be measured at different SBP (*Systolic blood pressure*), DBP (*Diastolic blood pressure*). Note: Age, BMI and lifestyle are major factors which affect the blood pressure.

*2) Question:* Are the HIDM been used as an appropriate solution technology upon other approaches, for example, as principle solution?

*Answer:* Ontology is one of the most generic architecting mechanisms to introduce hypertension specific diagnosis because interoperability between different hypertension diagnoses and risk levels are anticipated and formalised easily in the linguistic variable expression. The HIDM is dictated to achieve efficient hypertension diagnosis by using the formularised linguistic variable expression for specific hypertension and measure its interoperability.

*3) Question:* What architect ramifications will be the desired HIDM?

*Answer:* Due to the motivated contrivances, the aim of the desired HIDM architecture will be limited to specific hypertension diagnosis only, instead of complete EHR systems. The hypertension diagnosis ramification model contrivance logic will store into the machine learning technology database to be re-used or inherited to similar related hypertension diagnosis applications. In this way, schematic hypertension diagnosis will be characterised as a domain-specific HIDM.

*4) Question:* How to list all stakeholders and how to distribute their roles in the development process?

*Answer:* HIDM professionals will be the main stakeholders, whereas clinicians and providers will share their knowledge and patients will be the participants as data sources.

**5) Question:** How to ensure successful HIDM development and implementation when different types of stakeholders are engaged in the development process?

**Answer:** The HIDM development process will be controlled, managed and maintained by only HIDM professionals and they will collect knowledge and data from secondary (clinicians and providers) and third-party (patients) stakeholders. The conceptualisation knowledge will be used to build the model framework architecture and data will be used to test the approach. All stakeholders will actively engage during the development process, such as to refine and correct to ensure the interoperability.

All indications and answers described above declared that the HIDM development process will follow the specific HIDM model architecture and will use the appropriate hypertension data.

#### **8.4.2 Phase Two: Determining and ascertaining the necessity for Fuzziness in hypertension diagnosis**

Fuzzy-Ontology is defined as to handle vagueness, uncertainty, inaccuracy and imprecision. The objective of this section is to determine and ascertain the necessity for fuzziness in hypertension diagnosis. The eventual destination of this section is to ascertain the kind of approach needed to build, namely Fuzzy-Ontology or crisp ontology. In this step, the primary and secondary stakeholders should decide to ascertain if any needs of fuzziness exist in the hypertension diagnosis. To delaminate the proper decision, a list of operations need to be performed:

- **Understanding the data:** The need to provide a proper investigation over the target domain. First of all, raw hypertension data needs to be analysed to gain a complete understanding of the data and all the type of incidents

and to ascertain if any needs of fuzziness exist. Determining and noting if any vague or similar incident presents.

- *Implementing the Fuzziness where needed:* The secondary stakeholders will decide whether fuzziness will be included in the HIDM architecture, before the appearance of the Fuzzy-Ontology approaches. The crisp ontology is capable to handle general vague data and is widely implemented. But, crisp ontology often failed with complex vague data and in the matter of inaccurate, uncertain and inconsistent data. The crisp ontology is unable to deal with a high expectation of the degree of vagueness and its complexity. But Fuzzy-Ontology is capable to deal with these incidents where crisp ontology fails. Therefore, the secondary stakeholder justifies the incident and asks if the HIDM expert should extend the fuzziness.
- *Determining the type of Fuzziness:* According to the Fuzzy-Ontology definition, the fuzziness might already present in the ontology elements. Therefore, it is important to determine if any fuzziness exists and if so which kind of fuzziness in individual or blurry concept.

The determination of each and the specific fuzzy element may not possible but should be enough in the fuzzy set. This issue partially may be solved when re-using the Fuzzy-Ontology. Once the necessity of fuzziness can be determined and ascertained, logic needs to be implemented to obtain a special type of fuzzy element that exists in the semantic ontology. In EHRs as the target domain, two types of important features in the hypertension diagnosis are involved in modelling data:

- 1) Diagnosis;
- 2) Measure;

Therefore, all certain elements could be handled by the crisp ontology. But, the patient wants to hear the hypertension level in the linguistic specification, rather than numeric quantification. As linguistic specification often exposes vagueness, it

is not easy to map efficiently hypertension explicit numeric measure to linguistic specification. For instance, the hypertension diagnosis level is determined on three levels, namely mid, moderate and severe. The interval between the different levels is blurry and could be overlapped to the other level. According to this investigation, to manage and handle this blurry vagueness implicit to the patient's age and BMI, fuzziness needs to be applied. This analysis summarised that Fuzzy-Ontology will be more efficiently manage and handle the hypertension diagnosis domain than the crisp ontology along.

### 8.4.3 Phase Three: Identifying hypertension vagueness and impreciseness related data for Fuzziness

As Fuzzy-Ontology is associated with the model architecture for the proposed hypertension diagnosis in the EHRs domain, it is essential to establish solid cooperation between HIDM professionals and healthcare providers. Section two described that to specify the fuzzy associated elements, these need to be established by the healthcare professionals. To provide an accurate hypertension diagnosis and establish the boundary between fuzzy associated elements and particular data, HIDM professionals must collect the wisdom and recommendation from a healthcare professional, as follows:

- ***Appropriating the accurate data:*** The patient's raw data, such as various ages and BMI, can be apparently specified by the appropriate hypertension risk level. Those characteristics are directly affecting the patient's blood pressure. Different levels of hypertension diagnoses with numerical representation can be calculated apparently illustrating the patient's age and BMI category.
- ***Fuzzy associated components:*** The linguistic representation of hypertension diagnosis, such as *mid*, *moderate* and *severe*, contain vague meanings. This is because age and BMI could be described as *severe* to some extent, whereas it could also be labelled as *moderate* with a probability. The definitions for



linguistic classifications for the level of hypertension diagnosis should be fuzzified to meet the EHRs domain needs.

The knowledge base of the EHRs domain is accurately divided into two parts, as follows:

- a)* Precise information;
- b)* Fuzzy-related information.

Afterwards, they can be modelled with different treatments and healthcare services, respectively.

#### **8.4.4 Phase Four: Reapplying the appropriate subsisting HIDM resources**

The aim of the EHRs domain and excellence is the desired Fuzzy-Ontology, subsisting HIDM resources. It is concerning to check resources in online projects and publications to re-use them. Resources on not only the fuzzy elements or crisp ontology could be found, but also on Fuzzy-Ontology to re-use them in the potential approach. After the profound analysis and search for the resources in online projects and open publications, some elements have been identified to re-use them, such as hypertension, hypertension levels, and hypertension specifications and levels. An approach called Clinical Decision Support Systems (CDSSs) was found, that is built to assist healthcare professionals to diagnose their patients and relief the provider with some support together with the diagnosis.

The CDSSs ontology comprises a classification of conceptual wisdom, including relying on practical experiences in the case and rule base logic accordingly. Several model functions and classes for hypertension diagnosis involve inappropriate implementation as a part of the diagnosis.

### 8.4.5 Phase Five: Survey reapplying the appropriate subsisting Fuzzy-Ontology component resources

Since the CDSSs ontology selected as the ontology candidate to be re-used from phase four is a crisp ontology, then a conclusion can be drawn in this phase that only crisp ontology elements could be re-used. Specifically, CDSSs are tools, which are designed to assist clinicians for better clinical decision-making with knowledge and relevant clinical data of the patient, intelligently filtered and presented, to enhance health and healthcare ontology, which are selected to be reapplied. These are as follows:

- a) Category specification:* Representing appropriate hypertension risk level and status, when this specific data is captured. For hypertension, diagnosis specification could represent the level of hypertension of the particular patient.
- b) Diagnosis:* Defined as the category specification was diagnosed in the specific occurrence of hypertension.
- c) Contribution:* Defined as the specification of the latest result or diagnosis that could be achieved from the proposed approach. The result will be the patient hypertension status after diagnosis.

### 8.4.6 Phase Six: Appropriating the subsisting Fuzzy-Ontology components

As any subsisting of hypertension Fuzzy-Ontology components will not be reapplied, this section has been skipped.

### 8.4.7 Phase Seven: Identifying hypertension Fuzzy-Ontology components

This section is to represent, the way to identify appropriate and various Fuzzy-Ontology components to illustrate uncertain and vague data. All hypertension

related numeric representation will here be converted to linguistic specification. The determination of fuzzy data type of the blood pressure level and its breakdown that ensures the fuzzy description logic, are shown in table 8.5 (Saïod *et al.* 2019a), as follows:

Table: 8.5: Determination of fuzzy data type of the blood pressure level and its breakdown that ensures the fuzzy description logic (Saïod *et al.* 2018)

Category	Systolic		Diastolic
Hypertension	< 90 mmHg	And/or	< 60 mmHg
Optimal	< 120 mmHg	and	< 80 mmHg
Normal	120-129 mmHg	And/or	80-84 mmHg
High normal	130-139 mmHg	And/or	85-89 mmHg
Level one (Mid)	140-159 mmHg	And/or	90-99 mmHg
Level two (Moderate)	160-179 mmHg	And/or	100-109 mmHg
Level three (Severe)	$\geq 180$ mmHg	And/or	$\geq 110$ mmHg
Isolated systolic hypertension	$\geq 140$ mmHg	And	< 90 mmHg

Fuzzy data types defined in the fuzzy hypertension specific ontology are presented in table 8.6 (Saïod *et al.* 2019a), as follows:

Table: 8.6: Fuzzy data types determined and identified in the hypertension fuzzy description logic (Saïod *et al.* 2019a)

Fuzzy data type	Definition	Vague data modelled
MidSystolic = LeftShoulderSystolic (0,200,140,159) MidDiastolic = LeftShoulderDiastolic (0,120,90,99)	Denoting that the numeric value of the Mid should comply with leftShoulder membership function leftShoulderSystolic (0,200,140,159) and leftShoulderDiastolic (0,120,90,99)	Hypertension with its Systolic value range from 0-159 mmHg and Diastolic 0-99 mmHg could be regarded as small to some value. The value distribution

		complies with a leftShoulder membership function.
ModerateStageSystolic = TrapezoidalSystolic (130,139,160,179) ModerateStageDiastolic = TrapezoidalDiastolic (90,99,100,109)	Denoting that the numeric value of the Moderate should comply with trapezoidal membership function trapezoidalSystolic (130,139,160,179) and trapezoidalDiastolic (90,99,100,109)	Hypertension with its Systolic value range from 130-179 mmHg and Diastolic 90-109 mmHg could be regarded as small to some value. The value distribution complies with a trapezoidal membership function.
SevereSystolic = RightShoulderSystolic (160,179,180,200) SevereDiastolic = RightShoulderDiastolic (100,109,110,120)	Denoting that the numeric value of the Severe should comply with RightShoulder membership function RightShoulderSystolic (160,179,180,200) and RightShoulderDiastolic (100,109,110,120)	Hypertension with its Systolic value range from 160-200 mmHg and Diastolic 100-120 mmHg could be regarded as small to some value. The value distribution complies with a RightShoulder membership function.

Specifications for Fuzzy concepts defined in the Fuzzy hypertension specific ontology are also presented in table 8.3 (Saïod *et al.* 2019a), as follows:

Table: 8.7: Determination of appropriate fuzzy concepts in the hypertension specific diagnosis (Saïod *et al.* 2019a)

Fuzzy concept	Definition	Vague information modelled
HypertensionStage	Representing the super class of a set of sub-concepts, including MidStage, ModerateStage and SevereStage	Hypertension stage could be described by linguistic variables, such as mid-stage, moderate stage and severe stage
MidStage	Containing a collection of diagnosis whose level is assigned with the MidStage Fuzzy data type.  <i>MidStage =</i> <i>∃ hasNumericValue.MidLevel</i>	Hypertension diagnosis Mid stage ranging from Systolic: 140-159 mmHg and Diastolic: 90-99 mmHg is classified as mid complying with a leftshoulder membership function.

ModerateStage	<p>Containing a collection of diagnosis whose level is assigned with ModerateStage Fussy data type.</p> <p><i>ModerateStage</i> =  <math>\exists</math> <i>hasNumericValue.ModerateLevel</i></p>	<p>Hypertension diagnosis Moderate stage ranging Systolic: <math>\geq 180</math> mmHg and Diastolic: 100-109 mmHg is classified as moderate complying with a trapezoidal membership function.</p>
SevereStage	<p>Containing a collection of whose level is assigned with SevereStage Fussy data type. diagnosis</p> <p><i>SevereStage</i> =  <math>\exists</math> <i>hasNumericValue.SevereLevel</i></p>	<p>Hypertension diagnosis Severe stage ranging from Systolic: 160-179 mmHg and Diastolic: <math>\geq 110</math> mmHg is classified as severe complying with a rightshoulder membership function.</p>

The Fuzzy Description Logics (FuzzyDL) syntaxes are followed by the fuzzy concept expressions (Straccia 2013). The fuzzy data type is defined as the essential goal to standardise the fuzzy concept data and is provided in corresponding order. For example, in fuzzy data, a numeric representation date must be changed to linguistic representation, such as stage one hypertension diagnosis and also generate a probabilistic linguistic specification of the hypertension risk level, such as mid-stage.

It is imprecise as the fuzzy data characteristic of the crisp set has Numeric Value, and this must convert into the linguistic specification to specify the relationship between the different concepts, such as Mid, Moderate and Severe fuzzy data prefixes, for example as Mid-stage, Moderate-stage and Severe-stage.

Tables 8.5 and 8.6 provide detailed information by healthcare providers on the vague and imprecise boundary level among mid, moderate and severe hypertension stages, applying three different fuzzy value sets, defined as membership functions. Figure 8.4 demonstrates how the fuzzy data type represents Fuzzy-Ontology hypertension diagnosis in a certain specification.

### 8.4.8 Phase Eight: Identifying hypertension crisp ontology components

This phase is to define the model architecture of appropriate crisp ontology elements in the hypertension diagnosis domain. To avoid data inconsistency, the ontology precise information should be considered to create new crisp ontology elements from the traditional subsisting ontology. The appropriate Fuzzy-Ontology elements were already identified in step seven. Fuzzy input sets, encased in the Fuzzy hypertension specific Ontology to describe fuzzy data types, are shown in figure 8.5 (Saïod *et al.* 2019a), as follows:

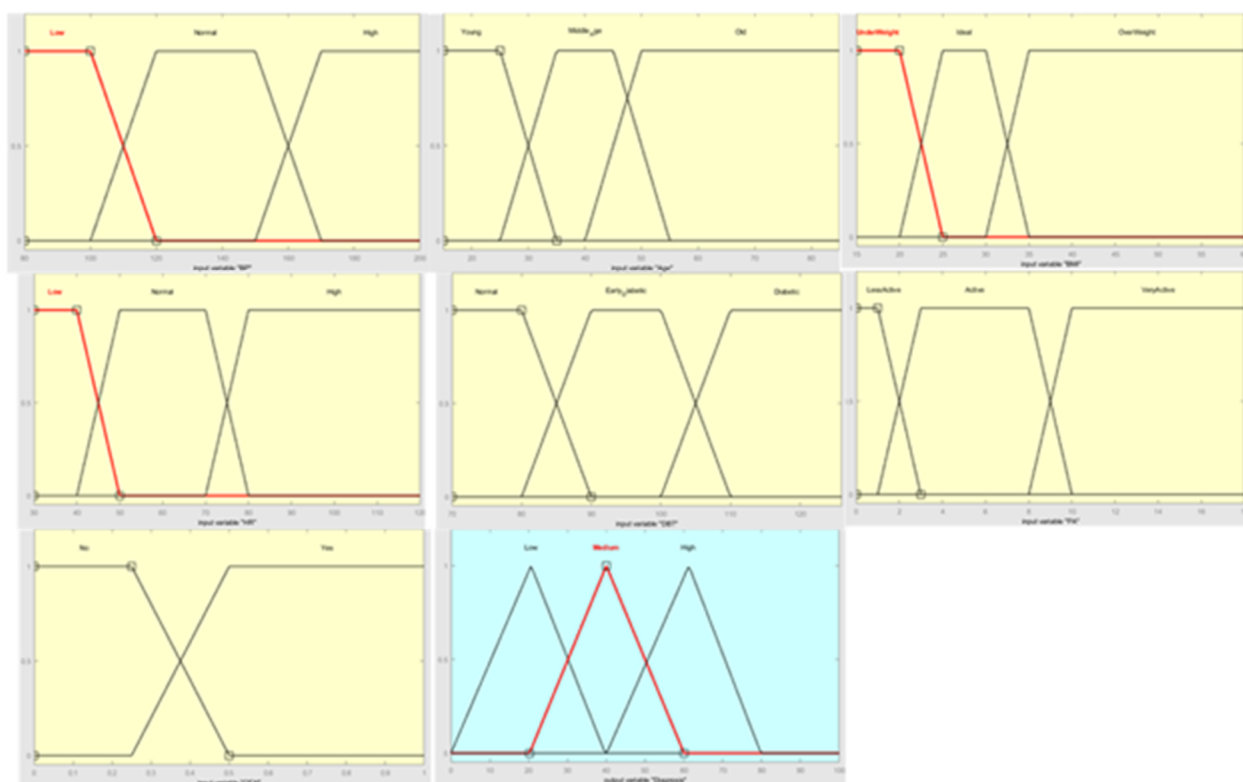


Figure 8.5: Fuzzy data types for hypertension diagnosis (Saïod *et al.* 2019a)

Table 8.8 (Saïod *et al.* 2019a) demonstrates the crisp ontology concept, data characteristics, and proposals, supplied by the healthcare professional, as follows:

Table: 8.8: The appropriate crisp ontology logics in the hypertension diagnosis  
(Saïod *et al.* 2019a)

Crisp Concept	Definition	Certain health data modelled
<i>HypertensionStage</i>	Defining a superclass of different stages of hypertension.	A specific age and BMI could be identified as a specific type. The type of a specific stage of hypertension is a significant feature for hypertension diagnosis to be considered during the study.
<i>Normal</i>	Representing Normal which generally the systolic blood pressure between $\leq 120$ mmHg and the diastolic blood pressure between $\leq 80$ mmHg.	Health expert thinks that the recognition of normal stage as normal and important to hypertension diagnosis.
<i>Hypertension</i>	Representing hypertension which generally the systolic blood pressure between 140-179 mmHg and the diastolic blood pressure between 90-109 mmHg.	Health expert thinks that the recognition of hypertension stage as Hypertension and important to hypertension diagnosis for the treatment.
<i>Hypertensive</i>	Representing hypertensive which generally the systolic blood pressure $\leq 180$ mmHg and the diastolic blood pressure $\leq 110$ mmHg.	The provider considers the fact of the hypertension stage as a Severe stage.

Table 8.8 is different from traditional hypertension diagnosis. Three extra patterns are associated with hypertension Fuzzy-Ontology because traditional hypertension ontology is inappropriate to deal with all the requirements for the healthcare domain. The argument behind the age, BMI and unknown, is that the healthcare professional has defined that three kinds of extra diagnoses are essential, namely normal, hypertension and hypertensive. They are knapping between each other in

all kinds of hypertension diagnoses. Table 8.9 (Saïod *et al.* 2019a) describes the description of the hypertension-related dataset, as follows:

Table 8.9: The description of the hypertension-related dataset (Saïod *et al.* 2019a)

No	Feature	Description	Data Type	Domain	Number of Entry
1	Age	Patient Age in year	Numeric	[15-85]	
2	Sex	Gender	Binary	[0,1]	
3	BPS	Blood Pressure Systolic	Numeric	[0-200]	35
4	BPD	Blood Pressure Diastolic	Numeric	[0-122]	35
5	BMI	Body Mass Index	Numeric	[0-67]	11

Table 8.10 (Saïod *et al.* 2019a) describes the hypertension Fuzzy dataset corresponding to feature with the numerical presentation, as follows:

Table 8.10: The hypertension Fuzzy dataset corresponding to feature with the numerical presentation (Saïod *et al.* 2019a)

No	Feature	Fuzzy Set	Data Intervals			
1	Age	Young	15	15	25	30
		Middle_Age	25	35	45	55
		Old	45	55	85	85
2	BP (Blood Pressure Systolic/Diastolic)	Low	80/20	80/20	100/55	120/65
		Normal	100/55	120/65	150/85	170/90
		High	150/85	170/90	200/120	200/120
3	BMI (Body Mass Index)	Under_weight	15	15	20	25
		Ideal	20	25	30	35



		Over_weight	30	35	40	40
4	HR (Heart Rate)	Low	30	30	40	50
		Normal	40	50	70	80
		High	70	80	100	100
5	DBT (Diabetic)	Normal	70	70	80	90
		Early_Diabetic	80	90	100	110
		Diabetic	100	110	120	120
6	PA (Physical Activity)	Less_Active	0.5	0.5	1	3
		Active	1	3	8	10
		Very_Active	8	10	16	16
7	GEN (Genetic)	No	0	0	0.25	0.5
		Yes	0.25	0.5	1	1

#### 8.4.9 Phase Nine: Formalisation of the model design

OWL2 is chosen to be used for this case approach. It is the formalism language to illustrate the ontology design architecture. The fuzzy OWL2 scope and ontology performer portage have been deployed to easily convert the theoretical concept into OWL2 supported signification, in this section. The ontology performer portage is used for visualised structures and realises comfortably for the modelled hypertension specific Fuzzy-Ontology. OWL or RDF is sufficient enough for the ontology performer portage and automatic code creation ontology. Many OWL files are approachable for online constructed hypertension specific Fuzzy-Ontology diagnosis. Figure 8.6 (Saïod *et al.* 2019a) demonstrates the overall visualisation of the structure of the hypertension specific Fuzzy-Ontology diagnosis, as follows:

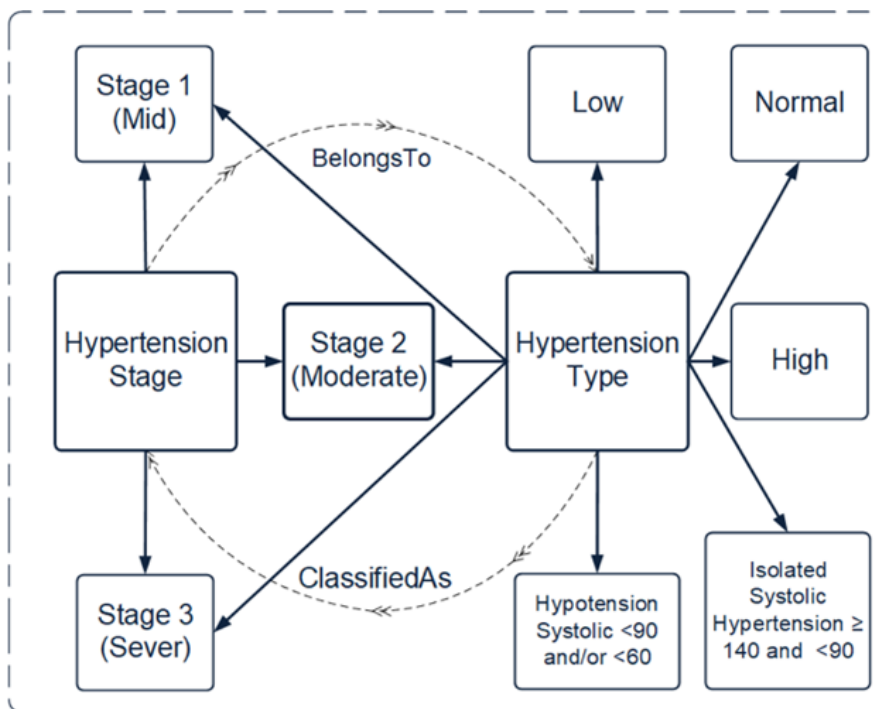


Figure 8.6: The overall visualised the structure of the Fuzzy Hypertension specific Ontology (Saiod *et al.* 2019a)

#### 8.4.10 Phase Ten: Hypertension diagnosis result affirmation

The affirmation is defined as the evidence of the approach appropriateness of the constructed approach consequences, in this case, the consequences of the HIDM. The appropriateness of the data consistency and accuracy characterisation are appraised by the fuzzy description logic reasoner. Other appropriateness of the HIDM consequences is subjectively verified by all stockholders who have been directly engaged in the approach evaluation procedure. The affirmation consequences are provided as follows:

- a) Data accuracy affirmation:* The approached HIDM hypertension diagnosis have been accurately illustrated and architected from the EHRs domain. Data have been collected from healthcare professionals and have followed accurately all actual instruction, including recommendations, to keep the correct boundary line between specific data and uncertain vague data. All

subsisting uncertain elements have been identified accurately to present accurately appropriate Fuzzy-Ontology elements, amalgamated by the appropriate fuzzy data set. In addition, the correct relationship has been established between Fuzzy-Ontology and crisp ontology elements, defined by the HIDM expert and healthcare professionals.

- b) *Data inconsistency affirmation:*** The fuzzy description logic reasoner has been introduced to define and verify the hypertension diagnosis for data consistency. And ontology has been introduced to observe the HIDM data architecture and data elements. No debatable concept and data exist in the EHRs domain for the hypertension diagnosis development process, as verified by the healthcare professional.
- c) *Data completeness affirmation:*** The hypertension diagnosis model has satisfied the entire specification of the HIDM contrivance and excellence, including the complete knowledgebase concept that has been specified in step one. Especially, all imprecise and vague data has been integrated and converted into the hypertension diagnosis representation.
- d) *Data logicalness affirmation:*** All controversial boundary lines between the specific and uncertain data are logically rationally identified by a healthcare professional. Different imprecise and vague data types and categories that have been operated by the fuzzy approximation concept have made sense to other HIDM experts and healthcare professionals.
- e) *Data reliability affirmation:*** The developed HIDM hypertension diagnosis is easily understandable and reliable for all stakeholders. The RHRs domain terminologies that have been identified in the HIDM are specified clear enough for the reliability.
- f) *Data validity affirmation:*** According to all stakeholders' numerous investigations and verifications, the EHRs domain terminology is enough to declare an accurate and consistent hypertension diagnosis. No data redundancy has been noticed or identified in the HIDM architecture or model.

### 8.4.11 Phase Eleven: Hypertension asseveration and notes

This section to introduce hypertension diagnosis based on HIDM is excluded as the main contribution approach fails out as the main focus of this study.

## 8.5 Mathematical simulation for hypertension diagnosis based on the Markov Chain Probability Model

This section presented a mathematical simulation for Hypertension Diagnosis Based on the Markov Chain Probability Model (MCPM). The MCPM is defined as a stochastic process, which has strong progressive characteristics and the future evolution relies only on the current circumstance situation. Figure 8.7 (Saïod *et al.* 2019a) shows the Markov Chain with respect to the different hypertension probabilities, as follows:

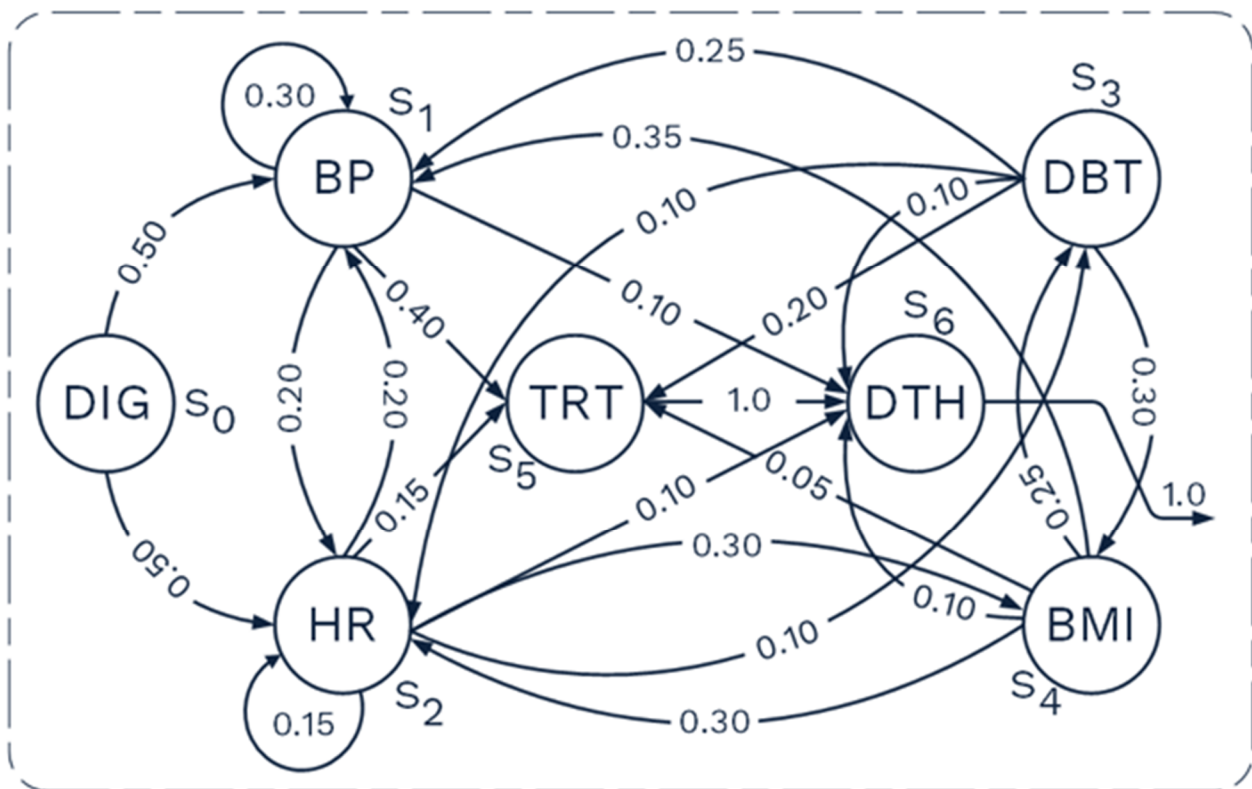


Figure 8.7 – The Markov Chain Probability link structure of the hypertension progression risk model (Saïod *et al.* 2019a)

Where,

- DIG – Start Diagnosis/Monitoring;
- BP – Blood Pressure;

<i>HR</i>	– Heart Rate;
<i>DBT</i>	– Diabetic;
<i>BMI</i>	– Body Mass Index;
<i>D</i>	– Death/Extremely Critical;
<i>EXT</i>	– Exit from Diagnosis/Monitoring;

In other words, the Markov process can be explained as it is a stochastic process by its present position. That is, the distinctions of events are independent of the history of the system. It means that the statement of the current circumstance completely captures all data, which could possibly impact the future enlargement of the procedure. The stochastic procedure defines all circumstantial conversion as probabilistic. At each phase, the procedure probably alters its circumstance from the present circumstance to another circumstance or remains in the same circumstance associated with the probability assignment.

Here the interval when each symptom reflects the visualisation of the condition is included. The probabilistic conversations and transitions are composed in the set of Markov state calculation from the original position to the final position. The visual probabilistic conversations and transitions comply with two aspects, detailed below:

- 1) ***Transitive aspect:*** The transitive aspect defines how a symptom passes from one condition to another condition. A mathematical simulation includes a conditional graph, where the other is describing the matrix probabilities passes will be provided. The multiple probabilistic conversations and transitions in a single time are infinitesimals of the upper sequence and possibly disregarded;
- 2) ***Temporary aspect:*** The temporary aspect is defined as the interval needed for the requirements on passes from one condition to the others. In other words, the probability of conversations or transitions in time  $\Delta t$  from one position to another position is conferred by acquiring into  $\Delta t$ ;

The interval is estimated with the point of views condition and referred to as transformation probabilities or simply perceived, the probabilities. The transformation probability represents by itself the average interval probabilities, passes from the condition of end diagnosing/monitoring of the condition and

before reception by them to the next probability condition during the same interval. The Markov model can be illustrated by means of a position conversation and transition diagram, which demonstrate the entire position and transition probabilities. The Markov chain model describes, as follows:

The hypertension diagnosis has a set of positions, which can be represented as  $S = \{S_0, S_1, S_2, \dots, S_r\}$ . The procedure begins in one of these positions and transfers coherently from one position to another position. Every transition is mentioned as a step. If the chain is present in the position  $a S_i$ , then it transfers to the position  $S_{aj}$  at the subsequent position with a probability defined by  $S_{ij}$  and this current position probability does not relay on the previous position. Therefore, the transition probability defines as  $S_{ij}$ . The transition procedure possibly remains in the same position and this happens with probability  $S_{ij}$ . The original probability allocation determined on S and describes the beginning position. Basically, this determination of the commencement position is committed in a specific position. Using the transition matrix S, the hypertension probability could be represented as  $S_{00}, \dots, S_{06}$ . The probability of the other two occasions could be represented as the position of starts of S. Generally, when the Markov chain has r position, then:

$$S_{06}^{(2)} = \sum_{k=0}^r S_{ik} S_{kj} \quad (8.1)$$

Therefore, the above description and induction can be easily implemented using the below proposition. The original probability allocation determined the set of positions. A Markov chain hypertension progression risk model can be defined as follows:

$$S_0(k+1) = 0 + 0 + 0 + 0 + 0 + 0 + 0; \quad (8.2)$$

$$S_1(k+1) = 0.5 \times S_0(k) + 0.3 \times S_1(k) + 0.2 \times S_2(k) + 0.25 \times S_3(k) + 0.35 \times S_4(k) + 0 + 0; \quad (8.3)$$

$$S_2(k+1) = 0.5 \times S_0(k) + 0.2 \times S_1(k) + 0.15 \times S_2(k) + 0.10 \times S_3(k) + 0.3 \times S_4(k) + 0 + 0; \quad (8.4)$$

$$S_3(k+1) = 0 + 0 + 0.1 \times S_2(k) + 0 + 0.25 \times S_4(k) + 0 + 0; \quad (8.5)$$

$$S_4(k+1) = 0 + 0 + 0.3 \times S_2(k) + 0.3 \times S_3(k) + 0 + 0 + 0; \quad (8.6)$$

$$S_5(k+1) = 0 + 0.4 \times S_1(k) + 0.15 \times S_2(k) + 0.2 \times S_3(k) + 0.05 \times S_4(k) + 0 + 0; \quad (8.7)$$

$$S_6(k+1) = 0 + 0.1 \times S_1(k) + 0.1 \times S_2(k) + 0.2 \times S_3(k) + 0.1 \times S_4(k) + 1.0 \times S_5(k) + 0; \quad (8.8)$$

So,

$$(S_0(k+1), S_1(k+1), S_2(k+1), S_3(k+1), S_4(k+1), S_5(k+1), S_6(k+1)) =$$

$$[S_0(k), S_1(k), S_2(k), S_3(k), S_4(k), S_5(k), S_6(k)] \times \begin{vmatrix} 0 & 0.5 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0.3 & 0.2 & 0 & 0 & 0.4 & 0.1 \\ 0 & 0.2 & 0.15 & 0.1 & 0.3 & 0.15 & 0.1 \\ 0 & 0.25 & 0.1 & 0 & 0.3 & 0.2 & 0.1 \\ 0 & 0.35 & 0.3 & 0.25 & 0 & 0.05 & 0.1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

When,  $k = 0$ ;

$$S_0 = 1;$$

$$S_1 = S_2 = S_3 = S_4 = S_5 = S_6 = 0;$$

When,  $k = 1$ ;

$$S_0 = 0;$$

$$S_1 = 0.5 \times S_0(k_0) = 0.5;$$

$$S_2 = 0.5 \times S_0(k_0) = 0.5;$$

$$S_3 = S_4 = S_5 = S_6 = 0;$$

When,  $k = 2$ ;

$$S_0 = 0;$$

$$S_1 = 0.3 \times S_1(k_1) + 0.2 \times S_2(k_1) = 0.25;$$

$$S_2 = 0.2 \times S_1(k_1) + 0.15 \times S_2(k_1) = 0.18;$$

$$S_3 = S_4 = S_5 = 0;$$

$$S_6 = 0.1 \times S_1(k_1) + 0.1 \times S_2(k_1) = 0.1;$$

When,  $k = 3$ ;

$$S_0 = 0;$$

$$S_1 = 0.25 \times S_1(k_2) + 0.2 \times S_2(k_2) = 0.1;$$

$$S_2 = 0.2 \times S_1(k_2) + 0.1 \times S_2(k_2) = 0.07;$$

$$S_3 = 0.1 \times S_2(k_2) = 0.02;$$

$$S_4 = S_5 = 0;$$

$$S_6 = 0.1 \times S_1(k_2) + 0.1 \times S_2(k_2) + 0.1 \times S_3(k_2) = 0.05;$$

When,  $k = 4$ ;

$$S_0 = 0;$$

$$S_1 = 0.35 \times S_1(k_3) + 0.2 \times S_2(k_3) = 0.05;$$

$$S_2 = 0.2 \times S_1(k_3) + 0.3 \times S_2(k_3) = 0.04;$$

$$S_3 = 0.1 \times S_2(k_3) + 0.25 \times S_4(k_3) = 0.01;$$

$$S_4 = 0.3 \times S_2(k_3) + 0.3 \times S_3(k_3) = 0.02;$$

$$S_5 = 0;$$



$$S_6 = 0.1 \times S_1(k_3) + 0.1 \times S_2(k_3) + 0.1 \times S_3(k_3) + 0.1 \times S_6(k_3) = 0.02;$$

When,  $k = 5$ ;

$$S_0 = 0;$$

$$S_1 = 0.3 \times S_1(k_4) + 0.2 \times S_2(k_4) + 0.25 \times S_3(k_4) + 0.35 \times S_4(k_4) = 0.06;$$

$$S_2 = 0.2 \times S_1(k_4) + 0.15 \times S_2(k_4) + 0.1 \times S_3(k_4) + 0.3 \times S_4(k_4) = 0.02;$$

$$S_3 = 0.1 \times S_2(k_4) + 0.25 \times S_4(k_4) = 0.01;$$

$$S_4 = 0.3 \times S_2(k_4) + 0.3 \times S_3(k_4) = 0.01;$$

$$S_5 = 0.4 \times S_1(k_4) + 0.15 \times S_2(k_4) + 0.2 \times S_3(k_4) + 0.05 \times S_4(k_4) = 0.03;$$

$$S_6 = 0.1 \times S_1(k_4) + 0.1 \times S_2(k_4) + 0.1 \times S_3(k_4) + 0.1 \times S_4(k_4) + 0.1 \times S_5(k_4) = 0.01;$$

Figure 8.8 (Saïod *et al.* 2019a) demonstrates the graphical representation of the Markov chain probability model link structure of hypertension progression risk when “BMI to BP = 0.35 and BMI to HR = 0.30”, as follows:

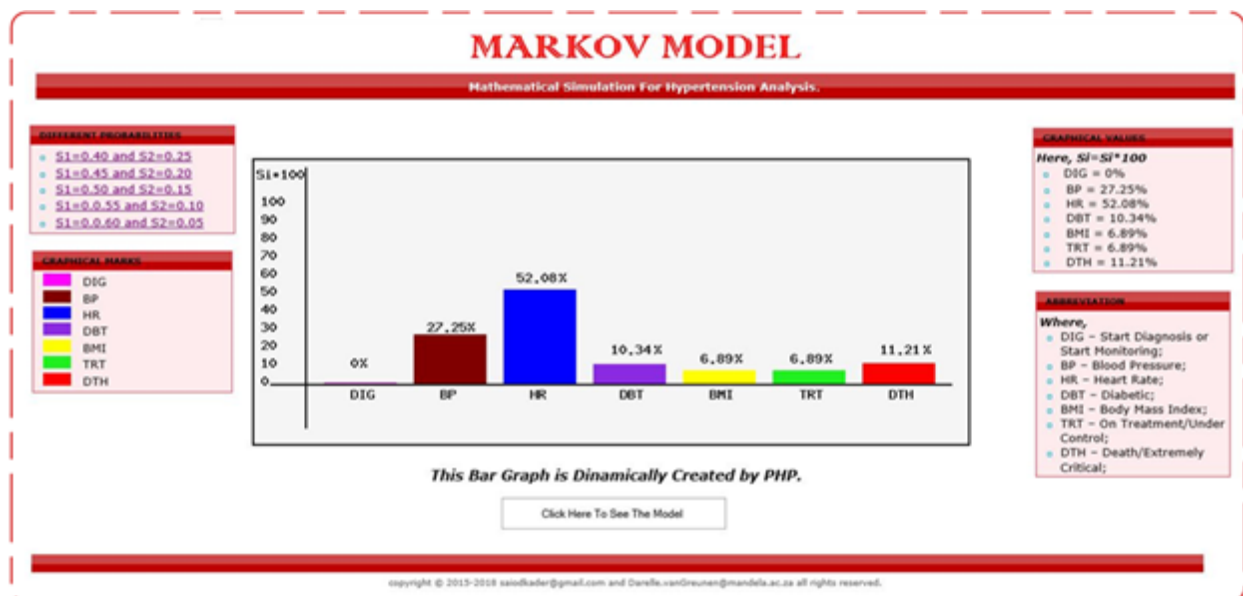


Figure 8.8: Graphical representation of the Markov chain Probability link structure of hypertension progression risk when “BMI to BP = 0.35 and BMI to HR = 0.30” (Saïod *et al.* 2019a)

The result of the graphical representation of the Markov chain Probability when “BMI to BP = 0.35 and BMI to HR = 0.30”. The mathematical simulation showed that when the “BMI to BP” and “BMI to HR” changes, the hypertension progression also changes accordingly, which increases the probability of death and vice versa. Some other conditions also affect hypertension diagnosis, such as lifestyle, smoke, alcohol, daily activity, work environment and conditions, including other diseases.

Table 8.11 (Saïod *et al.* 2019a) describes the different matrix probability simulations according to “BMI to BP” transmission in hypertension diagnosis, as follows:

Table 8.11: Different Matrix probability simulations according to “BMI to BP” transmission in hypertension diagnosis (Saïod *et al.* 2019a)

Conditions								
BMI to BP	BMI to HR	DIG	BP	HR	DBT	BMI	TRT	DTH
0,35	0,3	0%	27,25%	52,08%	10,34%	6,89%	6,89%	11,21%
0,4	0,25	0%	27,30%	52,19%	10,36%	6,91%	6,91%	11,23%
0,45	0,2	0%	27,36%	52,30%	10,38%	6,92%	6,92%	11,26%
0,5	0,15	0%	27,42%	52,41%	10,41%	6,94%	6,94%	11,28%
0,55	0,1	0%	27,25%	52,08%	10,34%	6,89%	6,89%	11,21%
0,6	0,05	0%	27,53%	52,62%	10,45%	6,97%	6,97%	11,33%

Figure 8.9 (Saïod *et al.* 2019a) describes the different matrix probability of hypertension progression risk simulation according to “BMI to BP” transmission, as follows:

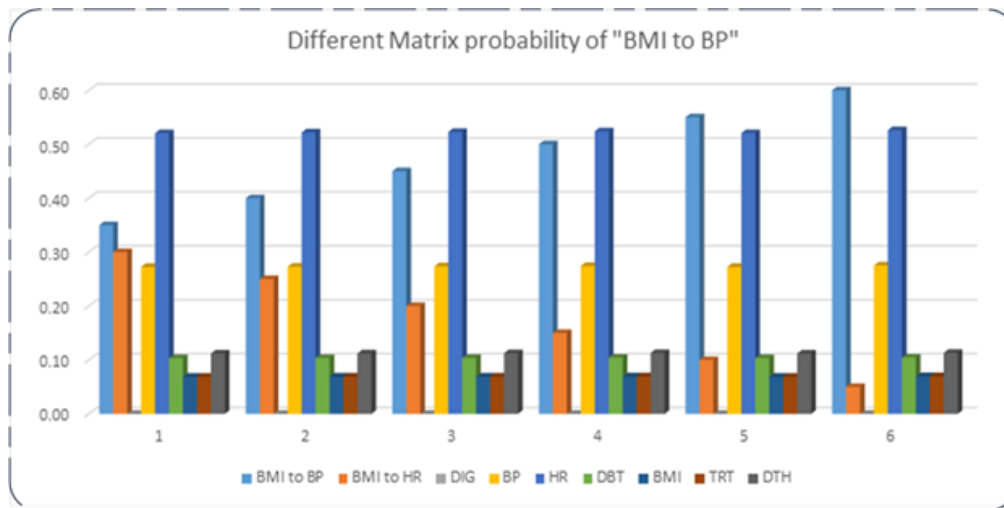


Figure 8.9: Different Matrix probability simulations according to “BMI to BP” transmission in hypertension diagnosis (Saïod *et al.* 2019a)

Different Matrix probabilities, the matrix probability model and the graphical representations have been included in the Appendix of “BMI to BP”.

## 8.6 The perfect matching

Every apex of the nodes is incident and to exactly one edge of the matching is an assignment of nodes, called perfect matching. The concept of perfect matching is  $n/2$  edge; it means that perfect matching is only possible on nodes with the even vertices number only. Complete matching or 1-factor is the other name of a perfect matching. Here, the dynamic Hungarian algorithm is presented, appropriate to optimally solve the assigned task in condition with changing edge costs, time and also improving the healthcare service.

The combinatorial optimisation algorithm of the Hungarian method is to solve the assignment issues in polynomial time, with expected later primal-dual methods. The assignment problem is widely-studied and exists in many application domains, known as the maximum weighted bipartite matching problem (Burkard *et al.* 1999). The Hungarian algorithm undertakes the existence of a bipartite graph,  $G = (H, P; E)$  that have illustrated in figure 8.9 (Saïod *et al.* 2017), where  $E$  is the set of edges and  $H$  and  $P$  are the sets of nodes in each baffler of the diagram.

Let us call a function  $y : (H \cup P) \rightarrow R$  a *potential* if,  $y(i) + y(j) \leq c(i, j)$  for each  $i \in H, j \in P$ . The potential value of  $y$  is  $\sum_{v \in H \cup P} y(v)$ .

*The time of each perfect matching is the latest value of each potential.*

The perfect matching of tight edges discovered by the Hungarian method: an edge  $ij$  is called tight for a potential  $y$ , if  $y(i) + y(j) = c(i, j)$ . Let us denote the subgraph of the tight edges by  $G_y$ . The time of a perfect matching in  $G_y$  (if there is one) equals the value of  $y$ .

Suppose, there are four hospitals (same group hospital) in a big city to which a model has assigned tasks on a one-to-one basis. The time of assigning a given resource to a given task is also known. An optimal assignment needs to be found, which minimises the total time to better service. The central hospital call centre receives emergency phone calls and decides from which hospital to send the ambulance and to which hospital the minimum distance is between the hospitals (H) and the patient call location (P). Figure 8.10 shows a bipartite graph of the Hungarian algorithm, as follows:

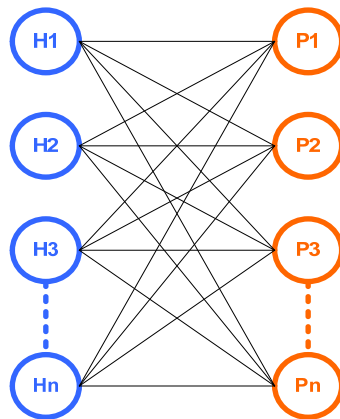


Figure 8.10: Bipartite graph of the Hungarian algorithm (Saiod *et al.* 2017)

A decision-making model is designed using the Hungarian algorithm to calculate how the ambulance should respond to the patient for emergency services to minimise the total time, as time is the biggest factor which can save the life of patients. The distance in kilometres (km) between the hospital H (ambulance

location) and the patient call location P are given in figure 8.11 (Saïod *et al.* 2017), as follows:

<b>P</b> \ <b>H</b>	<b>H1</b>	<b>H2</b>	<b>H3</b>	<b>H4</b>
<b>P1</b>	110	95	95	100
<b>P2</b>	55	105	75	85
<b>P3</b>	145	115	110	125
<b>P4</b>	165	130	115	135

Figure 8.11: Matrix of edge weights (Saïod *et al.* 2017)

*Step One - Subtract row minima:* This step is to determine the lowest element. Then in that row, subtract it from each element. In row 1 subtract 95, in row 2 subtract 55, in row 3 subtract 110 and in row 4 subtract 65 as the lowest element.

$$\begin{bmatrix} 110 & 95 & 95 & 100 \\ 55 & 105 & 75 & 85 \\ 145 & 115 & 110 & 125 \\ 65 & 130 & 115 & 135 \end{bmatrix} \sim \begin{bmatrix} 15 & 0 & 0 & 5 \\ 0 & 50 & 20 & 30 \\ 35 & 5 & 0 & 15 \\ 0 & 65 & 50 & 70 \end{bmatrix}$$

*Step Two - Subtract column minima:* In this step similarly as in step one for each column let us determine the lowest element, then subtract it from each element in that same column. In column 1 subtract 0, in column 2 subtract 0, in column 3 subtract 0 and in column 4 subtract 10.

$$\begin{bmatrix} 15 & 0 & 0 & 5 \\ 0 & 50 & 20 & 30 \\ 35 & 5 & 0 & 15 \\ 0 & 65 & 50 & 70 \end{bmatrix} \sim \begin{bmatrix} 15 & 0 & 0 & 0 \\ 0 & 50 & 20 & 25 \\ 35 & 5 & 0 & 10 \\ 0 & 65 & 50 & 65 \end{bmatrix}$$

*Step Three - Cover all zeros with a minimum number of lines:* This step is to cover all zeros in the resulting matrix. The minimum number of horizontal and vertical lines should be used to cover all zeros. An optimal assignment exists between the zeros if  $n$  lines are required. The algorithm stops.

$$\begin{bmatrix} 15 & 0 & 0 & 0 \\ 0 & 50 & 20 & 25 \\ 35 & 5 & 0 & 10 \\ 0 & 65 & 50 & 65 \end{bmatrix}$$

(The matrix above has three lines drawn through it: a horizontal line through the first row, and two vertical lines through the first and third columns.)

Step four will continue, as it required less than  $n^{\text{th}}$  lines.

*Step Four - Create additional zeros:* This step is to determine the smallest element (call it  $k$ ) in step three. This smallest element was not covered by a line. All uncovered elements must subtract by  $k$ , here is the smallest element. Then if the element is covered twice then add  $k$  to all the elements. We have to proceed to step five, as we have a minimal number of lines as less than 4.

*Step Five:* This step is to determine the smallest entry (5) that is not covered by any line. Therefore, in each uncovered row subtract 5.

$$\begin{bmatrix} 15 & 0 & 0 & 0 \\ 0 & 50 & 20 & 25 \\ 35 & 5 & 0 & 10 \\ 0 & 65 & 50 & 65 \end{bmatrix} \sim \begin{bmatrix} 15 & 0 & 0 & 0 \\ -5 & 45 & 15 & 20 \\ 30 & 0 & -5 & 5 \\ -5 & 60 & 45 & 60 \end{bmatrix}$$

Now add 5 to each covered column.

$$\begin{bmatrix} 15 & 0 & 0 & 0 \\ -5 & 45 & 15 & 20 \\ 30 & 0 & -5 & 5 \\ -5 & 60 & 45 & 60 \end{bmatrix} \sim \begin{bmatrix} 20 & 0 & 5 & 0 \\ 0 & 45 & 20 & 20 \\ 35 & 0 & 0 & 5 \\ 0 & 60 & 50 & 60 \end{bmatrix}$$

And now let us return to *step three*.

*Step Three:* Again, cover all the zeros in the resulting matrix. A minimum number of horizontal and vertical lines should be used to cover all zeros.

$$\begin{bmatrix} \cancel{20} & \cancel{0} & \cancel{5} & \cancel{0} \\ 0 & 45 & 20 & 20 \\ \cancel{35} & \cancel{0} & \cancel{0} & \cancel{5} \\ 0 & 60 & 50 & 60 \end{bmatrix}$$

*Step Four:* Since the minimal number of lines is less than 4, return to step five.

*Step Five:* Note that 20 is the smallest entry not covered by a line. Subtract 20 from each uncovered row.

$$\begin{bmatrix} 20 & 0 & 5 & 0 \\ 0 & 45 & 20 & 20 \\ 35 & 0 & 0 & 5 \\ 0 & 60 & 50 & 60 \end{bmatrix} \sim \begin{bmatrix} 20 & 0 & 5 & 0 \\ -20 & 25 & 0 & 0 \\ 35 & 0 & 0 & 5 \\ -20 & 40 & 30 & 40 \end{bmatrix}$$

Then add 20 to each covered column.

$$\begin{bmatrix} 20 & 0 & 5 & 0 \\ -20 & 25 & 0 & 0 \\ 35 & 0 & 0 & 5 \\ -20 & 40 & 30 & 40 \end{bmatrix} \sim \begin{bmatrix} 40 & 0 & 5 & 0 \\ 0 & 25 & 0 & 0 \\ 55 & 0 & 0 & 5 \\ 0 & 40 & 30 & 40 \end{bmatrix}$$

Now return to *step three*.

*Step Three:* Cover all the zeros in the matrix with the minimum number of horizontal or vertical lines.

$$\begin{bmatrix} \cancel{40} & \cancel{0} & \cancel{5} & \cancel{0} \\ \cancel{0} & 25 & \cancel{0} & \cancel{0} \\ \cancel{55} & \cancel{0} & \cancel{0} & \cancel{5} \\ \cancel{0} & \cancel{40} & \cancel{30} & \cancel{40} \end{bmatrix}$$

*Step Four:* Again, determine the smallest element of lines like 4. This smallest element was not covered by a line. The calculation is finished as an optimal assignment of zeros is possible.

$$\begin{bmatrix} 40 & 0 & 5 & \boxed{0} \\ 0 & 25 & \boxed{0} & 0 \\ 55 & \boxed{0} & 0 & 5 \\ \boxed{0} & 40 & 30 & 40 \end{bmatrix}$$

We have found zero as the total cost for this assignment. Therefore, it must be an optimal assignment.



Now, let us return to the original time matrix of the same assignment.

<b>110</b>	<b>95</b>	<b>95</b>	<b>100</b>
<b>55</b>	<b>105</b>	<b>75</b>	<b>85</b>
<b>145</b>	<b>115</b>	<b>110</b>	<b>125</b>
<b>65</b>	<b>130</b>	<b>115</b>	<b>135</b>

Therefore, the hospital should send ambulance H4 to Site P1, ambulance H3 to Site P2, ambulance H2 to Site P3 and ambulance H1 to Site P4.

### 8.6.1 The perfect matching analysis

Information and Communication Technologies (ICT) in healthcare organisations can be used in a beneficial way to address the key benefits and challenges faced by EHR systems and policymakers for them to increasingly recognise this potential. ICT enabled solutions to support the provision of effective, efficient and good quality services when implemented on a larger scale DBMS.

Healthcare policy makers and strategists inevitably will have to find some way in which to deliver more and more complex services to meet the increasing demand and expectations for the promotion and maintenance of health, treatment and care. A significantly essential component is the confirmation that healthcare professionals must actualise the expected benefits to ensure EHRs adoption. Associated with the specific impact of isolating and organisational factors in designating EHRs adoption, is called a knowledge base gap. Therefore, these need to be assessed on the adoption of EHRs in healthcare settings, the unique contributions of isolating and organisational factors, as well as the possible interrelations between these factors.

All experimentally measured units such as time, distance and motion, are continuous variables and calculated in standard deviations and units in standard

time formats. Time is estimated using a count of the incidences of an activity within a certain time period and reported as proportions. To facilitate comparisons across studies, taking into account the different sampling units, such as ambulance encounter versus ambulance total emergency service time, a relative time difference was calculated. The relative time difference was determined for each, considering the time it took to document using a computer, minus the time it took to document on paper, divided by the time it took to document on paper, producing a negative value if the EHRs was time efficient. Ninety-five percent confidence intervals were calculated for differences in means and proportions to assess the significance of reported differences when there was insufficient information to compute 95% confidence intervals.

The weighted averages were calculated for both types of sampling unit ambulance encounter and emergency service time, to accumulate for the changeability across the test studies. The following formulas have been used to calculate the weighted averages (Lise *et al.* 2005):

$$WA = \frac{\sum_{i=1}^n [SW(i) * RTD(i)]}{\sum_{i=1}^n SW(i)} \quad (8.9)$$

$$\text{In which, } (SW) = (n_{group1} + n_{group2}) \quad (8.10)$$

$$(RTD) = \frac{(\text{documentation time}_{group2} - \text{documentation time}_{group1})}{\text{documentation time}_{group1}} \quad (8.11)$$

*Where, WA – Weighted Average,*

*SW – Sampling Weight*

*RTD – Relative Time Difference*

The overall research identified that to achieve the benefits, depends on successful EHR systems implementation and use. Only DQ can provide confidence about the EHRs data to providers so that the benefits of using EHRs, such as best service, data accessibility, quality, safety measurement, improvement and reporting, can be seen.

## 8.7 The overall experiment result analysis

As shown in section 8.4, the hypertension diagnosis has been successfully handled by HIDM based on Fuzzy-Ontology. It also demonstrated successful implementation of the HIDM and addressing the DQ issues for hypertension diagnosis in the LSDB following the guideline accumulated by the proposed approaches. During the HIDM implementation process, each and every step has been specified for diaphanous contrivance and the specific to-do lists. As the development process has used the formal ontology associated with fuzzy logic, the appropriateness and efficiency achievement from the HIDM hypertension diagnosis could be assumed. The fundamental contrivance, the HIDM performance, is an incorporate statement, which needs to be constructed in a logical order as Fuzzy-Ontology.

The eventual contrivance of the proposed HIDM is to provide a methodological instruction to handle efficiently the big data as an accomplishment confirmation of appropriate achievement. In spite of this, as emphasised between indication and philosophical introduction, it is not easy to implement a universal qualitative and quantitative corresponding analysis with another subsistence approach.

Widely accepted circumstances confirm still quantitative limitation present in the entire subsisting apprological ontology (Carvalho *et al.* 2016), together with the contrivance for constructing Fuzzy-Ontology, crisp or probabilistic ontologies. Vicelike, METHONTOLOGY is the renowned approach, which does not comprise any evaluation, even if it allows the systematic approach for constructing crisp ontology from scratch. Another example, the applicability of NeON, has been proven in several exploration platforms but did not provide any meticulous appraisalment.

The possibility of the diligent approach is evaluating many used case crisp ontologies, except for providing any range of appraisalments. According to Carvalho *et al.* (2016) the latest probabilistic ontology development approach omits the appraisalment segment too. Therefore, this chapter seeks to just accommodate a

corresponding way of constructing a used case of hypertension diagnosis in a big data environment using the HIDM methodology. In as much as ontology is an existing approach, which cannot be appraised rigorously, it makes clear that generally all stakeholders including the HIDM professionals are selecting from existing approaches and simply combine appropriate approaches to achieve the best performance according to their needs.

According to the above-mentioned phenomenon, the ontology remained as unapprised evaluation approach, such as other existing consequential approaches along with Fuzzy-Ontology non-methodological elaboration approaches. Therefore, the following lineament of the proposed HIDM evaluated methodology could be anticipated to introduce affix in the Fuzzy-Ontology evaluation excursion:

- *Analogies of HIDM:* The proposed HIDM follows all generic rules and procedure as a formal Fuzzy-Ontology approach, excluding a few significant rules, such as using single or combined approaches, where needed and re-using the existing knowledge-based Fuzzy-Ontology elements in appropriate occurrences. As HIDM is the new way hybrid approach based on the Fuzzy-Ontology evaluation process, it may initially need additional time to understand the implementation process;
- *Methodological guideline of HIDM:* The proposed HIDM provides the first methodological guideline for elaborating the hybrid approach based on Fuzzy-Ontology from scratch. Analogies to other existing approaches show that HIDM is more mature, complete and compressive. For the non-methodological evaluation process, it is absolutely significant when an important step is easily excluded. A similar incident can be identified in existing hybrid evaluation approaches. For example, re-using the Fuzzy-Ontology elements is not a common consideration in present hybrid evaluation approaches;
- *Applicability and interoperability of HIDM:* The proposed HIDAs is a generic hybrid approach based on Fuzzy-Ontology and has higher applicability and interoperability compared to other present approaches. The

possibility of HIDM is to build own Fuzzy-Ontology elements to re-use them, rather than fuzzify each and every existing crisp ontology, which imposes extra resources for their uses. This reduces the applicability performance and provides a limitation on the domain-specific dependency, where the approaches were previously implemented. The main contribution of HIDAs is to grant a different generic solution approach from scratch for the big data domain that does not depend on existing domain specific ontology;

- *Limitation of HIDM:* Ontology management according to the different approaches related to conflict management in the process of knowledge integration, such as ontology matching/alignment, ontology merging or ontology mapping.

The knowledge base data has two different types in HIDAs:

- 1) Appropriate data;
- 2) Fuzzy related data;

This concept easily draws the borderline among them and accomplished them with the specific strategical methodology. Where Fuzzy-Ontology deals with vague data, an existing associative approach deals with appropriate data. Although the main contribution of HIDM is to illustrate a methodological guideline to develop a hybrid approach to deal with big data, this approach also has all the possibilities to deal with the crisp ontology development. The applicability and interoperability of the HIDM development have been demonstrated in the crisp ontology evaluation process. This feature declares that the HIDM approach can be used as a methodological guideline for crisp ontology development. In conclusion, the HIDM approach can be appropriate for both crisp ontology and Fuzzy-Ontology development as a guideline according to the extensive and generic features.

## 8.8 Summary

This chapter described the applicability of the HM based on Fuzzy-Ontology by illustrating its use in a hypothetical hypertension diagnosis project, the HIDM. Novel

Fuzzy-Ontology development approaches for big data, presented hybrid approaches as HIDM, including the perfect matching and mathematical simulation. The HIDM demonstrates the approaching directive for constructing the Fuzzy-Ontology approach for hypertension diagnosis in EHRs environment from the base. According to our implicit wisdom delegation, the HIDM has been considered to aim at the data standardisation to deal with DQ issues, including imprecise and vague data according to the lesson learned from existing different approaches. The entire HIDM construction process was divided into 11 development steps and determined and implemented where necessary in each step.

Figure 8.12 demonstrates the combined outcome of Chapter Three, Chapter Four, Chapter Five, Chapter Six, Chapter Seven and Chapter Eight, as follows:

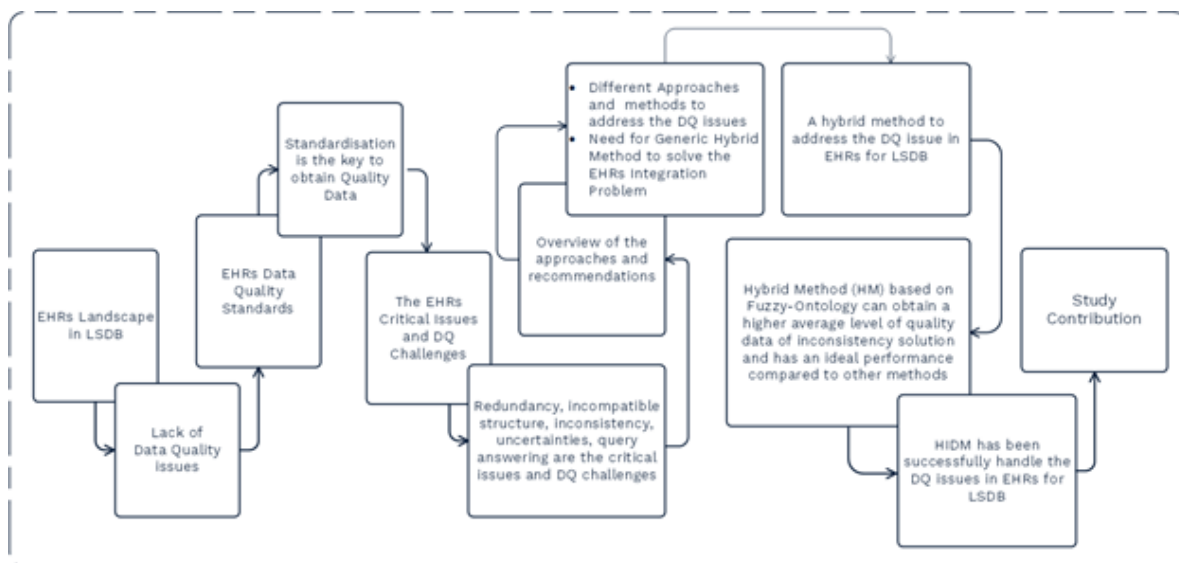
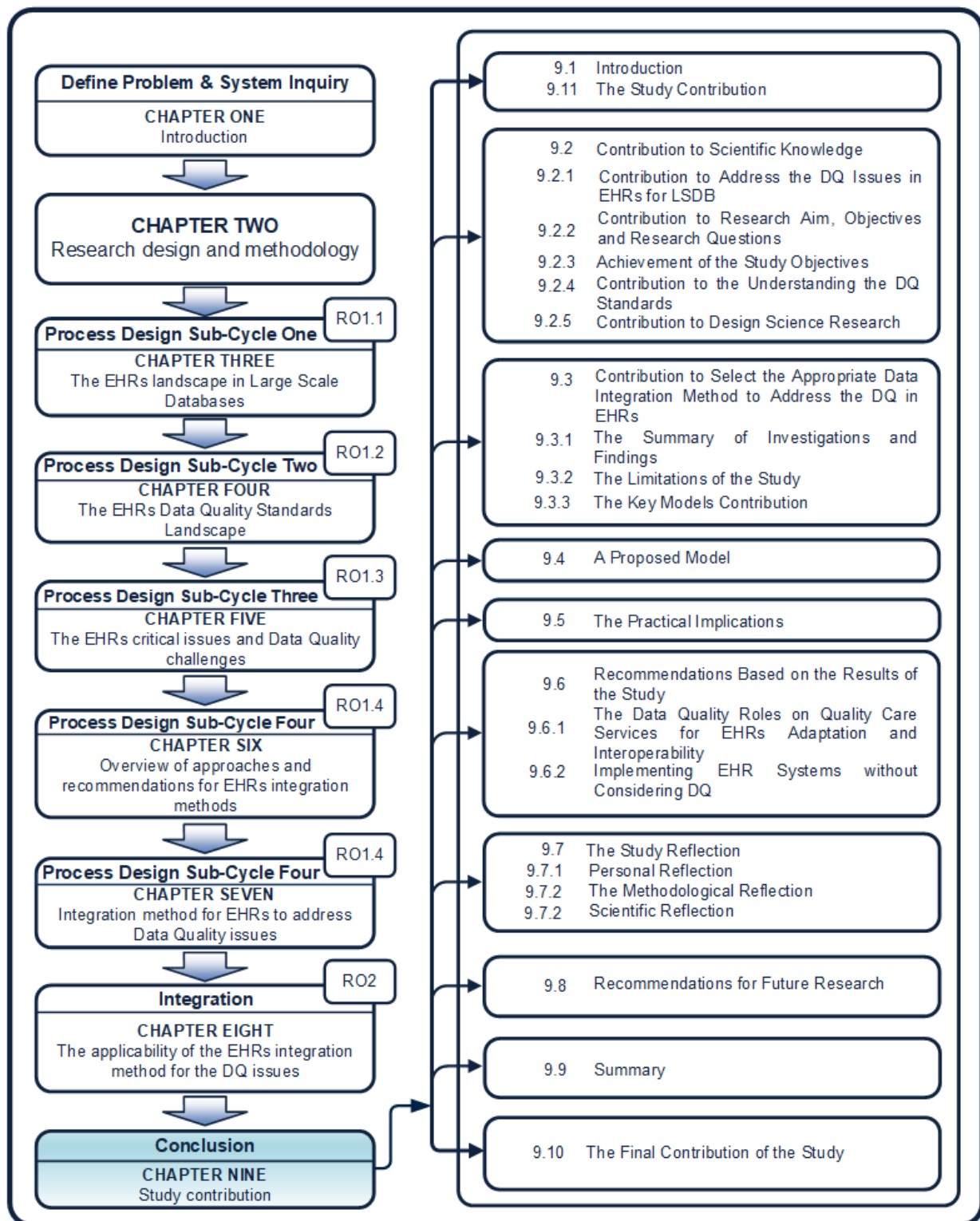


Figure 8.12: The combination outcome of Chapter Three, Chapter Four, Chapter Five, Chapter Six, Chapter Seven, and Chapter Eight

The HIDM provides a structured approach for the selection of an appropriate HIDM based on Fuzzy-Ontology that can support the interoperability of big health DI systems. This chapter demonstrated the usefulness of EHRs interoperability when DQ is perfect for the LSDB.

## CHAPTER NINE: Study contribution



Outline of the Chapter Nine

## CHAPTER NINE

### 9.1 Introduction

The main purpose of this chapter is to summarise the current study contributions and to present its contribution in comparison to the existing literature on the role of addressing the DQ issues in EHRs for the LSDB. It maps to the conclusion phase of the main DSR process, described in section 2.3.2.1 and highlighted in figure 9.1.

The aim of this chapter is to illustrate the applicability of the hybrid method based on Fuzzy-Ontology to address DQ issues, described in section 2.3.2.1 and highlighted in figure 8.1. This chapter focuses on the application of HM based on Fuzzy-Ontology intelligent systems to diagnose hypertension. The applicability of HM was done in three stages, firstly determining the usability of the proposed Hybrid Method (HM) based on Fuzzy-Ontology (see 8.2). *In section 8.3*, the specification for the proposed HIDM for EHRs integration was specified.

This was followed by an overview in section 8.4 of the real-world project: A Fuzzy Hypertension Specific Ontology, which in section 8.5 performed a mathematical simulation for hypertension progression analysis based on the Markov Chain probability model and in section 8.6 a similarity measurement was performed based on the Hungarian algorithm. The result of the analysis for perfect matching analysis was discussed in section 8.6.1, whereas section 8.7 provided the overall results of the analysis. A summary of the chapter is provided in section 8.8.

#### 9.1.1 The study contribution

This chapter involves the contribution to scientific knowledge to investigate and address the DQ issues associated with EHRs for LSDB.



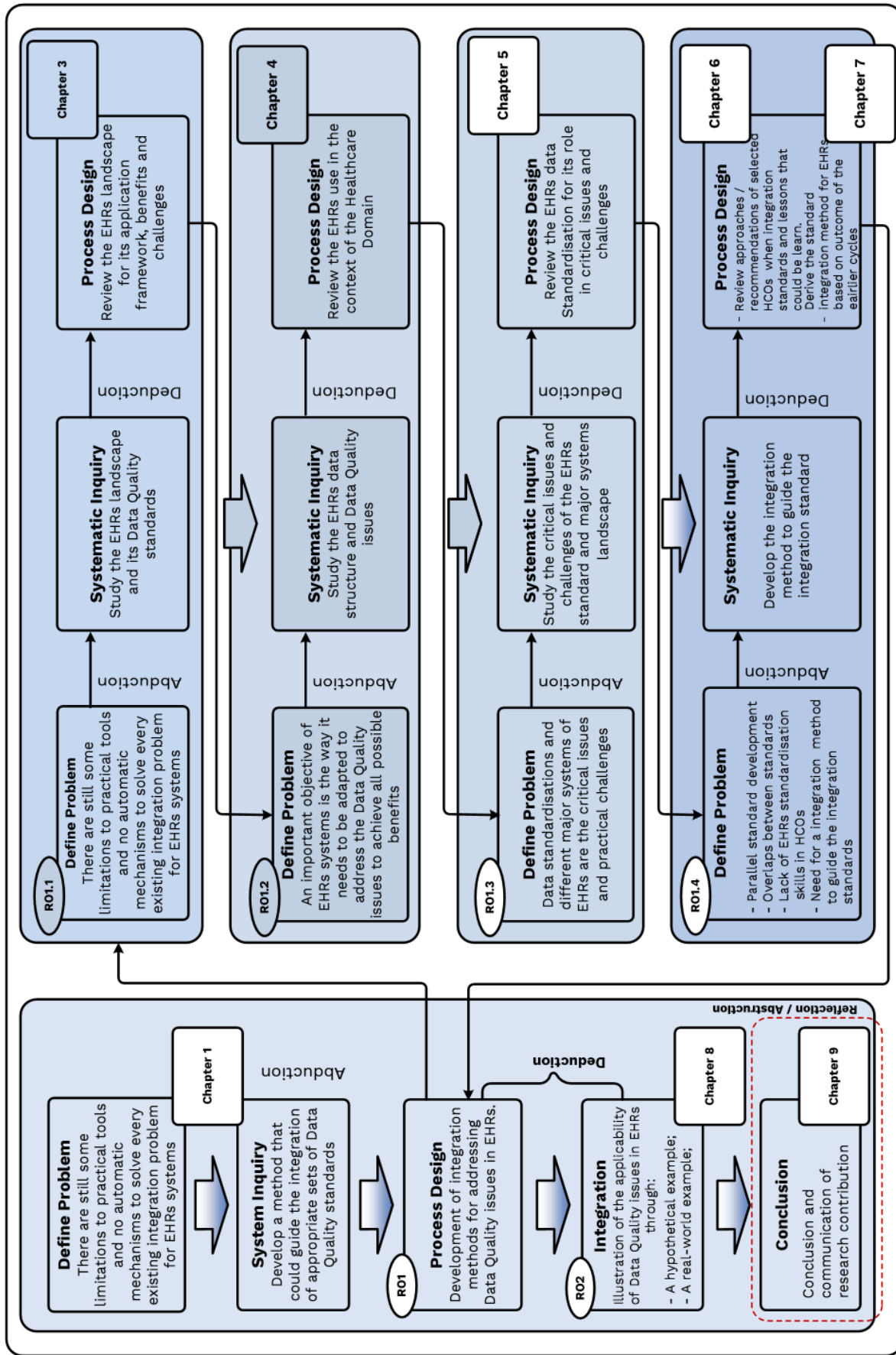


Figure 9.1: The position of Chapter 9 in the design science research process used in this study

The study contributions made by the research presented in this thesis to the scientific knowledge are discussed in section 9.2. The chapter contribution selecting the appropriate data integration method to address the DQ in EHRs is presented in section 9.3 and the summary is provided in section 9.5. Information in EHRs is being promoted for use in clinical decision support, patient registers, measurement and improvement of integration and quality of care services and translational research. To do this EHR-derived data product creators need to logically integrate patient data with information and knowledge from diverse sources and contexts.

The main contribution of this study is the improvement of a novel framework for an effective method for electronic health records to achieve its maximum benefits and reduce DQ challenges across HCOs to the minimum.

*The first contribution:* The applicability of the HAs based on Fuzzy-Ontology was described by illustrating its use in a hypothetical hypertension diagnosis project, the HIDM. Novel Fuzzy-Ontology development approaches for hypertension diagnosis, presented hybrid approaches as HIDM. The HIDM demonstrated the approaching directive for the constructing approach Fuzzy-Ontology for hypertension diagnosis in EHRs environment from the base.

According to our implicit wisdom delegation, the HIDM has been considered to aim at the data standardisation to deal with DQ issues, including imprecise and vague data according to the lessons learned from existing different approaches. The entire HIDM construction process was divided into 11 development steps and determined and implemented where necessary in each step. The HIDM provided a structured approach for the selection of an appropriate HAs, based on Fuzzy-Ontology that can support the interoperability of big DI systems. This contribution demonstrated the usefulness of HAs for big data communication.

*The second contribution:* The second contribution presented a mathematical simulation for hypertension diagnosis based on the Markov Chain Probability Model (MCPM). The MCPM has been defined as a stochastic process, which has strong progressive characteristics and the future evolution relies only on the current

circumstance situation. The transformation probability represented by itself the average interval hypertension risk level probabilities passes from the condition of end diagnosing/monitoring of the condition and before reception by them to the next probability condition during the same interval. The Markov model was illustrated by means of a position conversation and transition diagram, which demonstrated the entire position and transition probabilities.

*The third contribution:* The consensus method was also applied to solve the matching conflicts in EHRs integration. In practice, a dynamic Hungarian algorithm matching tool was implemented by combining PCP and consensus techniques. The EHRs consist of the following essential steps to achieve the goal: **the formal concept analysis, the conceptual clustering, the ontology generation, the Grid-File for multi-attribute search and the semantic representation conversion.** EHR technology became even more essential for modern healthcare services with the increasing communication network (Internet) and ICT technologies.

The aim of EHR systems are not only to improve the healthcare service and wellbeing, but it is also an indispensable demand to design a novel framework for EHRs services to reach beyond independence towards the sustainability of modern society and adaptation. Introducing EHR systems in healthcare service can, however, offer vast benefits to HCOs and society. The social and ethical acceptance is an important factor for the EHR systems adoption, such as services relying on the trust between patients and providers towards them. In this chapter, the possible benefits and challenges of data quality were discussed by introducing efficient EHR systems in HCOs.

The dynamic Hungarian algorithm showed how a decision-making system for the assignment problems with emergency services could save time and reduce the service costs. The results showed that both accuracy and completeness have a large impact on the approach's effectiveness. In real-time scenarios, the goal of the method algorithm is to efficiently integrate health data and repair inconsistent data instantly and accurately, when changes in the edge time and costs appear.

The overall scenario and challenges discussed above showed data quality in EHR systems, which demonstrated the method to be effective with regard to the accurate performance of the provider's service. This method presented the result of the principal theoretical characteristics considered to tackle thereafter any theoretical and practical problems for both qualitative and quantitative methodologies of implementing EHRs. The EHR system would not have any limits and the system could be modified efficiently to benefit scalability.

## **9.2 Contribution to scientific knowledge**

One of the main advantages of using EHRs in the HCO activity comes from their ability to provide useful information for decision-making to health professionals to support continuity of quality care and the DQ is the key component to enable interoperability. One of the challenges associated with EHRs is to integrate diverse heterogeneous data into a global schema, which satisfies the needs of users.

### **9.2.1 Contribution to address the DQ Issues in EHRs for LSDB**

This study presented in this research identified that fragmentation and inability of DQ in EHRs to integrate heterogeneous health data would improve the healthcare service and help clinicians to diagnose their patients. The data in EHRs are being flourished for use in healthcare decision-making, patient registers, providing quality healthcare service, measurement and characteristic diagnosis, improvement of integration and translational research. To achieve this the EHR-derived data systems contriver needs to consequentially integrate health data with information and knowledge from diverse heterogeneous sources.

Integrating heterogeneous data within an EHR using the hybrid method based on Fuzzy-Ontology algorithms has improved the accuracy of the diagnosis and compensates for suboptimal DQ and hence creates a dataset that is satisfying the needs of the users. The DQ has been identified as a key value as one of the

essential divers of interoperability. A number of challenges were identified as limiting DQ issues in EHRs for LSDB include, as follows:

- 1) The different data sources have several standards and different major systems, which causes the overlaps or contradiction between the standards;
- 2) Heterogeneous data sources are often incompatible and sometimes inconsistent in their data structures;
- 3) Lack of proper implementation guidance of the data standards, which affects widespread DQ. This issue impacts the interoperability when developers are implementing and integrating the EHR systems with different standards;

The HIDM to EHRs improves the integration of DQ and the potential for better use of quality data to improve the quality of care services. This fact has an important impact on the way in which data are introduced by healthcare professionals, on the way that data are recorded on databases and also on the heterogeneity found when trying to integrate data from diverse heterogeneous sources. Incrementally integrating health data using the HIDM methods, identified nearly 100% of quality data in real-world cases. Incrementally integrating the six datasets improved correctness, consistency, completeness, rationality, understandability and conciseness for heterogeneous EHRs. Manual validation and mathematical simulation confirmed the accuracy of the method when DQ is perfect in EHRs.

This study contributed to the solutions to address DQ in EHRs. It dealt with the challenges associated with DQ with EHRs implementation. The study addressed the challenges associated with DQ with combining the Fuzzy-Ontology and reduced the risk of data inconsistencies between different standards.

### **9.2.2 Contribution to the research aim, objectives and research questions**

The aim of this study was to investigate and address the DQ issues associated with EHRs for the LSDB. To achieve the aim, the following research objectives were addressed:

- ✓ To analyse the impact of EHRs formal concept analysis adoption on the research productivity;
- ✓ To examine research productivity using DQ conceptual clustering;
- ✓ To examine research productivity using DQ generation;
- ✓ To examine research productivity using traditional systems;
- ✓ To design a model on EHRs adoption for the increase of research productivity in the Grid-File for multi-attribute search and semantic representation conversion;

To achieve the above research objectives, the following the main research question was developed:

***The feasibility of introducing EHR systems in HCOs to improve the data quality in order to achieve all possible benefits pertaining to healthcare services.***

To effectively address this single research question, the following open sub-research questions have been addressed, as follows:

- ✓ To define the most meaningful associations among heterogeneous data sources that can be explored to improve DQ in EHRs for LSDB;
- ✓ To identify the kinds of integrity constraints specified in the global schema of data mapping that can be explored to improve DQ in EHRs for LSDB;
- ✓ To identify the uncertainties in the data integration that when minimised, resulting in an improved DQ in EHRs for LSDB;

### 9.2.3 Achievement of the study objectives

This section will present conclusions that have been reached for each of the objectives as stated above, detailed below:

*Objective One:* To analyse the impact of EHRs formal concept analysis adoption on the research productivity;

The EHRs interoperability was investigated on the health data of some 3000 patients with hypertension provided by different hospitals, clinics and a medical aid company in South Africa and computed the results in the range of the predefined limit by the domain experts.

A face-to-face survey was conducted with patients which involved male and female groups in different age groups with different hypertension conditions. Questionnaires and measures were distributed according to the ratio as described in Chapter Eight. Subjects were selected to obtain the parameters such as age (range between 20 to 60 years old), gender (male and female), BMI level, blood pressure (120/80mmHg) and heart rate. In addition, questions such as working background, medical history and lifestyle were asked in the distributed questionnaire for obtaining additional knowledge on the subjects. It was revealed that the DQ issue in EHRs had a significant impact on research productivity.

*Objective Two:* To examine research productivity using DQ conceptual clustering;

DQ improvement in EHRs was investigated and addressed using the HIDM method based on Fuzzy-Ontology where the research took place. A HIDM model was generated using MATLAB R2014B and the mathematical simulation of hypertension progression analysis was based on the Markov Chain probability model performed using PHP5, MySQL under Apache server in WAMP5. There another perfect matching was provided using the dynamic Hungarian algorithm, appropriate to optimally solve the assigned task in condition with changing edge costs and time, as well as improving the healthcare service.

Two book chapters and one conference paper have been published according to the study which has been checked by the Turnitin. Therefore, it can be concluded that using a HIDM methodology based on Fuzzy-Ontology, the disease progression analysis using the Markov Chain probability model and similarity measurement

based on the Hungarian algorithm increases researcher performance in terms of research productivity.

*Objective Three:* To examine research productivity using traditional systems;

During the study, the positive impact in EHRs has been shown that are referred by the DQ handled by the HIDM methodology and the comparison with the traditional system has been provided. Therefore, it can be concluded that, when using the HIDM methodology to address the DQ issues in EHRs, quality care productivity is very high.

*Objective Four:* To examine research productivity using DQ Generation;

The DQ in EHRs has been investigated and addressed over HCOs where the research took place. A HIDM model based on Fuzzy Ontology was generated using MATLAB R2014B and mathematical simulation of hypertension diagnosis based on the Markov chain probability model performed using PHP5, MySQL under Apache server in WAMP5, which clearly showed that DQ in EHRs had low significance in quality care service. Therefore, it can be concluded that using EHRs without DQ does not have a high significance on quality care service over HCOs.

*Objective Five:* To design a model on EHRs adoption for the increase of research productivity in the Grid-File for multi-attribute search and semantic representation conversion;

The joint impact of HIDM methodology in EHRs adoption and interoperability clearly showed a growth in DQ level that increased the quality of care service over HCOs. This objective was achieved by combining a hybrid method based on Fuzzy-Ontology, the disease progression analysis using the Markov Chain probability model and similarity measurement based on the Hungarian algorithm.

*Finally,* the model showed that the HIDM adoption and interoperability together enhanced the DQ that impact the quality healthcare services across HCOs.



## 9.2.4 Contribution to understanding the DQ standards

One of the challenges associated with DQ relates to different data sources having several standards and different major systems. Some of them are oriented to support only the patients, providers and HCOs demographic data. Some of them aim to enable the exchange of patient health data. Therefore, the studies contributed and focused to improve the understanding of the data pattern to reduce the inconsistency when integrating.

EHRs are usually used for purposes other than healthcare delivery, namely research or management. The health DBMS tend to be more reliable with IT development. However, the DQ issue associated with EHRs in the LSDB has become more relevant than ever as the utilisation of the DBMS is rapidly increasing both in magnitude and importance. The classification of those issues according to those standards consists of, as below:

- ***Fitness for use:*** DQ is relative to each objective and can be identified as fitness for use. This concept states that data can be considered of appropriate quality for one purpose but may not hold sufficient quality for another propose. For example, a clinical DBMS can qualify for economic analyses, but may not have been sufficiently qualified for a clinical study.
- ***Data quality:*** DQ is defined as data that is accurate, complete, relevant, timely, sufficiently detailed, appropriately represented (for example, consistently coded using a clinical coding system) and it should retain sufficient contextual information to support decision-making.

The DQ dimension is classified into four categories, as follows:

- ***Accuracy:*** Accuracy is defined as the degree of correctness and precision with which real-world data are represented;
- ***Completeness:*** Completeness is defined as the degree to which all relevant data are recorded;

- **Consistency:** Consistency is defined as the degree to which data satisfy specified constraints and business rules;
- **Timeliness:** Timeliness is defined as the degree to which the recorded data are up-to-date;

In another perspective, it is possible to analyse DQ concerning three roles, namely production, custodian and consumer, as follows:

- **Data producers:** Data producers are those that generate data (*for example, medical, nursing or administrative staff*).
- **Data custodians:** Data custodians are those that provide and manage computing resources for storing and processing data (*for example, database administrators and computer scientists*).
- **Data consumers:** Data consumers are those who use data in medical care (*for example, physicians, researchers and managers*).

The DQ standards classification schemes in this study also map to the conceptual knowledge contribution category.

### 9.2.5 Contribution to Design Science Research

The study presented in this thesis contributes to the DSR body of knowledge by creating a specific form of DSR artefact, as a method.

The HIDM for EHRs to address DQ issues in the LSDB is a generic method aimed to guide the EHRs integration of an appropriate set of methods and logical algorithm to support the interoperability of EHR systems. Irrespective of the absence of consensus regarding the number of DSR outputs focusing on the method from the artefacts, the HIDM is a form of a method artefact adding to the pipe of methods as DSR outputs. It fits the prescriptive knowledge contribution. In theory, every individual scientist is capable of being his/her most severe critic and

his/her own writings are expected to be discussed with real care and seriousness as well as the objections against his/her own novel ideas (Toulmin *et al.* 1979).

Hevner *et al.* (2004) have stated that Design Science Research for Information Systems must produce one or more artefacts. According to the definition, an artefact is an innovation that defines ideas, practices, technical capabilities and products, through which the analysis, design, implementation and use of Information Systems can be effectively accomplished.

Hevner *et al.* (2004) and March *et al.* (1995) have further stated that the artefacts for Design Science Research are constructs, models, methods and instantiations. Vaishnavi *et al.* (2004) have observed that in addition to the production of artefacts, design science research should produce better theories for the field of research. Constructs form the conceptual vocabulary of the field of study. Figure 9.2 (adopted according to Vaishnavi *et al.* 2015 and Aken *et al.* 2016) describes the DSR model that has been used in this study, as follows:

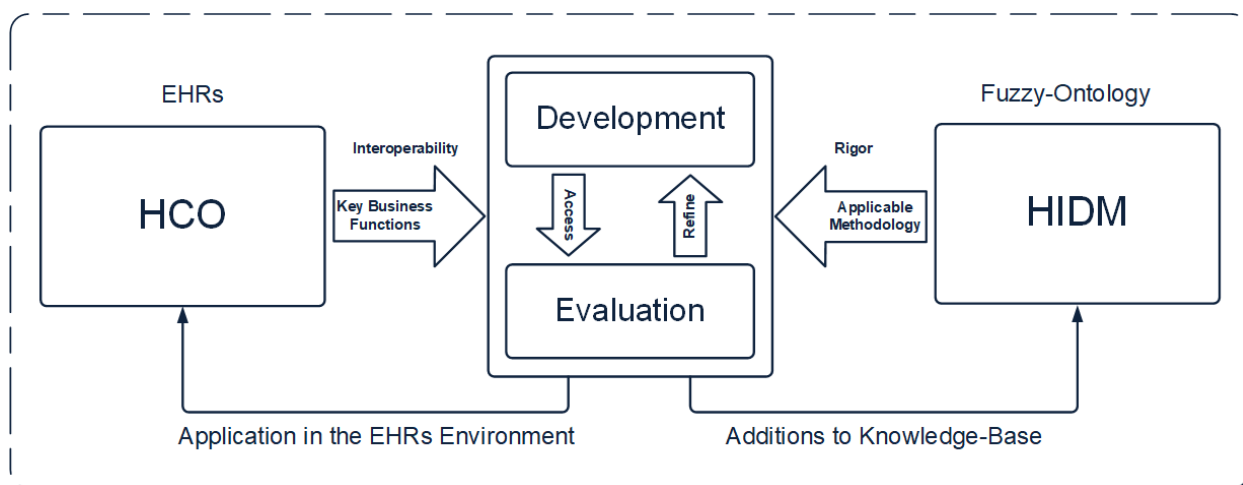


Figure 9.2: Design Science Research cycles adopted according to Vaishnavi *et al.* (2015) and Aken *et al.* (2016)

Constructs make up the language used to define and communicate the problems and solutions in the field of study. For Design Science Research, the term ‘model’ is used to refer to the set of propositions that specify relationships between the constructs. Methods are definitions of the processes that need to be achieved. A method may be stated as a set of steps to perform a given task or a method may

be specified as a formal computational algorithm. Instantiations are the actual implementations of the models and methods to demonstrate that they actually work. “*Better Theories*” provide an increased understanding arising from the study of the created artefacts.

The HIDM is a generic appropriate hybrid method aimed at guiding to address DQ issues in EHRs for LSDB to support the interoperability of health information integration systems. Irrespective of the absence of consensus regarding the number of DSR output focusing on the method from the artefacts, HIDM is a form of a **method** artefact adding to the assemblage of methods as DSR output. It fits the prescriptive knowledge contribution describe by Kuechler *et al.* (2004) and Gregor *et al.* (2013).

### **9.3 Contribution to select the appropriate data integration method to address the DQ in EHRs**

With the growing concern over the fragmentation of DQ issues in EHRs and their impact on healthcare systems, DQ is the key component to enable the quality of care and interoperability of EHRs. However, the DQ is fraught with diverse challenges, including different standards evaluated by similar organisations and inconsistency in the data standards. The study presented in this research contributes to the HIDM based on Fuzzy-Ontology to address the DQ issues in EHRs fragmentation for LSDB. Chapter Six highlighted the limitation in the number of publicly available methods to address the DQ issues in EHRs.

The research presented in this study adds to the publicly available body of knowledge on methods to guide the addressing the DQ issues to support EHRs’ interoperability.

Although other methods are possibly available, lack of access to these methods in the open source makes it arduous to learn from their approaches. As illustrated in figure 9.3, the HIDM consisted of five phases, as below:

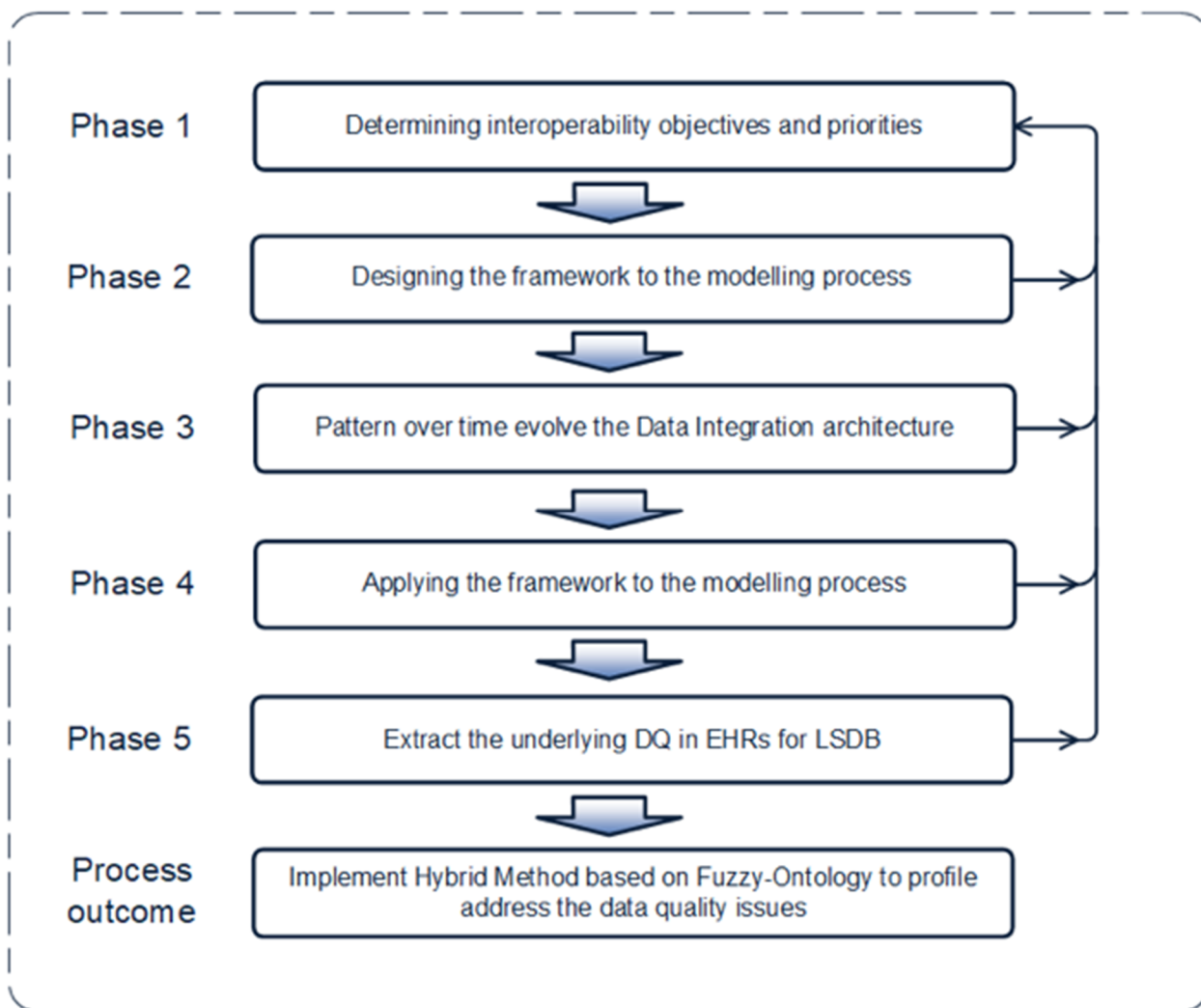


Figure 9.3 Proposed hybrid method Fuzzy-Ontology (researcher source)

The HM based on Fuzzy-Ontology consisted of five phases, as illustrated in Figure 7.2, detailed below:

- a) *Phase One – Determining interoperability objectives and priorities:* The determination of interoperability objectives described the impact of DQ in EHRs of the healthcare systems. It is a fact that the DQ objectives are the highest priority in EHRs that drive the quality care service. It has also been identified that the limited resources are directed at the area of greatest need.
- b) *Phase Two – Designing the framework to the modelling process:* The modelling process framework specified the description of the healthcare service process so that the required appropriate quality data sharing can

be determined. In this way, the EHR systems can automatically be provided and shared independently across data. The modelling process specifications should be driven by healthcare priorities.

- c) Phase Three – Pattern over time evolves the data integration architecture:* The EHRs Integration architecture guides the data standard that will be required to enable the EHRs integration across the healthcare organisations.
- d) Phase Four –Applying the framework to the modelling process:* Providing support the EHRs integration require a combination of different standards across healthcare organisations, which could often lead to issues of conflict between different combined standards.
- e) Phase Five – Extract the underlying DQ in EHRs for LSDB:* The health standard-based IHE profiles provide a standards-based framework for the exchange of EHRs integration service to support the quality of care. With the emergence of the EHRs as a pervasive healthcare information technology, new opportunities and challenges for use of clinical data for quality measurements arise with respect to DQ, data availability and comparability.

DQ of EHRs was sufficient to be used for the calculation of quality indicators, although comparability to the survey data was problematic. Standardisation is needed, not only to be able to compare different data collection methods properly but also to compare practices with different EHRs. EHRs have the option to administrate narrative data, but natural language processing tools are needed to quantify these text boxes. Such development can narrow the comparability gap between scoring quality indicators based on EHR and based on survey health data.

### 9.3.1 The summary of investigations and findings

The aim of this study was to investigate and address the DQ issues associated with Electronic Health Records (EHRs) for Large Scale Databases (LSD). Based on the results, the following conclusions are drawn, as follows:

- ✓ The integrated data often show inconsistency because different HCOs have several standards and different major systems, which have emerged as critical issues and practical challenges,
- ✓ EHRs are inherently difficult to coherently manage incompatible and sometimes inconsistent data structures from diverse heterogeneous sources.
- ✓ It is inherently difficult to coherently manage data from diverse heterogeneous sources.
- ✓ A logic analysis of inconsistent data analysis was described and the way to prevent hackers (see section 6.3.2).
- ✓ An inefficient EHRs system is unable to reduce data redundancy and prevent system application failures.
- ✓ It is practically challenging to integrate diverse data into a global schema, which satisfies the needs of the users.
- ✓ The efficient management of EHR systems using an existing DBMS presents a challenge because of the incompatibility and sometimes inconsistency of data structures.
- ✓ There is no common methodological approach currently in existence to effectively solve every data integration problem.
- ✓ Most of the existing EHRs data integration methods, such as peer-to-peer, data warehouses, middleware, data grid, data mining, semantic and ontology, actually establish semantic connections between heterogeneous data sources. Although these methods offer some advantages in some aspects, they do not provide a coherent mechanism to solve every data integration problem.
- ✓ None of them pays strong attention to data inconsistency, which has been a long-standing DQ challenge in health database environments.
- ✓ Existing literature shows that several techniques and major EHR systems currently exist to deal with DQ issues, which historically have faced DBMS. After a profound analysis of various cutting-edge commercial accomplishments existing on the software market and an intensive review

of the literature, some limitations to practical tools for EHR systems still appear.

- ✓ Physical access to diverse information sources of robust support is provided, but only if these are standard database structure tables.
- ✓ Two major barriers and challenges in the way of successful EHRs implementation, are, namely the human barriers (for example, professional and belief) and the financial barriers (for example, available money and a funding opportunity). The human factors become even more important as the benefits are only expected after the implementation and use of EHR systems.
- ✓ The difficulty involves the way in which to practically combine data from disparate, incompatible, inconsistent and typically heterogeneous data sources.
- ✓ The other difficult objective in EHRs is that data has a structure, which is usually complex and cannot be treated as a simple string of bytes.
- ✓ Often data inconsistency occurs because the data structure may depend on other structures, therefore, in a distributed system such as data management is very difficult.
- ✓ Another important aspect of an EHR integration system is whether the system is able to materialise data, which are retrieved from diverse sources through data mappings.
- ✓ The eradication of DQ issues in EHRs will benefit the integration of electronic health records and systems in the LSDB.
- ✓ The query answering in the context of data exchange is the final important issue for DQ.

### 9.3.2 The limitations of the study

Research studies usually have some limitations that may raise doubts as to the validity and reliability of the findings.



One major limitation of the study was that, due to the limited time and cost, the actual EHRs adaptation and interoperability could only take place over a certain period of time. In future studies, more time for testing, investigating more real-world issues and implementation should be allocated, which could enhance the study.

Another limitation was the limited data quantity, which needs to be implemented for the HIDM methodology and monitor the performance over time.

### 9.3.3 The key model's contribution

The main contribution of this study is that it has conceptualised new methodology to address the DQ issues in EHRs for LSDB and empirically investigated its relationship with research productivity.

*Firstly*, this study developed the inputs inspiring to conceive that the HIDM has significance on research productivity (see figure 9.4), as follows:

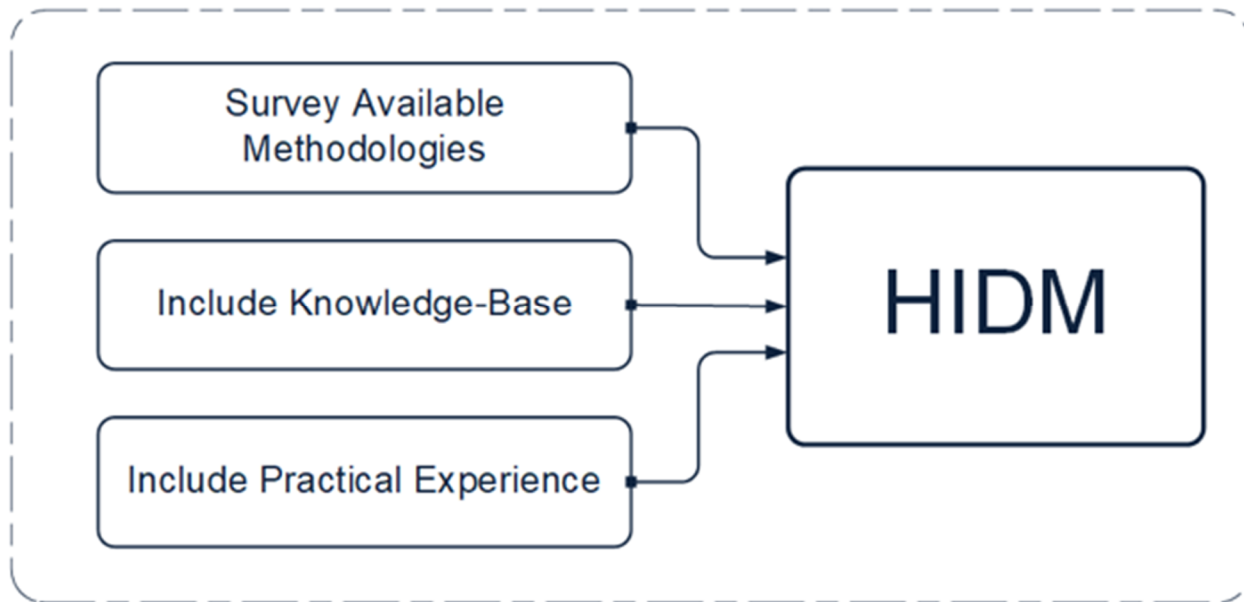


Figure 9.4: Inputs inspiring to conceive the HIDM (Saïod *et al.* 2019a)

*Secondly*, this study constructed the structure of the proposed system –HIDM for EHRs based on Fuzzy-Ontology (see figure 9.5) for the LSDB.

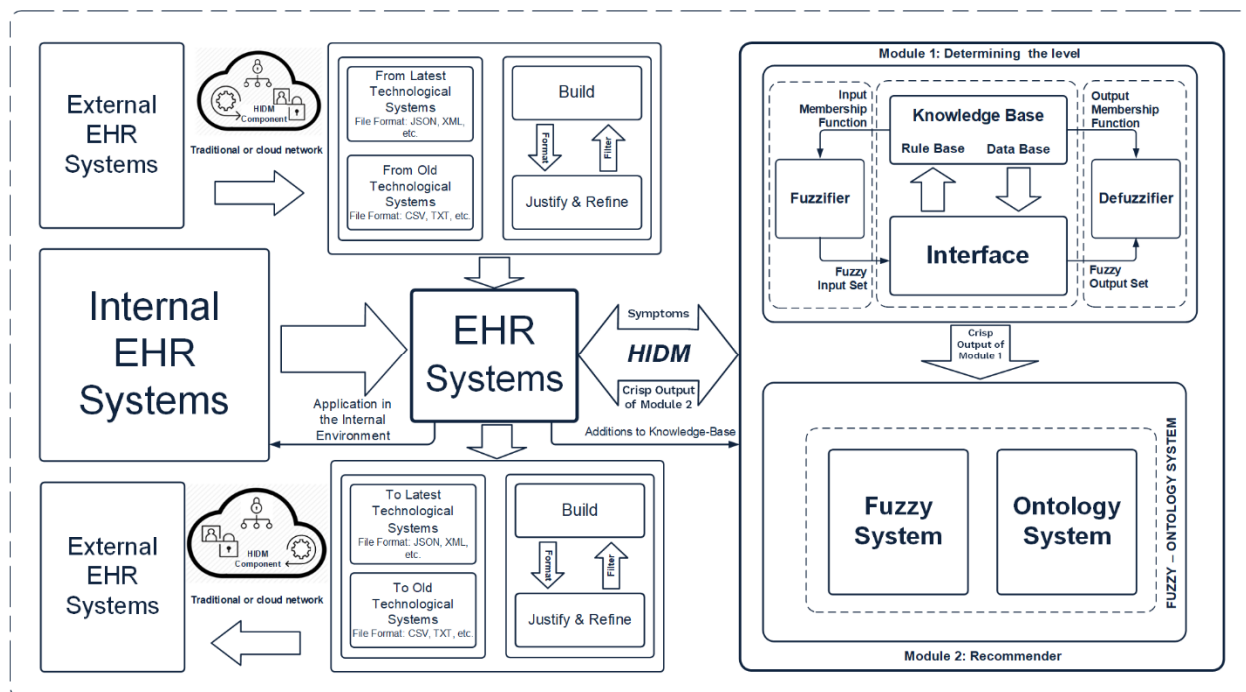


Figure 9.5: The complete HIDM structure based on Fuzzy-Ontology for EHRs integration systems (researcher source)

*Thirdly*, this study found that the overall visualised structure of the Fuzzy Hypertension specific Ontology has significance on research productivity (see figure 9.6), as follows:

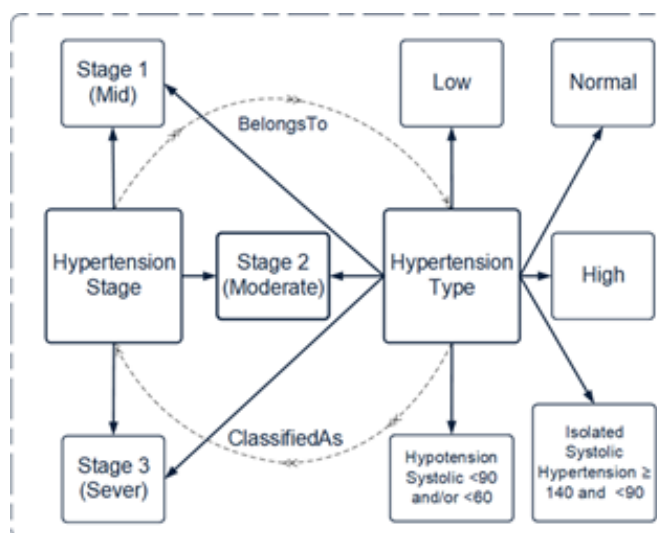


Figure 9.6: The overall visualised the structure of the Fuzzy Hypertension specific Ontology (Saïod *et al.* 2019a)

*Fourthly*, this study demonstrated the impact of the DQ in EHRs for LSDB, the different matrix probability simulation according to “BMI to BP” transmission in hypertension diagnosis (see figure 9.7) based on the Markov Probability Chain Model, as follows:

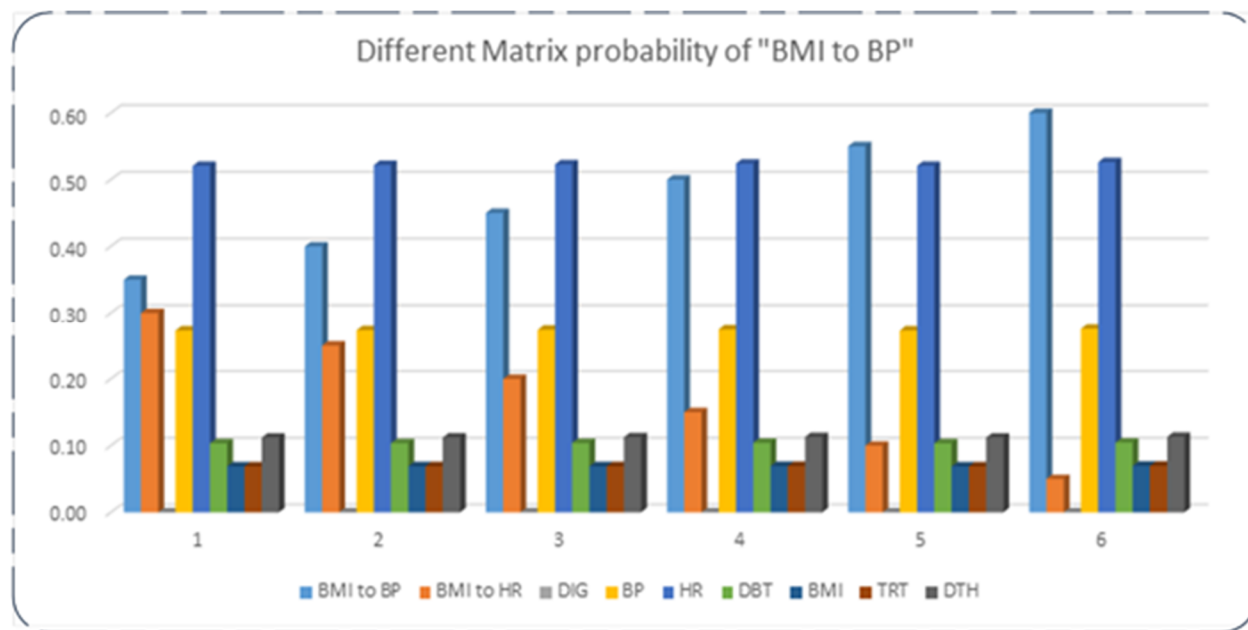


Figure 9.7: Different Matrix probability simulation according to “BMI to BP” transmission in hypertension diagnosis (Saïod *et al.* 2019a)

*Fifthly*, this study demonstrated the incident and to exactly one edge of the matching impact of the DQ in EHRs for LSDB. Using the Hungarian algorithm have presented the appropriate to optimally solve the assigned task in condition with changing edge costs, time as well as improving the healthcare services (see figure 9.8), as follows:

110	95	95	100
55	105	75	85
145	115	110	125
65	130	115	135

Figure 9.8: The perfect matching result of the Hungarian algorithm (Saïod *et al.* 2017)

## 9.4 A proposed model

The research found empirical evidence that different constructs, the HIDM for EHRs based on Fuzzy-Ontology, can reduce the DQ issues and increase the quality of healthcare services at HCOs. This study empirically confirmed the HIDM methodological contribution to the EHRs adaptation and interoperability to the theory of technology acceptance, as indicated in figure 9.9 (Saïod *et al.* 2019a), as follows:

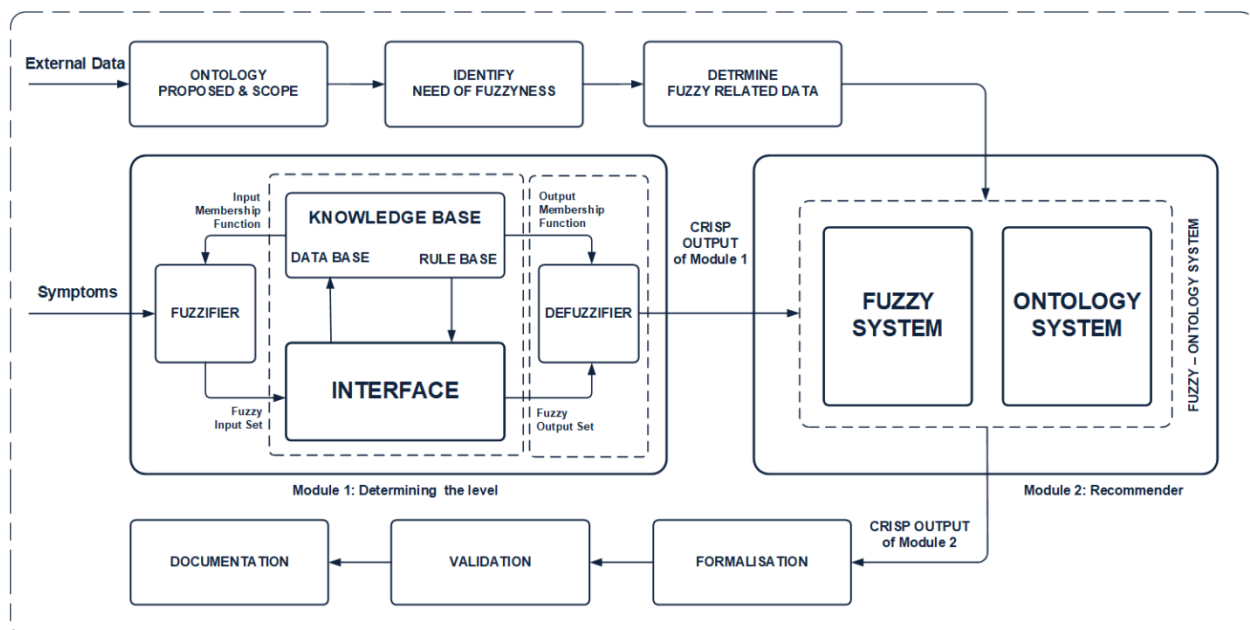


Figure 9.9: Structure of the proposed system – Hybrid Integration Development Methodology (HIDM) for EHRs based on Fuzzy-Ontology (Saïod *et al.* 2019a)

## 9.5 The practical implications

A practical implication of this research is linked to healthcare services to increase the quality of healthcare service among HCOs. However, this study identified that DQ highly impacts the EHRs adaptation and interoperability and positively influences healthcare service quality. It means that the better the DQ among EHRs adaptation and EHRs interoperability with proper healthcare staff training, the higher the quality care service. This study finding may also have implications for changes and interventions demanding to move away from current DQ issues in EHRs without training, using a traditional system (without using research methodology and training) and several standards across HCOs for Healthcare Quality Service (HQS) productivity.

## 9.6 Recommendations based on the results of the study

This section contains the recommendations based on findings from this study that enable clinicians to improve HQS productivity, as follows:

### 9.6.1 The data quality roles on quality care services for EHRs adaptation and interoperability

The role of DQ has been recognised worldwide as a key to successful EHRs adoption and interoperability. The respondents of this study have investigated and implemented the HIDM methodology to increase DQ in EHRs to provide HQS. In this regard, IT professionals with HCOs need to formulate strategies on EHRs adoption and interoperability for the improvement of HQS.

### 9.6.2 Implementing EHR systems without considering DQ

The role of using EHR systems without considering DQ is one of the key issues of unsuccessful EHRs implementation. According to respondents, this has poor

effects on HQS as compared to when DQ is in order. The findings indicated that, when implementing the EHR systems without considering the DQ, the significance was very poor for EHRs adaptation and interoperability. IT professional including all stakeholders, therefore, need to formulate strategies to motivate clinicians, including each healthcare staff member, who has access to the EHR systems to be trained in each level to be aware of DQ impacts in EHRs, instead of using EHR systems without training.

## **9.7 The study reflection**

This section represents a reflection on the study from three perspectives. A personal reflection is presented in section 9.7.1. The methodological reflection of the appropriateness of the DSR paradigm for the research is presented in section 9.7.2 and a scientific reflection is presented in section 9.7.3.

### **9.7.1 Personal reflection**

The study presented in this research provided me with the opportunity to integrate my knowledge of the health domain, including other several domains as a computer engineer, with my 17 years of software development experiences. During my professional careers as a software developer, I became aware of the importance of the DQ issues in any application environment. If I look at the current situation and to the past, healthcare information was all paper-based and/or using traditional systems and this might still be found in many HCOs, especially in developing countries. This is time-consuming, but time is one of the important factors in HQS to save patients' lives and this is always an issue with information accuracy.

Having changed from paper-based health data to EHRs has brought about a wide range of improvement in healthcare services. In my professional careers, I have been faced with many real big issues impacting to the DQ issues, where I was continuously seeking the way to address the DQ issues, to apply the DQ benefits

to the application environment. My interest in the research presented in this thesis was initiated when **Prof Dr. Darelle van Greunen** faced difficult and complicated DQ issues in EHRs. In the period that followed, I had to investigate the issues and had read several publications on eHealth and EHRs until the final decision to embark on the research.

### 9.7.2 The methodological reflection

This section provides my own reflection on the appropriateness of the chosen research paradigm and research process.

At the start of the research presented in this study, the interpretive and DSR paradigm were considered for their potential suitability as well as the qualitative and quantitative research methodology. However, the interpretive research paradigm was found to be inadequate, since it involves the interpretation of a research context from a researcher's perspective.

I have found the DSR paradigm to be the most appropriate to guide the research process as it is primarily concerned with the creation of artefacts to solve the given problem. The research presented in this thesis involved the development of a method form of the artefact to guide the investigation and address the DQ issues in EHRs for LSDBMS. In addition, the guidelines emanate from the fundamental principles of DSR. The principles are based on the premise that the acquisition of knowledge and understanding the root cause of the problem, as well as to design its solution, revolves around the development and application of the artefact. The seven guidelines are summarised below (Henver *et al.* 2014):

1. ***Design as an artefact:*** The underlying purpose of DSR is the creation of an artefact either in the form of a construct, a model, a method or an instantiation. Therefore, a DSR project must have as its outcome a viable artefact in one of these forms;

2. *Problem relevance:* The goal of DSR should be the development and implementation of solutions to the important and relevant “*health business*” problem;
3. *Design evaluation:* Evaluation is an integral part of DSR. Hence, the utility, quality and efficiency of a design artefact must be demonstrated through well-executed evaluation methods. A designed artefact can be evaluated using observational, analytical, experimental, testing or descriptive methods;
4. *Research contribution:* Effective DSR must provide clear and verifiable contributions. The contribution could be based on the novelty, generality or significance of the research;
5. *Research rigour:* This refers to how the research was conducted. DSR should apply rigorous methods in the development and evaluation of the designed artefact;
6. *Design as a search process:* DSR intrinsically involves iterations through the development and evaluation cycles, as evaluation provides valuable feedback to the development cycle. To arrive at a satisfying solution that is *good enough*, the design science researcher should utilise all available means to reach the desired solution;
7. *Communication of research:* The result of DSR must be effectively communicated to relevant audiences (technology and management-oriented). This is to enable the realisation of benefits that could accrue from the implementation of the artefacts, as well as to build a cumulative knowledge base for future extension of the artefacts;

### 9.7.3 Scientific reflection

This section provides my own reflection on the lessons learned from the research presented in this study to the scientific body of knowledge.

The problem space for the study presented in this thesis is well-known and attempts have been made by many developed countries to address DQ issues in



EHRs. However, given the limitation of skills and experience to understand the DQ impact in the entire EHR systems, special attention is needed on data when capturing, inserting and importing data to EHR systems and all source data must be validated through the validation system. Therefore, EHRs should be integrated through HIDM methodology.

The primary goal of the research presented in this thesis was to investigate and address DQ issues in EHRs for the LSDB that could provide interoperability of EHR systems and help healthcare professionals to provide quality care services. The literature review and analysis, as well as discussions with experts on eHealth DQ standardisation, confirmed the need for such a method. The main contribution of this study was the development of the HIDM system based on Fuzzy-Ontology to address DQ issues in EHRs with combined features to search, extract, filter, clean and integrate data, to ensure that users can coherently create new consistent data sets. The method contributed by providing a structured way to address the challenges associated with DQ issues in EHRs for LSDBMS.

## 9.8 Recommendations for future research

During the course of conducting the research presented in this study, the following future research possibilities were identified:

*The first consideration:* The generic Hybrid Integration Development Methodology (HIDM) based on Fuzzy-Ontology, developed in the research presented in this thesis, is aimed to address the DQ issues in EHRs for the LSDB. Its usefulness and performance were demonstrated by a real-life project based on HIDM – a Fuzzy Hypertension specific Ontology.

*The second consideration:* Further research is required to determine the applicability of HIDM in HCOs, such as a hospital, clinics, laboratories and the medical aids domain in a real-world situation.

*The third consideration:* Implement the mathematical simulation based on the Markov Probability Chain Model for diagnosis of the hypertension risk level, where some other related consideration such as age, BMI, HR, sugar level, physical activity and genetics, should be considered. This method can be used to measuring risk level probability for another diagnosis.

*The fourth consideration:* Implementation of a similarity measurement will be useful for a similar case to improve the service accuracy and will reduce the service cost including saving time, which has a high impact on saving patient lives.

*The fifth consideration:* Further research is also required for the completeness of HIDM.

*The sixth consideration:* The HIDM was especially aimed to address the DQ issues in EHRs for LSDB for the healthcare domain. Further study is required to determine its applicability beyond the healthcare domain including smaller DBMS.

*The seventh consideration:* It would be more interesting to examine the impact of DQ issues in EHRs (*but keep on high alert as it may impact the risk of the patient's life*) on HQS productivity of HCO using real-life health data, rather than a survey-based data.

*The eighth consideration:* It will also be interesting to examine the results when the number of participants is increased and the number of HCOs to be included in the research are higher.

*The ninth consideration:* The focus of the research presented in this study was on addressing the DQ issues in EHRs in the LSDB to guide the IT professionals including the HCOs stakeholders that support interoperability. Further research is required to determine the DQ governance structure required for the implementation of HDIM.

*The tenth consideration:* This study has identified the needs of guideline the low and medium-income countries on how to determine the HCOs interoperability

goals and identify the minimum set of DQ standards. The potential of the generic HDMI has been demonstrated by its use in guiding the development of the South African *National Health Normative Standards Framework for Interoperability in eHealth (HNSF)*.

## 9.9 Summary

This chapter described the contribution of the study, including the HM based on Fuzzy-Ontology by illustrating its use in a hypothetical hypertension diagnosis project, the HIDM.

The mathematical simulation demonstrated the way to use the Markov Chain Probability Model to measure hypertension risk levels to save patient lives. The perfect matching for the similarity measurement demonstrated how to save time and service costs, which can save patients' lives, as time is very important for the emergency services and the smooth running of the HCOs business. Figure 9.10 demonstrates the combined outcome of Chapter Three, Chapter Four, Chapter Five, Chapter Six, Chapter Seven, Chapter Eight and Chapter Nine, as follows:

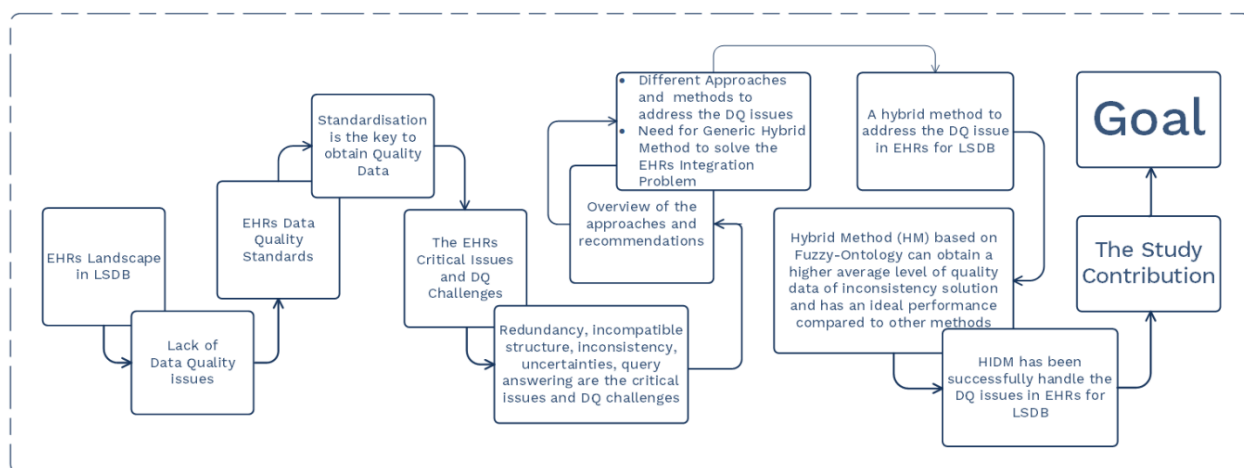


Figure 9.10: The combination outcome of Chapter Three, Chapter Four, Chapter Five, Chapter Six, Chapter Seven, Chapter Eight and Chapter Nine

The use of interoperable EHR systems features has the potential to transform efficient healthcare service, including the safety and quality of healthcare services

to the patient, and lowering the overall healthcare cost. The adaptation of appropriate EHR systems is vital to the goal of interoperable EHR systems. The research presented in this thesis is a step towards fulfilling this need. It is hoped and believed that this study will be useful and maximise the EHRs benefit of HCO's investment and improve the quality of healthcare services.

## **9.10 The final contribution of the study**

This chapter marks the end of the DSR process. It provided a summary of the research contributions made by the research presented in this study. It addresses the gap of the DQ issues in EHRs to determine the way to deduce and address the DQ issues for EHRs interoperability. This study contributes by adding to existing methodology and solutions to address DQ issues in EHR systems. It also contributes to the analyses of the needs of HCOs, including selecting the proper EHR standards to support interoperability.

The results of DQ issues in EHRs for LSDB conducted by this study confirmed findings from existing literature on the impact of HCOs. The present study contributes to knowledge by adding DQ issues in the EHRs for LSDB concept that positively affect HQS provided the proper impact of DQ for the HQS. This research is novel compared to current existing research in the nexus DQ issues in EHRs. When implementing the EHR system for HCOs from the proposed HIDM methodological model, including mathematical simulation and perfect matching, it showed that this is intended to boost HQS and allow healthcare professionals, including HCOs, to improve their quality of care services.

## REFERENCES

- Abdel, N. H. Z., Mohammed, E. and Seham, A. E. (2015), Electronic Health Records: Applications, Techniques and Challenges; International Journal of Computer Applications (0975 – 8887) Volume 119 – No.14, June 2015
- Abdullah Gani, Aisha Siddiqa, Shahaboddin Shamshirband, Fariza Hanum - A survey on indexing techniques for big data: taxonomy and performance evaluation, 26 March 2015, pp 1-44, DOI: 10.1007/s10115-015-0830-y, Print ISSN: 0219-1377, Online ISSN: 0219-3116
- Abram, D. (1996). The spell of the sensations. New York: Pantheon
- Abdolhadi, N., Mohammad, M. S., Abbas, A. B. (2012). Landfill site selection by decision-making tools based on fuzzy multi-attribute decision-making method. DOI 10.2007/s12665-011-1137-2
- Ackoff, Russell L. (1961). The Design of Social Research, Chicago: University of Chicago Press
- AHIMA's physician practice council resolution on quality data and documentation in the EHR. Available at :  
[http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1\\_035781.hcsp?dDocName=bok1\\_035781](http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_035781.hcsp?dDocName=bok1_035781) Date: 2007 (Accessed September 09, 2017).
- Ahmed MK. and Nasruddin H. (2017) - A note on “A novel approach to multi attribute group decision making based on trapezoidal interval type-2 fuzzy soft sets”.  
<http://dx.doi.org/10.1016/j.apm.2016.04.014>
- Aken JVA., Chandrasekaran AB. and Joop H. (2016) - Conducting and publishing design science research Inaugural essay of the design science department of the Journal of Operations Management.  
<https://doi.org/10.1016/j.jom.2016.06.004>

- Aken VJE. (2005). "Management research as a design science: Articulating the research products of mode 2 knowledge production in management". *Br J Manage.* 2005; 16(1): 19–36.
- Alejandro C., Cristian M., Alejandro Z., Daniela G. and Silvia S. (2017) - Persisting big-data: The NoSQL landscape. DOI: <http://dx.doi.org/10.1016/j.is.2016.07.009>
- Alex R, Cristiano ADC., Rodrigo DRR and Kleinner SFDO (2017) -Personal Health Records: A Systematic Literature Review. DOI: 10.2196/jmir.5876
- Alexopoulos P., Wallace M., Kafentzis K. and Askounis D. (2012) "IKARUSOnto: A methodology to develop fuzzy ontologies from crisp ones," *Knowl. Inf. Syst.*, vol. 32, no. 3, pp. 667-695, Sep. 2012.
- Al-Sakran and Hasan O. (2015) - FRAMEWORK ARCHITECTURE FOR IMPROVING HEALTHCARE INFORMATION SYSTEMS USING AGENT TECHNOLOGY. *International Journal of Managing Information Technology* Volume: 7 Issue 1 (2015) ISSN: 0975-5926
- Amatayakul M. (2007). *Electronic Health Records: A Practical Guide for Professionals and Organizations*. 3rd ed. Chicago: American Health Information Management Association;
- Amy W., Dajun D. and Tolu O. (2016). A cross-sectional and spatial analysis of the prevalence of multimorbidity and its association with socioeconomic disadvantage in South Africa: A comparison between 2008 and 2012. DOI: <http://dx.doi.org/10.1016/j.socscimed.2016.06.055>
- Anagnostopoulos, I., Zeadally, S. & Exposito, E. *J Supercomput* (2016) - Handling big data: research challenges and future directions. 72: 1494. DOI: 10.1007/s11227-016-1677-z
- Andrew H. B., Milton C. W., Elisabeth A.L.F., Jonathan K., Mark J.S. and David P. (2012). *Model Parameter Estimation and Uncertainty: A Report of the ISPOR-*

SMDM Modeling Good Research Practices Task Force-6. DOI: <http://dx.doi.org/10.1016/j.jval.2012.04.014>

Arts, D.G.T., Keizer, N.F.D. and Scheffer, G-J. (2002). Defining and improving data quality in medical registries: A literature review, case study, and generic framework. *Journal of the American Medical Informatics Association* 9: 600-611

Ashwin B., Raghuram T., Reza SMS., Fatemeh N., Daniel AB and Kayvan N. (2015) - Big Data Analytics in Healthcare. DOI: <http://dx.doi.org/10.1155/2015/370194>

Athman B., Boualem B., Ahmed K. and Elmagarmid (2012). – Interconnecting Heterogeneous Information Systems, 2012, ISBN: 978-1-4613-7546-3, DOI: 10.1007/978-1-4615-5567-4

Aspden P. (2004). - Patient Safety Achieving a New Standard for Care. Washington, D.C: National Academies Press.

Aygün, R. S. and Yazici, A. (2004). Modeling and Management of Fuzzy Information in Multimedia Database Applications. *Multimedia Tools Appl.* vol. 24, pp. 29-56.

Bardach E. and Patashnik EM. (2015). A practical guide for policy analysis: The eightfold path to more effective problem solving. ISBN 978-1-4833-5946-5

Baron RJ, Fabens EL, Schiffman M, Wolf E. Electronic Health Records: Just around the Corner? Or over the Cliff? *Ann Intern Med.* 2005;143(3):222–6.

Baškarada S. and Koronios A. (2014). "A Critical Success Factors Framework for Information Quality Management". *Information Systems Management.* 31 (4): 1–20. doi:10.1080/10580530.2014.958023.

Beata MS. (2017) - Local Public Enterprise Business Model as Multiple Value Creation System. DOI: 10.4018/978-1-5225-2215-7.ch008

- Bernard, H.R. (2011). "Research Methods in Anthropology" 5th edition, AltaMira Press, p.7
- Bernhard, T. (2013). Dependencies in Relational Databases, December 1, ISBN: 9783663120186
- Bertino e., Dai c. and Kantarcioglu M. (2009). The challenge of assuring data trustworthiness. In X. Zhou, HaruoYokota, K. Deng, and Q. Liu, editors, Proceedings of the 14th International Conference on Database Systems for Advanced Applications (DASFAA 2009), Brisbane, Australia, 21-23 April 2009, volume 5463 of Lecture Notes in Computer Science, pages 22–33. Springer, 2009
- Bertoa M, Vallecillo A. An Ontology for Software Measurement. In: Calero C, Ruiz F, Piattini M, editors. Ontologies for Software Engineering and Software Technology. Heidelberg: Springer; 2006. pp. 175–196.
- Bizer, C., Heath, T., Berners-Lee, T. (2009). - Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS), pp.3.
- Bland, C. J., Center, B. A., Finstad, D. A., Risbey, K. R. and Staples, J. G. (2005). A theoretical, practical, predictive model of faculty and department research productivity, Academic Medicine, 80 (3): 225-237.
- Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med. 2010;363:501–504. doi: 10.1056/NEJMp1006114.
- Bobillo F. and Straccia U. (2011). "Fuzzy ontology representation using OWL 2," Int. J. Approx. Reasoning, vol. 52, no. 7, pp. 1073-1094, Oct. 2011.
- Bobillo F., Delgado M. and Gómez-Romero J. (2012). "DeLorean: A reasoner for fuzzy OWL 2," Expert Syst. Appl., vol. 39, no. 1, pp. 258-272.



- Bobillo F. and Straccia U. (2016). "The fuzzy ontology reasoner fuzzyDL," *Knowl.-Based Syst.*, vol. 95, pp. 12-34.
- Broder A. (2002). A taxonomy of web search. *ACM Sigir forum*. 2002; 36:3–10. ACM. DOI: 10.1145/792550.792552
- Brown RB. (2006). *Doing Your Dissertation in Business and Management: The Reality of Research and Writing*, Sage Publications
- Bryman, A. and Bell, E. (2015). *Business research methods* (3rd edition). Oxford, United Kingdom: Oxford University Press.
- Bunge, M. (1967b). *Scientific Research II. The Search for Truth*. Berlin: Springer-Verlag
- Burns, R.B. (1997). *Introduction to research methods* (3rd Ed). Longman Australia: Melbourne.
- Bryman, A. (2004). *Quantity and Quality in Social Research*. London: Routledge. First published in 1988.
- Carlos, B., Roberto, Y., Fernando, B., Sergio, I., Jorge, B., Eduardo, M., Raquel, T., Ángel, LG. (2016). *Emerging Semantic-Based Applications*; DOI: 10.1007/978-3-319-16658-2\_4
- Carol C. and Steven M. (2016). *DATABASE SYSTEMS - Design, Implementation and Management*. ISBN: 978-1-305-86679-9
- Carvalho R. N., Laskey K. B. and Costa P. C. G. D. (2016). "Uncertainty modeling process for semantic technology," *PeerJ Comput. Sci.*, vol. 2, p. e77, Aug. 2016.
- Chan, K., Fowles, J., & Weiner, J. (2010). Electronic health records and the reliability and validity of quality measures: a review of the literature. *Med Care Res Rev*, 67(5), 503-527.

- Chaowei Y., Qunying H., Zhenlong L., Kai L. and Fei H. (2016) - Big Data and cloud computing: innovation opportunities and challenges. <http://dx.doi.org/10.1080/17538947.2016.1239771>
- Cheng, H. (2014). Analysis of panel data, Third Edition, ISBN: 978-1-107-65763-2 August 4-6, 2010, Las Vegas, Nevada, USA 978-1-4244-8098-2/10/\$26.00 ©2010 IEEE: 207
- Christian, B., Sabine, H. S., Silke R.(2010). Data Bases, the Base for Data Mining, Data Mining in Crystallography, Volume 134. pp 135-16,7, DOI: 10.1007/430\_2009\_5
- Christina EM., Joseph AH, Adeline P., Rebecca ER. and Frances EB. (2014) - Simulated Electronic Health Record (Sim-EHR) Curriculum: Teaching EHR Skills and Use of the EHR for Disease Management and Prevention. doi: 10.1097/ACM.0000000000000149
- Christoph B. and Carlo M. (2014). - Foundations of Information and Knowledge Systems; 8th International Symposium, FolKS 2014. Bordeaux, France, March 3-7, 2014. Proceedings ISSN 0302-9743 e-ISSN 1611-3349. ISBN 978-3-319-04938-0 e-ISSN 978-3-319-04939-7; DOI 10.1007/978-3-319-04939-7
- Cimino JJ. (2013). Improving the electronic health record are clinicians getting what they wished for? JAMA. JAMA. 2013;309(10):991-992. DOI:10.1001/jama.2013.890
- Codd E.F. (2009). The relational model for database management. Version 2. Boston, MA: Addison-Wesley Longman Publishing Co, Inc,
- Cois, A., Ehrlich, R., 2014. Analysing the socioeconomic determinants of hypertension in South Africa: a structural equation modelling approach. BMC Public Health 14, 414. <http://dx.doi.org/10.1186/1471-2458-14-414> .
- Corrao J., Natalie, Robinson AG, Swiernik MA and Naeim A. (2010). "Importance of testing for usability when selecting and implementing an electronic health or

- medical record system," *Journal of Oncology Practice*, vol. 6, no. 3, pp. 120-124, 2010.
- Creswell, J. W., and Plano Clark, V. L. (2011). *Designing and Conducting Mixed Methods Research*. California, USA: SAGE Publications.
- Cross V. V. (2014). "Fuzzy ontologies: The state of the art," in *Proc. IEEE Conf. Norbert Wiener 21st Century (21CW)*, Jun. 2014, pp. 1-8.
- Cynthia SS., FAAN RN., Matthew F., Heidi B., Jennifer SF., (2016). William AC. and Lisle SH. Providing primary care using an inter-professional collaborative practice model: What clinicians have learned. <https://doi.org/10.1016/j.profnurs.2016.11.004>
- Danielle G. T., DeKeizer N. F. and Scheffer G. J., (2002). *Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study, and Generic Framework*. DOI: 10.1197/jamia.M1087
- Daniel K., Jorn K., Geoffrey E. and Florian M. (2010) ISBN: 978-3-905673-77-7
- David DD. and Monideepa T. (2015) - *Understanding information exchange in healthcare operations: Evidence from hospitals and patients*. <http://dx.doi.org/10.1016/j.jom.2014.12.003>
- Dean B. B., Lam J., Natoli J. L., Butler Q., Aguilar D. and Nordyke R. J. (2009). Review: use of electronic medical records for health outcomes research: a literature review. *Med Care Res Rev*. 66(6):611–638.
- Dick R. S., Steen E. B., Detmer D. E. (1997). *Institute of Medicine. The Computer-Based Patient Record: An Essential Technology for Health Care*. 2nd ed. Washington, DC: National Academies Press.
- Dietz J. L. G. (2006) *Enterprise Ontology – Theory and Methodology*. Springer, Berlin, Heidelberg

- Dolin, R. H., Alschuler L., Beebe C., Biron P. V., Boyer S. L., Essin D., Kimber E., Lincoln T. and Mattison J. E. (2001). The HL7 Clinical Document Architecture. *J Am Med Inform Assoc* 8 (6):552–569.
- “E1384 Standard Guide on Content and Structure of Electronic Health Records.” American Society for Testing and Materials. Available at <http://www.astm.org>
- Edward H., Shortliffe J. J., Cimino D. (2013) - *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*, 2013, Springer Science & Business Media, ISBN: 9781447144748
- Eisenstein E. L., Collins R., Cracknell B. S., Podesta O., Reid E. D., Sandercock P., Shakhov Y., Terrin M. L., Sellers M. A., Califf R. M., Granger C. B. and Diaz R. (2008) Sensible approaches for reducing clinical trial costs. *Clin Trials*. 2008;5:75–84. doi: 10.1177/1740774507087551.
- Ellis T. J. and Levy Y. (2010). - *A Guide for Novice Researchers: Design and Development Research Methods*. Proceedings of Informing Science & IT Education Conference (InSITE) 2010
- Ellen, RF. (1984). Introduction. In RF Ellen (Ed.), *Ethnographic Research: A guide to general conduct (research methods in social anthropology)* (pp. 1-12). London: Academic Press
- Evangelos T. (2013). *Multi-criteria Decision Making Methods: A Comparative Study*. ISBN: 9781475731576
- Ewa O. (2013), *Incomplete Information: Rough Set Analysis*. ISBN: 9783790818888
- Eysenbach G. (2001). - What is e-health? *J Med Internet Res* 2001;3(2):e20, doi:10.2196/jmir.3.2.e20
- Fan w., Geerts F., Jia X. and Kementsietsidis A. (2008) “Conditional functional dependencies for capturing data inconsistencies,” *TODS*, vol. 33, no. 2.

- Fernando B. and Umberto S. (2013). Aggregation operators for fuzzy ontologies, Volume 13, Issue 9, September 2013, Pages 3816–3830, DOI:10.1016/j.asoc.2013.05.008
- Fernández M., Gomez-Perez A. and Juristo N. (1997). "METHONTOLOGY: From ontological art towards ontological engineering, Stanford Univ.," Stanford, CA, USA, Tech. Rep. SS-97-06, 1997.
- Frank V. H. (2014). Handbook Of Knowledge Representation: Communication, Information science; ISBN: eISBN 9781467222433
- Fudholi D. H., Maneerat N., Varakulsiripunth R. and Kato Y., (2009). "Application of Protégé, SWRL and SQWRL in fuzzy ontology-based menu recommendation," in Proc. Int. Symp. Intell. Process. Commun. Syst. (ISPACS), pp. 631-634.
- Francky C. S., Wuytack G.E. (2013) Custom Memory Management Methodology: Exploration of Memory Organisation for Embedded Multimedia System Design, de Greef Florin Banica Lode Nachtergaele Arnout Vandecappelle March 9, 2013, ISBN: 9781475728491
- Galanter, W.L., Hier, D.B., Jao, C., & Sarne, D. (2010). Computerized physician order entry of medications and clinical decision support can improve problem list documentation compliance. *Int J Med Inform*, 79(5), 332-338. doi: 10.1016/j.ijmedinf.2008.05.005
- Gartner (2007). "Forecast: Data quality tools, worldwide, 2006-2011,"
- Gerard FL., Gabriel O., Jonathan G., Alexey V., Mary H., Pierre G., Gene W., Gary G., Nigel Q., Michiel B., Scott P., Sim R., Noha G., Robert K. and Andrew H. (2013) – Integrated environmental modeling: A vision and roadmap for the future, Volume 39, January 2013, Pages 3–23, DOI:10.1016/j.envsoft.2012.09.006
- Ghorbel H., Bahri A. and Bouaziz R. (2010). "Fuzzy ontologies building method: Fuzzy ontomethodology," in Proc. Annu. Meeting North Amer. Fuzzy Inf. Process. Soc. (NAFIPS), 2010, pp. 1-8.

- Goddard, W. & Melville, S. (2004) "Research Methodology: An Introduction" 2nd edition, Blackwell Publishing
- Gregor S. and Hevner AR. (2013). POSITIONING AND PRESENTING DESIGN SCIENCE RESEARCH FOR MAXIMUM IMPACT. MIS Quarterly Vol. 37 No. 2, pp. 337-355/June 2013
- Grosshandler JA, Tulbert B., Kaufmann MD., Bhatia A. and Brodell RT. (2010) "The electronic medical record in dermatology," Archives of Dermatology, vol. 146, no. 9, pp. 1031-1036, 2010, 10.1001/archdermatol.2010.229.
- Gilson, A., Giraldi, Fabio, P., Bruno, S., Vinicius, F., and Dutra, M. L. (2005). Data Integration Middleware System for Scientific Visualization. , RJ, ZIP 25651-070, Brazil.
- Granquist, L. and Kovar, J.G. (1997): Editing of Survey Data: How much is enough? in Survey Measurement and Process Quality, New York: Wiley,
- Gribble, S., Halevy, A., Ives, Z., Rodrig, M. and Suciu, D. (2001). What can databases do for peer-to-peer? In Proceedings of the Fourth International Workshop on the Web and Databases, WebDB.
- Guba, E. and Lincoln, Y. (1998). – Competing paradigms in qualitative research, in N Denzin & Y Lincoln (eds), The landscape of qualitative research – theories and issues, Sage Publications, California, pp. 195-220
- Guba E., & Lincoln, Y., (1994). Competing paradigms in qualitative research. In Denzin, N. & Lincoln, Y (Eds.) Handbook on qualitative research. Thousand Oaks, Ca: Sage. 105-118.
- Hai, B. T., Trong, H. D. and Ngoc, T. N. (2013). *A Hybrid Method for Fuzzy Ontology Integration*. An International Journal, 44, pp. 133-154.
- Haibo C., Lingling X., Peng W., Peng Z. and Haibin Y. (2017). Discrete manufacturing ontology development. Industrial Technology (ICIT), 2017 IEEE International Conference on. DOI: 10.1109/ICIT.2017.7915568

- Hammond, W. E. (2002). Overview of Health Care Data Standards. Commissioned paper for IOM Committee on Data Standards for Patient Safety.
- Hammond W.E., Jaffe C. and Kush RD. (2009) Healthcare standards development. The value of nurturing collaboration. J AHIMA; 80: 44–50; quiz 1–2.
- Hampe, Holly M., Keeling T., Fontana M. and Balcik D. (2017). Impacting Care and Treatment of the Burn Patient Conversion to Electronic Documentation. Critical Care Nursing Quarterly. 40(1):8-15, January/March 2017. DOI: 10.1097/CNQ.0000000000000135
- Han PK, Klein WM, Arora NK. Varieties of uncertainty in health care: a conceptual taxonomy. Med Decis Making 2011;31:828–38; DOI: 10.1177/0272989X11393976
- Hartwood M., Procter R., Rouncefield M. and Slack R. (2003). Making a Case in Medical Work: Implications for the Electronic Medical Record. Comput Supported Coop Work. 2003; 12:241–266. DOI: 10.1023/A:1025055829026
- HCPCS Code range. "HCPCS Codes" URL: <https://coder.aapc.com/hcpcs-codes> access: 2017-07-30.
- Heron, J. & Reason, P. (1997). A participatory inquiry paradigm. Qualitative Inquiry. 3 (3) 274-294.
- Heron, J. (1996) Co-operative inquiry. London: Sage.
- Herzog, Th. N., Scheuren, F.J. and Winkler, W.E. (2007). Data Quality and Record Linkage Techniques. DOI: <http://www.springer.com/978-0-387-69502-0>
- Hevner, A., March S., Park J. and Ram S. (2004). "Design Science in Information Systems Research".Affiliated Journals › MISQ › Vol. 28.
- Hevner, A., & Chatterjee, S. (2010). Design Research in Information Systems Theory and Practice. Springer.
- Hoberman S. (2009). Data Modeling Made Simple: A Practical Guide for Business and IT Professionals. 2nd ed. Bradley Beach, NJ: Technics Publications; 2009.

- Hoffman S. and Podgurski A. (2009). "Finding a Cure: The Case for Regulation and Oversight of Electronic Health Record Systems," 120; Case Legal Studies Research Paper No. 08-13; Harvard Journal of Law and Technology, Vol. 22, No. 1, 2008 available at: [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1122426](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1122426)
- Hogan, W. R., and Wagner, M. M. (1997). Accuracy of data in computerbased patient records. *Journal of the American Medical Informatics Association* 4(5):342–355.
- Holly J. L., Dean F. S., Luci K. L., Michael L. P., Jacqueline A. P. and Reuben R. M. (2013). Understanding differences in electronic health record (EHR) use: linking individual physicians' perceptions of uncertainty and EHR use patterns in ambulatory care; 2014 Jan; 21(1): 73–81. Published online 2013 May 22. doi: 10.1136/amiajnl-2012-00137
- Hongyi M. and Yang S. (2017) - A Way to Understand Inpatients Based on the Electronic Medical Records in the Big Data Environment. DOI: <https://doi.org/10.1155/2017/9185686>
- HIMSS. EHR definition. [Accessed August 15, 2017]. [http://www.himss.org/ASP/topics\\_ehr.asp](http://www.himss.org/ASP/topics_ehr.asp) .
- Hirak K., Hasin AA., Nazrul H. Swarup R. and Dhruva KB. (2016) - Big data analytics in bioinformatics: architectures, techniques, tools and issues. DOI: 10.1007/s13721-016-0135-4
- Hirschtick R. (2006). A piece of my mind. Copy-and-paste. *JAMA*. 295(20):2335–2336. doi: 10.1001/jama.295.20.2335
- Hsiao CJ. H. E. (2012) Use and Characteristics of Electronic Health Record Systems Among Office-Based Physician Practices, United States, 2001-2012. US Department of Health; Human Services, Centers for Disease Control; Prevention, National Center for Health Statistics, United States.



- livari, J. (2007). "A Paradigmatic Analysis of Information Systems as a Design Science, *Scandinavian Journal of Information Systems*, 19(2), 2007.
- Illhoi, Y., Patricia, A., Miroslav, M., Keila, PH., Rajitha, G., Jia-Fu, C. and Lei, H. (2012). *Data Mining in Healthcare and Biomedicine: A Survey of the Literature*, DOI: 10.1007/s10916-011-9710-5
- Iraklis L. and Aykut Ö. (2015) - Selection of the best maintenance approach in the maritime industry under fuzzy multiple attributive group decision-making environment. DOI: 10.1177/1475090215569819
- Jackson N., Atar D., Borentain M., Breithardt G., Eickels V. M., Endres M., Fraass U., Friede T., Hannachi H., Janmohamed S., Kreuzer J., Landray M., Lautsch D., Floch C. L., Mol P., Naci H., Samani N., Svensson A., Thorstensen C., Tijssen J., Vandzhura V., Zalewski A. and Kirchhof P. (2016). Improving clinical trials for cardiovascular diseases: a position paper from the Cardiovascular Roundtable of the European Society of Cardiology. *Eur Heart J.* 2016;37:747–754. doi: 10.1093/eurheartj/ehv213.
- Jarrar M. and Meersman R. (2009). "Ontology engineering - The DOGMA approach," in *Adv. Web Semantics I*, vol. 4891, T. S. Dillon, E. Chang, R. Meersman, and K. Sycara, Eds. Berlin, Germany: Springer, 2009, pp. 7-34.
- Jason JS., Jennifer H., Nancy RW. (2016) - Function-specific Design Principles for the Electronic Health Record. *Function-specific Design Principles for the Electronic Health Record*. DOI 10.1177/1541931213601133
- Jayant, M., Shawn, R., Jeffery, SC., Xin (Luna), D., David, K., Cong, Y. and Alon H. (2007). – *Webscale Data Integration: You can only afford to Pay As You Go*. CIDR.
- Jennifer P., Carolyn McG., Nathan P. and Andrew J. (2015) - Enabling the integration of clinical event and physiological data for real-time and retrospective analysis. DOI: 10.1007/s10257-014-0232-9

- Jens, M., Stefan, O., Daniel, F., Christoph, H., Janina, K. and Wolfgang, H. (2012). Efficient data management in a large-scale epidemiology research project, *Computer Methods and Programs in Biomedicine*; <http://dx.doi.org/10.1016/j.cmpb.2010.12.016>
- Jha A, DesRoches CM, Campbell EG, Donelan K., Rao SR, Ferris TF, Shields A., Rosenbaum S. and Blumenthal D. (2009). Use of Electronic Health Records in US hospitals. *N Engl J Med.*; 360:1628–1638. DOI: 10.1056/NEJMsa0900592.
- Jian M., Shujun M. and Zhao Y. (2009). – RESTful Web Services: a Solution for Distributed Data Integration. 978-1-4244-4507-3/09/\$25.00 ©2009 IEEE
- Jiawei H., Micheline K., Jian P. (2011) - Data Mining: Concepts and Techniques: Edition 3, June 9, 2011, ISBN: 9780123814807
- Joachims T. (2002). Optimizing search engines using clickthrough data. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*; 2002. pp. 133–142. ACM. DOI: 10.1145/775047.775067
- John Dudovskiy. (2016). *The Ultimate Guide to Writing a Dissertation in Business Studies: A Step-by-Step Assistance*
- John P., Julien M., Laurent L. and Jaakko L. (2012) - Quality Analysis of Sensors Data for Personal Health Records on Mobile Devices. Chapter: Pervasive Health Knowledge Management, Part of the series *Healthcare Delivery in the Information Age* pp 103-133, DOI: 10.1007/978-1-4614-4514-2\_10
- John W. Creswell (2013) *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, March 14, 2013, ISBN: 9781483321479
- Johnson K.E., Kamineneni A., Fuller S., Olmstead D. and Wernli K.J. (2014). How the Provenance of Electronic Health Record Data Matters for Research: A Case Example Using System Mapping. *eGEMs (Generating Evidence & Methods to improve patient outcomes) [Internet]* 2014. Apr 16, [cited 2016 Mar 13];2(1).

- Jonas P., Dmitry Il., Sergei O. K., and Guido D. (2013). - Formal concept analysis in knowledge processing: A survey on applications, Volume 40, Issue 16, 15 November 2013, Pages 6538–6560, DOI: 10.1016/j.eswa.2013.05.009
- Jones S. S., Heaton P., Friedberg M. W. and Schneider E. C. (2011). “Today's ‘Meaningful Use’ Standard for Medication Orders by Hospitals May Save Few Lives; Later Levels May Do More” Health Affairs. 2011; 30(10):2005–12.
- Juhani I. (2007). “A Paradigmatic Analysis of Information Systems as a Design Science, Scandinavian Journal of Information Systems, 19(2), 2007. Available at: <https://www.researchgate.net/publication/201168950>
- Juran J.M. and Godfrey A. B. (1999). Juran’s Quality Control Handbook. 5th ed. New York: McGraw-Hill; ISBN 0-07-034003-X
- Kamil, K. and Kazimierz, S. (2010). Transparent Integration of Distributed Resources within Object-Oriented Database Grid. Faculty of Electrical, Electronic, Computer and Control Engineering of the Technical University of Lodz
- Karan R. S., Asif A. T., Shubham S. P., Sudeep S. K. and Sarang S. (2016). Establishment of Electronic Health Records in Developing Countries; International Journal of Computer Applications (0975 – 8887); Volume 136 – No.11, February 2016
- Keith W., Hipel and McLeod A. I. (1994). Time Series Modelling of Water Resources and Environmental Systems. ISBN: 978-0-444-89270-6
- Kuechler B, Vaishnavi V. (2008). "On theory development in design science research: Anatomy of a research project". European Journal of Information Systems. 17(5): 489–504.
- Kim TY. and Matney SA. (2017) - Informatics-Related Standards and Standards-Setting Organizations; Health Informatics: An Interprofessional Approach, Second edition. ISBN: 978-0-323-40231-6

Kok OM, Başoğlu N (2012) "Exploring the success systems adoption." IEEE 1: 1471-1477.

Kopcke F., Trinczek B., Majeed R. W., Schreiweis B., Wenk J., Leusch T., Ganslandt T., Ohmann C., Bergh B., Rohrig R., Dugas M. and Prokosch H. U. (2013) Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. BMC Med Inform Decis Mak. 2013;13:37. doi: 10.1186/1472-6947-13-37.

Kothari C. R. (2004). Research Methodology, Methods and Techniques. - ISBN (13): 978-81-224-2488-1

Kotis K. and Vouros G. A. (2006). "Human-centered ontology engineering: The HCOME methodology," Knowl. Inf. Syst., vol. 10, no. 1, pp. 109-131, Jul. 2006.

Krusinska E., Slowinski and Stefanowski J. (1992). Discriminant versus rough set approach to vague data analysis. *Application Stochastic Models Data Analysis* 8:43 56.

Kuhn, T. (1962). The Structure of Scientific Revolutions. University of Chicago Press.

Kukafka R., Ancker JS., Chan C., Chelico J., Khan S., Mortoti S., Natarajan K., Presley K. and Stephens K. (2007) - Redesigning electronic health record systems to support public health. J Biomed Inform.40(4):398-409.

Kumar R. and Tomkins A. (2009). A characterization of online search behavior. IEEE Data Eng. Bull. 2009;32(2):3-11. DOI: 10.1145/1772690.1772748

Lamine B., Amir H., Parisa G., Emmanuel A., Samy T. and Mohamed H. (2016). Hybrid Reasoning-Based Medical Platform to Assist Clinicians in their Clinical Reasoning Process Submitted on 27 Jan 2016

Lakshmana G. N. V., Muralikrishnan S. and Sivaraman G. (2011). Multi-Criteria decision-making method based on interval-valued intuitionistic fuzzy sets. Expert Syst. 38, 1464-1467. DOI: <https://doi.org/10.1016/j.eswa.2010.07.055>

- Lee, N. & Lings, I. (2008) "Doing Business Research: A Guide to Theory and Practice"  
SAGE Publications
- Lehrman W. G., Elliott M. N., Goldstein E. (2010). "Characteristics of Hospitals  
Demonstrating Superior Performance in Patient Experience and Clinical  
Process Measures of Care" *Medical Care*. 2010;67(1):38–55
- Leonardo, C. B., Jéssica, O. de S., Fábio, R. J., Caio, S. C., Márcio, R. de C., Vânia, P.  
de A., N. and Regina, B. de A. (2015). Methodology for Data and Information  
Quality Assessment in the Context of Emergency Situational Awareness, DOI:  
10.1007/s10209-016-0473-0
- Liaw S, Taggart J, Dennis S, Yeo A: Data quality and fitness for purpose of routinely  
collected data – a general practice case study from an electronic Practice-  
Based Research Network (ePBRN). In AMIA 2011 Annual Symposium Improving  
Health: Informatics and IT Changing the World; October 22–26, 2011.  
Washington DC, US: AMIA; 2011:785–94. PMID: 22195136
- LexisNexis Risk Solution. (2017) Why Data Quality is the Greatest Challenge and  
Opportunity for Health Care.  
[https://www.lexisnexis.com/risk/downloads/whitepaper/Data-Quality-  
POV.pdf](https://www.lexisnexis.com/risk/downloads/whitepaper/Data-Quality-POV.pdf) Last access: 2017-09-09
- Li, Y. (2016) "Expatriate Manager's Adaption and Knowledge Acquisition: Personal  
Development in Multi-National Companies in China" Springer Publications
- Lijun M., Steve EB., Philip VT., Michael W. McD. and Penny KS. (2017) - Inherent  
functional dependence among cochlear dose surrogates for stereotactic  
radiosurgery of vestibular schwannomas.  
<http://dx.doi.org/10.1016/j.prro.2016.08.005>
- Lu Z., Kim W. and Wilbur W.J. (2009) Evaluation of query expansion using MeSH in  
PubMed. *Information retrieval*. 2009;12(1):69–80. DOI: 10.1007/s10791-008-  
9074-8

- Lukasiewicz T. and Straccia U. (2008) "Managing uncertainty and vagueness in description logics for the Semantic Web," Web Semant. Sci. Services Agents World Wide Web
- Maedche, A., Motik, B., Silva, N. and Volz, R. (2002). Mafra - a mapping framework for distributed ontologies. DOI: 10.1007/3-540-45810-7\_23, Print ISBN: 978-3-540-44268-4, Online ISBN: 978-3-540-45810
- Maio DP, White LM, Bleakney R., Menezes RJ and Theodoropoulos J. (2014). Diagnostic Accuracy of an iPhone DICOM Viewer for the Interpretation of Magnetic Resonance Imaging of the Knee. *Clinical Journal of Sport Medicine*. 24, 4(2014), 308-314. DOI:10.1097/JSM.0000000000000005.
- Malcolm P. Atkinson, Peter Buneman, Ronald Morrison - Data Types and Persistence, December 6, 2012, ISBN: 9783642615566
- Mancia G., Fagard R., Narkiewicz K., Redón J., Zanchetti A., Böhm M., Christiaens T., Cifkova R., De Backer G., Dominiczak A., Galderisi M., Grobbee DE., Jaarsma T., Kirchhof P., Kjeldsen SE., Laurent S., Manolis AJ., Nilsson PM., Ruilope LM., Schmieder RE., Sirnes PA., Sleight P., Viigimaa M., Waeber B. and Zannad F. (2013). 2013 ESH/ESC Guidelines for the management of arterial hypertension: the Task Force for the management of arterial hypertension of the European Society of Hypertension (ESH) and of the European Society of Cardiology (ESC). DOI: 10.1097/01.hjh.0000431740.32696.cc.
- March, S. T. and Storey, V. C. (2008). Design Science in the Information Systems Discipline: An introduction to the special issue on design science research, *MIS Quarterly*, Vol. 32(4), pp. 725–730.
- March ST. and Smith GF. (1995). Design and natural science research on information technology. *Decision Support Systems*, 15(4), pp. 251–266.
- Mario LV., Boyang L., Christopher V. and Denys P. (2016) - Documenting Database Usages and Schema Constraints in Database-Centric Applications.

Proceedings of the 25th International Symposium on Software Testing and Analysis ISBN: 978-1-4503-4390-9 doi:10.1145/2931037.2931072

Markle Foundation (2011) "Markle survey: the public and doctors agree on importance of specific privacy protections for health I.T", Markle Survey on Health in a Networked Life.

Markel A. (2010). Copy and Paste of Electronic Health Records: A Modern Medical Illness. *Am J Med.* 123(5):e9. DOI: 10.1016/j.amjmed.2009.10.012.

Marne DV and Simone B. (2017) An Action Design Research Approach within Enterprise Engineering. DOI 10.1007/s11213-016-9390-7

Marta Z. and Diego GS. (2013). – A service oriented architecture to provide data mining services for non-expert data miners, Volume 55, Issue 1, April 2013, Pages 399–411, DOI:10.1016/j.dss. 2012.05.045

Martin RC. Juuso IB., Lesley HC., Sylvie D., Ian F., Fleur F., Samantha G., Salim J., Jörg K., Mark L., Alexander M., Seleen O., Jill PP., Mary RS., Wendy GS., Martin T., Faiez Z. and Andrew Z. (2016) - Electronic health records to facilitate clinical research. DOI: 10.1007/s00392-016-1025-6.

Matthias J., Maurizio L., Yannis V., Panos V. (2013) - Fundamentals of Data Warehouses: Edition 2 March 9, 2013, SBN: 9783662051535

Mary D. B. and Harry R. (2011). AHIMA. "Protecting Patient Information after a Facility Closure (2011 update)." *Journal of AHIMA* (Updated August 2011)

McLoughlin IP., Garrety K. and Wilson R (2017) - The Digitalization of Healthcare: Electronic Records and the Disruption of Moral Orders. ISBN: 978-0-19-874413-9

Mead CN. (2006). Data interchange standards in healthcare IT—computable semantic interoperability: now possible but still difficult, do we really need a better mousetrap? *J Healthc Inf Manag*; 20: 71–8.

- Mei Q., Zhou D. and Church K. (2008). Web Age, Information Management. Query suggestion using hitting time. Proceeding of the 17th ACM conference on Information and knowledge management; 2008. pp. 469–478. ACM. DOI: 10.1007/978-3-642-38562-9
- Mendel J. M. (2016). A comparison of three approaches for estimating (synthesizing) an interval type-2 fuzzy set model of a linguistic term for computing with words. *Granul. Comput.* 1, 59–69. DOI: <http://dx.doi.org/10.1007/s41066-015-0009-7>
- Mertens, D.M. (2005). *Research methods in education and psychology: Integrating diversity with quantitative and qualitative approaches.* (2nd ed.)
- Mingers, J. (2003). – A classification of the philosophical assumptions of management science methods. *Journal of the Operational Research Society*, Vol. 54, PP. 559–70.
- Mohamed K. (2013) - Barriers to Health Information Systems and Electronic Medical Records Implementation. A Field Study of Saudi Arabian Hospitals. *Procedia Computer Science* 21 ( 2013 ) 335 – 342  
<http://dx.doi.org/10.1016/j.procs.2013.09.044>
- Mohamedali F ., Oussena S. and Roth-Berghofer T. (2016) - Application of Complex Event Processing Techniques to Big Data Related to Healthcare: A Systematic Literature. ISBN9781522502944
- Mohcine M., Driss B. and Cui T. (2016) - Temporal data representation, normalization, extraction, and reasoning: A review from clinical domain.  
<http://dx.doi.org/10.1016/j.cmpb.2016.02.007>
- Momoh, A., Roy, R., and Shehab, E. (2010). Challenges in enterprise resource planning implementation: state-of-the-art. *Business Process Management Journal* Vol. 16 No. 4, Emerald Group Publishing Limited, DOI 10.1108/14637151011065919, pp. 537-565.



- Monette, DR, Sullivan, TJ, DeJong, CR, 2005, Applied Social Research. A Tool for the Human Services, 6th edition
- Moon, F., Juan, M., Campos, I. and Luis, M. (2016). Personalizing xml information retrieval; ISBN: 9788491250548
- Morgan, DL. (2007). Paradigms lost and paradigms regained. Journal of Mixed Methods Research. 1(1), 48-76.
- Näppilä, T. (2013) – Serving Sophisticated Ad Hoc Information Needs Based on Beforehand Unknown, Autonomous, and Heterogeneous XML Data Sources, 2013, ISBN: 978-951-44-9285-3
- Narender R., Yunlin Z., Taher K. and Todd B. (2014) – Leveraging advanced data analytics, machine learning, and metrology models to enable critical dimension metrology solutions for advanced integrated circuit nodes, 13(4), 041415 (Dec 29, 2014). DOI:10.1117/1.JMM.13.4.041415
- Natalya F. N. and Deborah L. MG. (2001) "Ontology development 101: A guide to creating your first ontology" Stanford Med. Informat., Stanford, CA, USA, Tech. Rep. SMI-2001-0880, 2001
- Niclas S., Joana V., Rong C., Hans B. and Sabine K. (2016) - How to improve vital sign data quality for use in clinical decision support systems? A qualitative study in nine Swedish emergency departments; DOI: 10.1186/s12911-016-0305-4
- Nirase, F. A., Azreen, A., Shyamala, D., Masrah, A. and Azmi, M. (2016). An integrated method of associative classification and neuro-fuzzy approach for effective mammographic classification. DOI: 10.1007/s00521-016-2290-z
- Nicole GW. and Chunhua W. (2013) Methods and dimensions of electronic health record data quality assessment: enabling re-use for clinical research, Journal of the American Medical Informatics Association, Volume 20, Issue 1, 1 January 2013, Pages 144–151, <https://doi.org/10.1136/amiajnl-2011-000681>

- Niyato, D., Xiao, L. and Wang, P. 2011. Machine-to-machine communications for home energy management system in smart grid. *Communications Magazine*, IEEE.
- Noteboom CB, Motorny SP, Qureshi S, Sarnikar S (2014) "Meaningful use of electronic health records for physician collaboration: A patient centred health care perspective." *IEEE* 1: 656-666.
- Noy N. F. and McGuinness D. L. (2001). "Ontology development 101: A guide to creating your First ontology," *Stanford Med. Informat.*, Stanford, CA, USA, Tech. Rep. SMI-2001-0880, 2001.
- Nunamaker JF., Chen M. and Purdin TDM. (1991). Systems Development in Information Systems Research. *Journal of Management Information Systems* 7, 3, 89-106.
- Nuno, L., Seyma, N. S. and Jorge, B. (2015). A Survey on Data Quality: Classifying Poor Data. Conference Paper. DOI: 10.1109/PRDC.2015.41
- O'Donnell H., Kaushal R. and Siegler E. (2008). Physicians attitudes towards copy and pasting in electronic note writing. *AMIA Annu Symp Proc*.
- O'Leary Z. (2004). "The essential guide to doing research". Sage Publication. ISBN 0-7619-4199-1
- Orrill, C. H., Hannafin, M. J. and Glazer, E. M. (2003). – Disciplined inquiry and the study of emerging technology. In D. H. Jonassen (Ed.), *Handbook of research for educational communications and technology* (2<sup>nd</sup> ed., pp. 335)
- O'zsu M. T. and Valduriez P. (1999). *Principles of Distributed Database Systems* (2nd edition). Prentice-Hall, 1999.
- Özgür, N. B. (2007). *An Intelligent Fuzzy Object Oriented Database Framework for Video Database Applications*. Computer Engineering Department, METU, Ankara, Turkey.

- Park J. H., Park I. Y., Kwun Y. C. and Tan X. (2011). Extension of the TOPSIS method for decision making problems under interval-valued intuitionistic fuzzy environment. *Appl. Math. Model.* 2011, 35, 2544–2556. DOI: <https://doi.org/10.1016/j.apm.2010.11.025>
- Paul J. and Erik P. (2014). *Research Strategies and Methods*, Chapter An Introduction to Design Science pp 39-73. DOI: 10.1007/978-3-319-10632-8\_3
- Paul JA., Gwendolyn B. (2017). Practical Steps for the Utilization of Action Research in Your Organization: A qualitative approach for non-academic research. *International Journal of Human Resource Studies* ISSN 2162-3058 2017, Vol. 7, No. 2
- PAWLAK Z. (1982). "Rough sets". *International Journal of Parallel Programming*. 11 (5): 341–356. doi:10.1007/BF01001956
- PAWLAK Z. (1991). *Rough Sets: Theoretical Aspects of Reasoning About Data*. Dordrecht: Kluwer Academic Publishing. ISBN 0-7923-1472-7.
- PAWLAK Z. (2010). *ROUGH SET THEORY AND ITS APPLICATIONS TO DATA ANALYSIS*. DOI: <http://dx.doi.org/10.1080/019697298125470>
- Pertti Järvinen (2007) *Action Research is Similar to Design Science*. DOI 10.1007/s11135-005-5427-1
- Pérez I. J., Wikström R., Mezei J., Carlsson C. and Herrera-Viedma E. (2013). A new consensus model for group decision making using fuzzy ontology, September 2013, Volume 17, Issue 9, pp 1617-1627, DOI: 10.1007/s00500-012-0975-5, Print ISSN: 1432-7643, Online ISSN: 1433-7479
- Peter J. and Han B. (2015). A Multilevel Design Model: the mutual relationship between product-service system development and societal change processes <https://doi.org/10.1016/j.jclepro.2014.06.043>

- Peter S., Claudia P. and Don E. D. (2009). CRITICAL ISSUES FOR ELECTRONIC HEALTH RECORDS, CONSIDERATIONS FROM AN EXPERT WORKSHOP, The Nuffield Trust for research and policy studies in health services.
- Philip C CL. and Chun-Yang Z. (2014) – Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, 2014, DOI:10.1016/j.ins.2014.01.015
- Philipp O., Olga L., Marten S. and Udo B. (2009). Outline of a Design Science Research Process. DESRIST'09, May 7-8, 2009, Malvern, PA, USA. Copyright 2009 ACM 978-1-60558-408-9/09/05...\$5.00
- Phillips Win, Fleming David. Ethical Concerns in the Use of Electronic Medical Records. *Modern Medicine*. 2009;106:328.
- Phillips W. and Fleming D. (2009). "Ethical Concerns in the Use of Electronic Medical Records." *Mo Med*. 2009 Sep-Oct;106(5):328-33.
- Peirce, C. S. (1931-1935). *Collected Papers of Charles Sanders Peirce*, Vols. 1-6, Harshorne, C. and P. Weiss, Eds. Cambridge, MA, Harvard University Press.
- Polonsky, M. J., & Waller, D. S. (2005). *Designing and managing a research project: A business student's guide*. Thousand Oaks, CA: Sage Publications.
- Popper, K. R. (1961). – *The logic of scientific discovery*. New York: Science Editions. ISBN 0-415-27843-0 (hbk), ISBN 0-415-27844-9 (pbk)
- Qiu RC. and Antonik P. (2017) - *Smart Grid Using Big Data Analytics: A Random Matrix Theory Approach*. ISBN: 9781118494059
- Ralph K. and Margy R. (2013). - *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, 01 Jul 2013, ISBN: 9781118732281
- Raphae C., Michael E. and Noémie E. (2013). Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation

strategies; Published online 2013 Jan 16. DOI: 10.1186/1471-2105-14-10; PMCID: PMC3599108

Ray M. C., Robert JK., Young OK. (2014) – Understanding the paradigm shift to computational social science in the presence of big data, Volume 63, July 2014, Pages 67–80, DOI:10.1016/j.dss.2013.08.008

Reb, Y., Pan, J. Z., and Zhao, Y. 2010. *Soundness Preserving Approximation for TBox Reasoning*. In the Proc. of the 25th AAAI Conference Conference (AAAI2010).

Reason, P. (Ed). (1988) *Human Inquiry in Action*. London: Sage.

Reason, P. (Ed). (1994). *Participation in Human Inquiry*. London: Sage

Reigeluth, C. M., and Frick, T. W. (1999). - Formative research: A methodology for creating and improving design theories. In C. M. Reigeluth (Ed.), *Instructional-design theories and models* (Vol. II, pp. 633–651).

Rezaeibagha F. (2013). “Privacy and data security of electronic patient records (ep) sharing”, Master of Science.

Richard A. Bloomfield JR., Felipe PW., Joshua CM. and Kenneth DM. (2016) - Opening the Duke electronic health record to apps: Implementing SMART on FHIR. <http://dx.doi.org/10.1016/j.ijmedinf.2016.12.005>

Richard L. Baskerville A. and Trevor WH. (2016). A Critical Perspective on Action Research as a Method for Information Systems Research. DOI: 10.1007/978-3-319-29269-4\_7

Risto, S., Olli J., Hanna, K. V. and Harri, H. (2011) Managing one master data – challenges and preconditions. Journal: *Industrial Management & Data Systems*, Volume: 111, Number: 1, Copyright © Emerald Group Publishing Limited ISSN: 0263-5577, pp. 146-162.

Rowley J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *J Inf Sci* 2007; 33(2):163–80.

- Roy H. (2010) – Data Compression in digital system, 2012, ISBN: 978-1-4613-7764-1, DOI: 10.1007/978-1-4615-6031-9
- Sallie K., Gizem K., Mark O., Aaron S. and Stephanie S. (2016) - The Evolution of Data Quality: Understanding the Transdisciplinary Origins of Data Quality Concepts and Approaches. DOI:10.1146/annurev-statistics-060116-054114
- Saiod AK, Darelle VG and Veldsman A. (2017). Electronic Health Records: Benefits and Challenges for Data Quality. © Springer International Publishing AG 2017. DOI: 10.1007/978-3-319-58280-1\_6
- Saiod AK and Darelle VG. (2019a). Novel Hybrid Approaches for Big Data Recommendations (Chapter Five). Book details: Big Data Recommender Systems Recent Trends and Advances. Editors: Osman Khalid, Samee U. Khan, Albert Y. Zomaya. Product Code: PBPC035A, ISBN: 978-1-78561-975-5
- Saiod AK and Darelle VG. (2019 b). Cloud Integration for eHealth Data. ICICIS-2019: Proceedings of 4th International Conference on the Internet, Cyber Security and Information Systems 2019. Kalpa Publications in Computing, Volume 12, 2019, Pages 300-309.
- Sanchez, E. and Yamanoi, T. (2006). Fuzzy Ontologies for the Semantic Web. 7th International Conference on Flexible Query Answering Systems (FQAS 2006), Milan, Italy. Pp. 7–10.
- SARFRAZ NB., MERVAT AB. and MUHAMMAD NB. (2016) - IDENTIFYING AND ANALYZING THE TRANSIENT AND PERMANENT BARRIERS FOR BIG DATA. Journal of Engineering Science and Technology Vol. 11, No. 12 (2016) 1793 - 1807© School of Engineering, Taylor’s University
- Saunders, M., Lewis, P. & Thornhill, A. (2012) “Research Methods for Business Students” 6th edition, Pearson Education Limited
- Selcuk CK. and Maria LS. (2010) – Data management for Multimedia Retrieval, 2010, ISBN: 978-0-521-88739-7

- Schmitt KF. and Wofford DA. (2002) – Financial analysis projects clear returns from electronic medical records. *Healthc Financ Manage.* 56(1):52–57.
- Sha F. and Guo-bing F. (2016). A Multiple Attribute Decision-Making Method Based on Exponential Fuzzy Numbers. DOI: 10.3390/mca21020019
- Shaker, E. S., Mohammed, E. and Riad A.M. (2015). A fuzzy-ontology-oriented case-based reasoning framework for semantic diabetes diagnosis. Volume 65, Issue 3, Pages 179–208, DOI: <http://dx.doi.org/10.1016/j.artmed.2015.08.003>
- Shannon T. (2009). "National clinical programmes: Aligning process improvements with information technologies. strategic framework proposal," University College London, Report, 2009.
- Shirley G., Oliver M. and Stefan S. (2013) - Reflection, Abstraction And Theorizing In Design And Development Research, ECIS 2013 Completed Research. Paper 74.
- Sheenam, Mishra D., Zhang D. and Mahalik NP. (2016) - Bio-inspired technology systems and data analysis methods for classification and subsequent decision making intended for automated systems, DOI: <http://dx.doi.org/10.1504/IJCVR.2016.079398>
- Shvaiko, P., Tas L., and Euzenat, J. (2013) – Ontology Matching: State of the Art and Future Challenges, 2013, DOI: 10.1109/TKDE.2011.253
- Siegler E.L. and Adelman R. (2009) Copy and paste: a remediable hazard of electronic health records. *Am J Med*; 122:495–6
- Silberschatz, A., Korth, H., and Sudarshan, S. (2006) *Database System Concepts*, Beijing: Higher Education Press
- Silow-Carrol S., Edwards JN and Rodin D. (2012) "Using electronic health records to improve quality and efficiency: the experiences of leading hospitals," *The Commonwealth Fund*, vol. 1608, no. 17, pp. 1-40, 2012.

- Skolimowski, H. (1992). *Living philosophy: Eco-philosophy as a tree of life*. London: Arkana
- Skowron A. and C. Rauszer. (1992). The discernibility matrices and functions. In *Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory*, ed. R. Slowinski, 331-362. Boston: Kluwer Academic Publishers.
- Sonia B., Domenico B., Francesco G. and Mirko O. (2011) – Data Integration, Pages pp 441-476, Copyright 2011, DOI: 10.1007/978-3-642-15865-0\_14, Print ISBN: 978-3-642-15864-3, Online ISBN: 978-3-642-15865-0
- Stetson P. D., Morrison F.P. and Bakken S. (2008). Preliminary development of the physician document quality instrument. *J Am Med Inform Assoc*; 15:534–41
- Steven V. (2008). Accessibility and clarity: The most neglected dimensions of quality? Committee for the Coordination of Statistical Activities. Conference on Data Quality for International Organizations, Rome, Italy
- Stoilos G., Straccia U., Stamou G. and Pan J. Z. (2006). "General concept inclusions in fuzzy description logics," in *Proc. 17th Eur. Conf. Artif. Intell. (ECAI)*, Riva del Garda, Italy, p. 456.
- Straccia U. (2009). "A minimal deductive system for general fuzzy RDF," in *Web Reasoning Rule Syst.*, vol. 5837, A. Polleres and T. Swift, Eds. Berlin, Germany: Springer, pp. 166-181.
- Straccia U. (2013). *Foundations of Fuzzy Logic and Semantic Web Languages*. London, U.K.: Chapman & Hall
- Steyn, K., 2006. Hypertension in South Africa. In: Steyn, K., Fourie, J., Temple, N.(Eds.), *Chronic Diseases of Lifestyle in South Africa Since 1995e2005*. South African Medical Research Council, Cape Town, pp. 80e96.



- Suarez-Figueroa M. C. (2010) ``NeOn methodology for building ontology networks: Specification, scheduling and re-use," M.S. thesis, Universidad Polit cnica de Madrid, Tech. Univ. of Madrid, Spain, 2010.
- Sultan A, Elankayer S. and Vallipuram M. (2016) - A survey on data leakage prevention systems. Journal of Network and Computer Applications. DOI: <http://dx.doi.org/10.1016/j.jnca.2016.01.008>
- Sure Y., Staab S. and Studer R. (2004). ``On-to-knowledge methodology (OTKM)," in Handbook Ontologies, S. Staab and R. Studer, Eds. Berlin, Germany: Springer, 2004, pp. 117-132.
- Tania, C., Genoveva, V., Jose, L. and Zechinelli, M. (2004). - Building Multimedia Data Warehouses from Distributed Data. e-Gnosis [online], Vol. 2, Art.10.
- Tao CA., Arif K., Markus S. and Ganesh V. (2010) - iBLOB: Complex Object Management in Databases through Intelligent Binary Large Objects, 2010, DOI: 10.1007/978-3-642-16092-9\_10, Print ISBN: 978-3-642-16091-2, Online ISBN: 978-3-642-16092-9
- Terre B. and Durrheim K. (1999). - Research in practice. Cape Town: UCT Press
- Tho Q. T., Hui S. C., FongA. C. M. and Cao T. H. (2006). ``Automatic fuzzy ontology generation for semantic Web," IEEE Trans. Knowl. Data Eng., vol. 18, no. 6, pp. 842-856, Jun. 2006
- Toshiaki K., Mark DW., Gos M., Shuichi K., Atsuko Y., Mitsuteru N., Yasunori Y., Shinobu O., Kenta O. and 57 more (2013). - The 3rd DBCLS BioHackathon: improving life science data integration with Semantic Web technologies, Journal of Biomedical Semantics December 2013, DOI: 10.1186/2041-1480-4-6, Online ISSN: 2041-1480
- Toulmin, S., R. Rieke and A. Janik (1979). An Introduction to Reasoning. Macmillan, New York.

Umberto D'A., Miguel GE., Stefano M. and Ignazio S. (2015) - TMDs: Evolution, modeling, precision. DOI: <http://dx.doi.org/10.1051/epjconf/20158502003>

Umut H. I., Sait G. - A multiple attribute decision model to compare the firms' occupational health and safety management perspectives. Safety Science 91 (2017) 221–231  
<http://dx.doi.org/10.1016/j.ssci.2016.08.018>

Uthayan K. R., Anandha Mala G. S. - Hybrid Ontology for Semantic Information Retrieval Model Using Keyword Matching Indexing System, Volume 2015 (2015), Article ID 414910, 9 pages, <http://dx.doi.org/10.1155/2015/414910>

Vaishnavi, V. and Kuechler, W. (2004). "Design Research in Information Systems" January 20, 2004, last updated January 18, 2006. URL:

<http://www.isworld.org/Researchdesign/drisISworld.htm>

Vaishnavi, V., and Keuchler, W. (2013). Design Science Research in Information Systems. Retrieved Sept 14, 2014, from Association for Information Systems: <http://www.desrist.org/desgn-research-in-information-systems/>

Vaishnavi VK. and Kuechler W. (2015) Design science research methods and patterns: innovating information and communication technology. ISBN-13: 978-1-4987-1526-3(eBook - PDF)

Veli N. S., Dipak K., Pierre L., Alan R., Jean M. R., Karl A. S., Gyorgy S., Bedirhan U., Martti V. and Pieter E. Z. (2009). Semantic Interoperability for Better Health and Safer Healthcare. Deployment and Research Roadmap for Europe. Catalogue number: KK-80-09-453-EN-C; ISBN-13 : 978-92-79-11139-6; DOI : 10.2759/38514

Vladimir Z. and John G. (2015) - A systematic approach to reliability assessment in integrated databases, 14th Apr 2015, DOI 10.1007/s10844-015-0359-2

Vojt'as, P. (2006). A fuzzy EL Description aggregation for web consulting. Knowledge-Based Systems (IPMU 2006)

- Vrandecic D., Pinto S., Tempich C. and Sure Y. (2005). "The DILIGENT knowledge processes," J. Knowl. Manage., vol. 9, no. 5, pp. 85-96, Oct. 2005.
- Walker J, Leveille SG, Ngo L, Vodicka E, Darer JD, et al. (2012) "Inviting patients to read their doctor's notes: patients and doctors look ahead." *Anna Int Med* 155: 811-819.
- Walliman, N. S. & Walliman N. (2011). "Research methods: the basics" Taylor and Francis
- Wang RY. and Strong DM. (2015). Beyond Accuracy: What Data Quality Means to Data Consumers. DOI: <http://dx.doi.org/10.1080/07421222.1996.11518099>
- Wang J. Q. (2006). Multi-Criteria Interval Intuitionistic Fuzzy Decision-making Approach with Incomplete Certain Information. *Control Decis.* 21, 1253–1256, 1263.
- Wang, C., Zhang, J. and Qin, L. (2016). Design & Research of Legal Affairs Information Service Platform Based on UIMA and Semantics. ISSN: 2233-7857 IJFGCN
- Wei G.W. (2008). A Method of Interval-Valued Intuitionistic Fuzzy Multiple Attributes Decision Making with Incomplete Attribute Weight Information. *Chin. J. Manag.* 5, 208–211, 217.
- Weiskopf, N.G., & Weng, C. (2012). Methods and dimensions of electronic health record data quality assessment: enabling re-use for clinical research. *J Am Med Inform Assoc*, 20(1), 144-151. DOI: 10.1136/amianjnl-2011-00681
- Welman, J. C. and Kruger, S. J. 2003. *Research Methodology*. Second Edition. Cape Town: Oxford University Press.
- Wenfei F., Floris G., Shuai M. and Heiko M. (2010). Detecting Inconsistencies in Distributed Data. DOI: 978-1-4244-5446-4/10

- Weng, C. C., Hao, H., Carolina, O. L. U. and Yong, C. (2013). Benefits and Challenges of Electronic Health Record System on Stakeholders: A Qualitative Study of Outpatient Physicians, DOI: 10.1007/s10916-013-9960-5
- Werner R. M., Kolstad J. T., Stuart E. A. and Polsky D. (2011). “The Effect of Pay-for-Performance in Hospitals: Lessons from Quality Improvement” Health Affairs. 2011;30(4):690–8.
- William N. (2005). “Your research project”, 2nd edition. Sage.
- Wilson, J. (2010) “Essentials of Business Research: A Guide to Doing Your Research Project” SAGE Publications
- William B., Christoph T., Grzegorz S., Adrian S. and Biraj P. (2017) - Towards Trust and Governance in Integrated Health and Social Care Platforms DOI: 10.1007/978-3-319-47617-9\_11
- Wincup E. (2017) - Criminological research: Understanding qualitative methods. ISBN: 978-1-4462-0913-4
- Winter R. (2008) Design Science Research in Europe. In: European Journal Of Information Systems 17(5), pp. 470–475
- Wullianallur R. and Viju R. (2014) - Big data analytics in healthcare: promise and potential. DOI: 10.1186/2047-2501-2-3© Raghupathi and Raghupathi; licensee BioMed Central Ltd. 2014
- Xiaoxia Y., Daniel RD., Justin GT., Jeffrey WF. (2015) - Development of a physiologically based pharmacokinetic model for assessment of human exposure to bisphenol A. DOI: <http://dx.doi.org/10.1016/j.taap.2015.10.016>
- Xin L., José-Fernán M. and Gregorio R. (2016) - A New Fuzzy Ontology Development Methodology (FODM) Proposal. DOI: 10.1109/ACCESS.2016.2621756

- Xu Z. S. (2007). Models for multiple attribute decision-making with intuitionistic fuzzy information. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 15, 285–297. DOI: <https://doi.org/10.1142/S0218488507004686>
- Yi P., Yong Z., Yu T. and Shiming L. (2011). – An incident information management framework based on data integration, data mining, and multi-criteria decision making, Volume 51, Issue 2, May 2011, Pages 316–327, DOI:10.1016/j.dss.2010.11.025
- Yuchu Q., Wenlong L., Qunfen Q., Xiaojun L., Yanru Z., Paul J. S. and Xiangqian J. (2016). Status, Comparison, and Issues of Computer-Aided Design Model Data Exchange Methods Based on Standardized Neutral Files and Web Ontology Language File, Paper No: JCISE-16-1047; DOI: 10.1115/1.4034325
- Yuguang W. and Yanan L. (2016) - Ruifa Hu An intuitionistic fuzzy multi-attribute decision making model for the acceptance of genetically modified foods based on IFHA operator. DOI:10.1109/CCDC.2016.7531880
- Yusuf, M. K. and Azlan, A. 2012. *Comparative Study of Techniques in Reducing Inconsistent Data*, International Journal of database Theory and Application. Vol.5, No. 1.
- Zadeh L. A. (1965). "Fuzzy sets," *Inf. Control*, vol. 8, no. 3, pp. 338-353
- Zhang F., Ma Z., Yan L., Cheng J. (2013). Construction of fuzzy OWL ontologies from fuzzy EER models: a semantics-preserving approach *Fuzzy Sets Syst.*, 229 (2013), pp. 1-32
- Zbigniew S. (2004). *An Introduction to Rough Set Theory and Its Applications*. ICENCO'2004, December 27-30, 2004, Cairo, Egypt
- Zhang R., Pakhomov S., McInnes BT. and Melton GB. (2011). Evaluating Measures of Redundancy in Clinical Texts. *Proc AMIA: 2011*;1612–1620

- Zhao, H. and Ram, S. (2008). Entity matching across heterogeneous data sources: An approach based on constrained cascade generalization. *Data & Knowledge Engineering*.
- Zhonggui, M., Chengyao, W. and Zongjie, W. 2011. Research on three-layered metadata model for oil-gas data integration. *The processing of IEEE CCIS2011*.
- Zhongqi Z., Aming Z., and Gang X. (2012). – Improved Protein Hydrogen/Deuterium Exchange Mass Spectrometry Platform with Fully Automated Data Processing, *Anal. Chem.*, 2012, 84 (11), pp 4942–4949, DOI: 10.1021/ac300535r
- Zhongqin B., Feifei X., Jingsheng L. and Teng J. (2016) - Attribute reduction in decision-theoretic rough set model based on minimum decision cost. DOI: 10.1002/cpe.3830

## **APPENDIX A: Fuzzy Hypertension Diagnosis using MATLAB**

Appendix A contains a copy of Fuzzy Hypertension Diagnosis using MATHLAB

The figure A1 describes Fuzzy model for Hypertension Diagnosis, as follows:

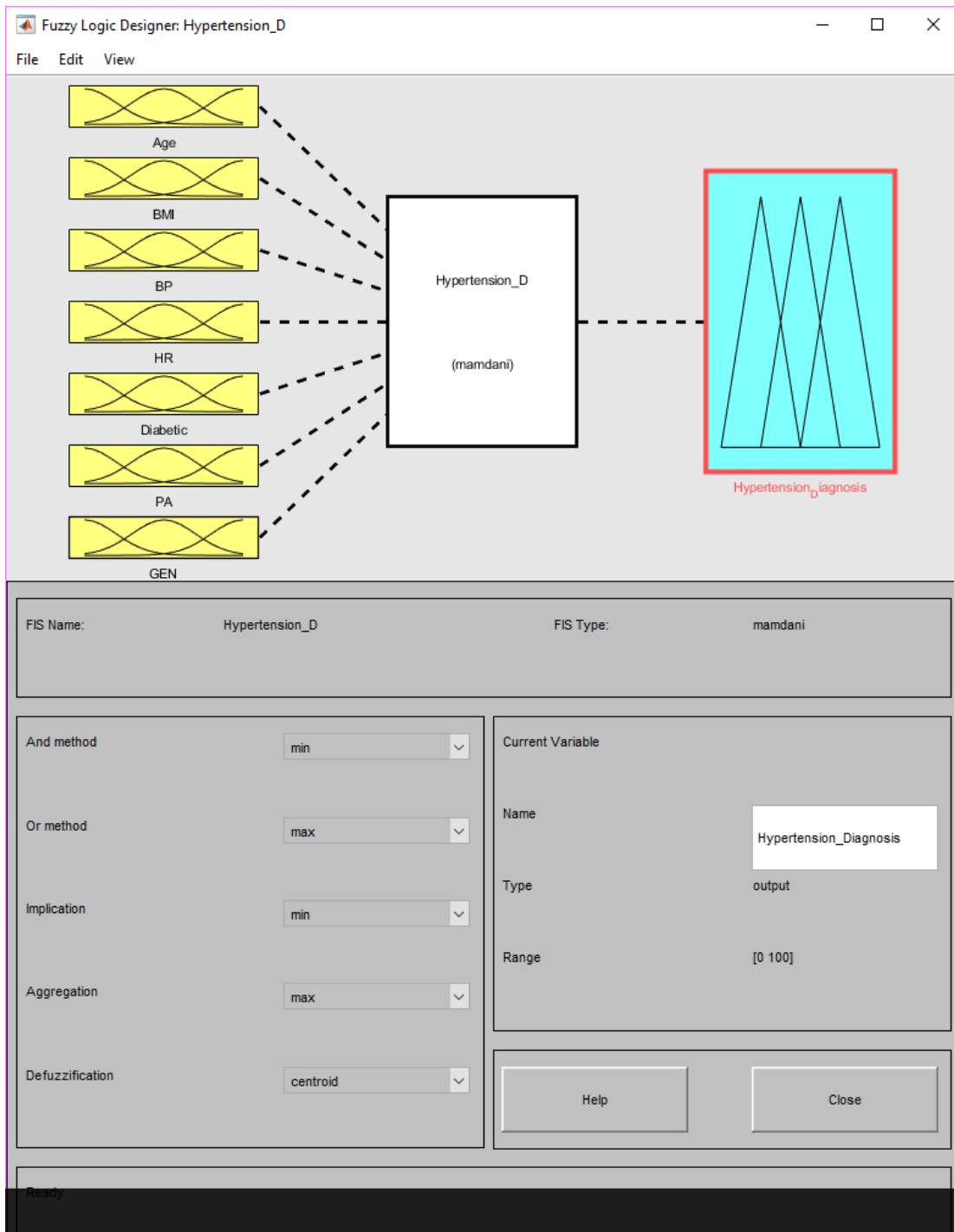


Figure A1: Fuzzy model for Hypertension Diagnosis.



The figure A2 describes Fuzzy Age input details for Hypertension Diagnosis, as follows:

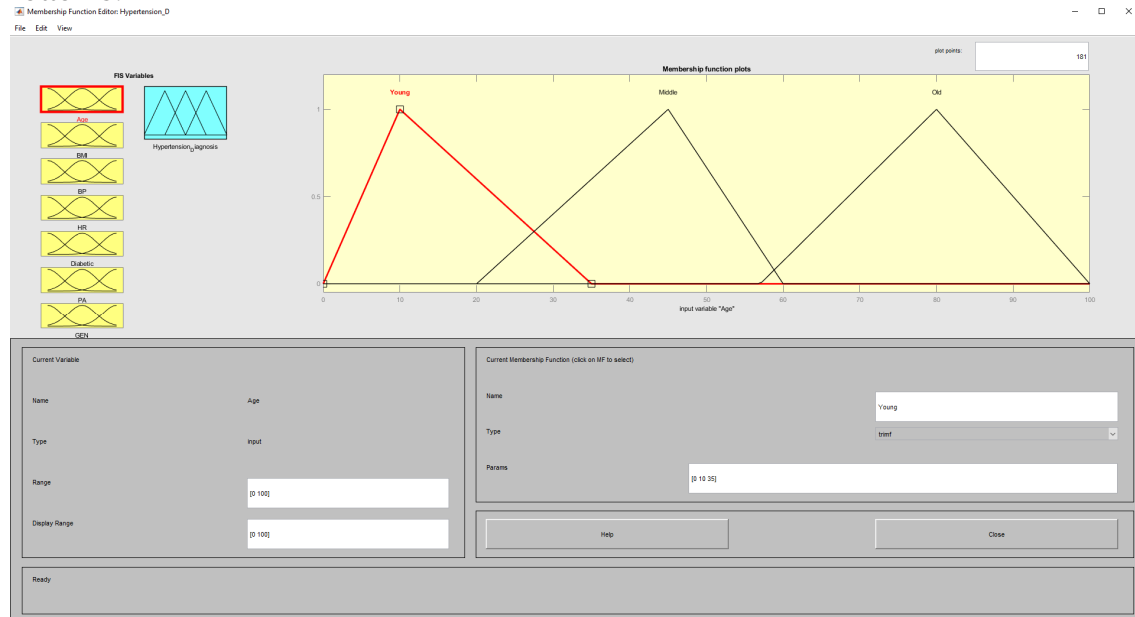


Figure A2: Fuzzy Age input details for Hypertension Diagnosis.

The figure A3 describes Fuzzy BMI input details for Hypertension Diagnosis, as follows:

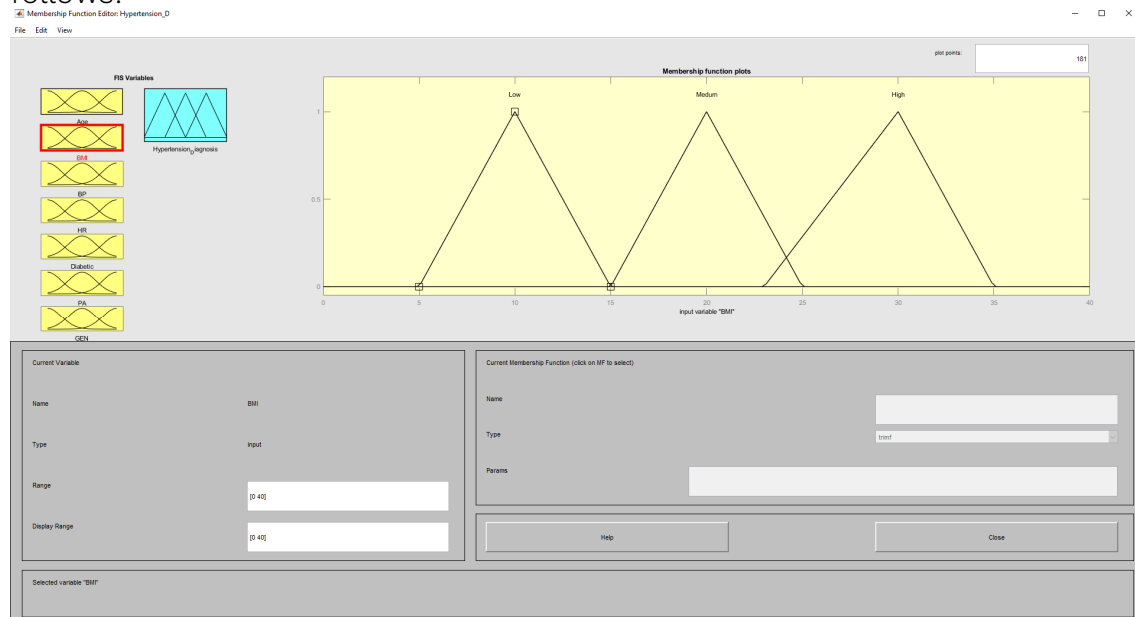


Figure A3: Fuzzy BMI input details for Hypertension Diagnosis.

The figure A4 describes Fuzzy BP input details for Hypertension Diagnosis, as follows:

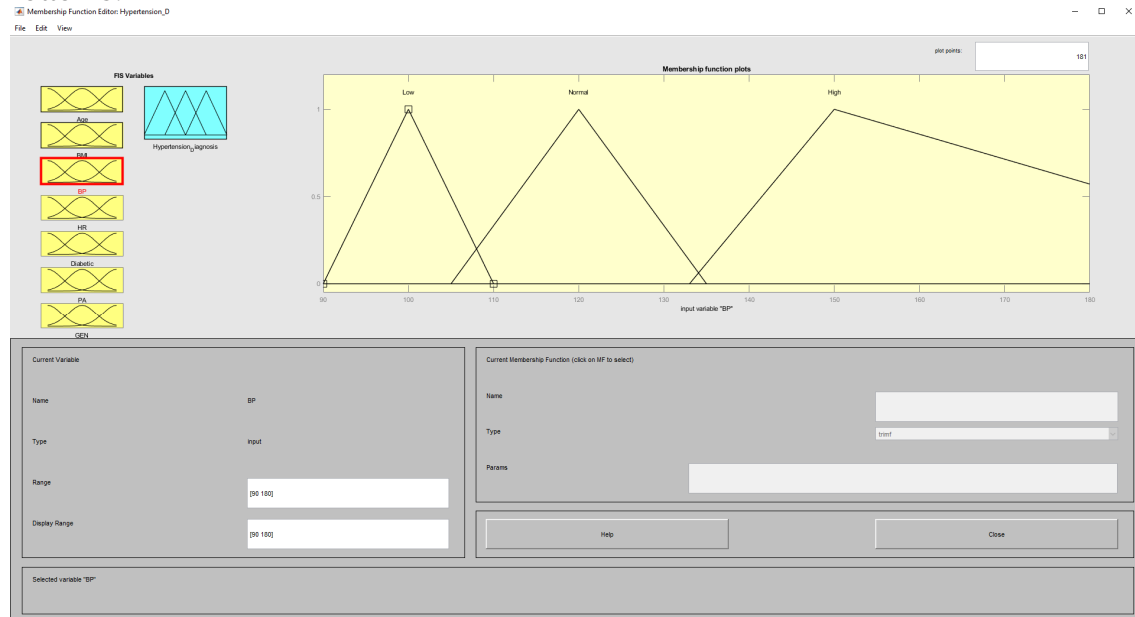


Figure A4: Fuzzy BP input details for Hypertension Diagnosis.

The figure A5 describes Fuzzy HR input details for Hypertension Diagnosis, as follows:

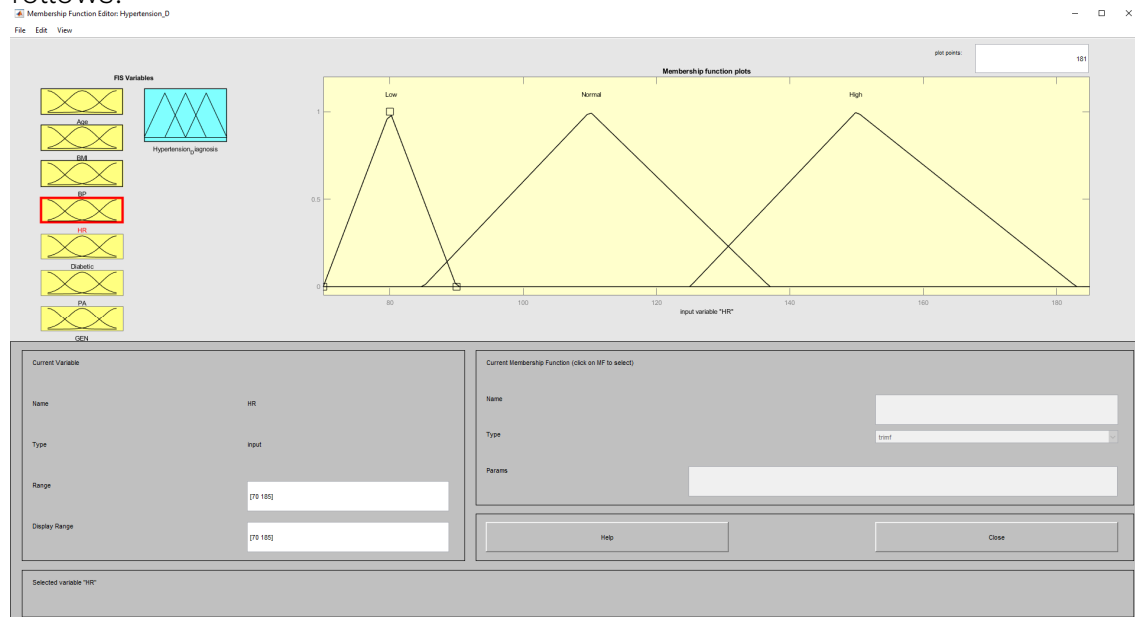


Figure A5: Fuzzy HR input details for Hypertension Diagnosis.

The figure A6 describes Fuzzy Diabetic input details for Hypertension Diagnosis, as follows:

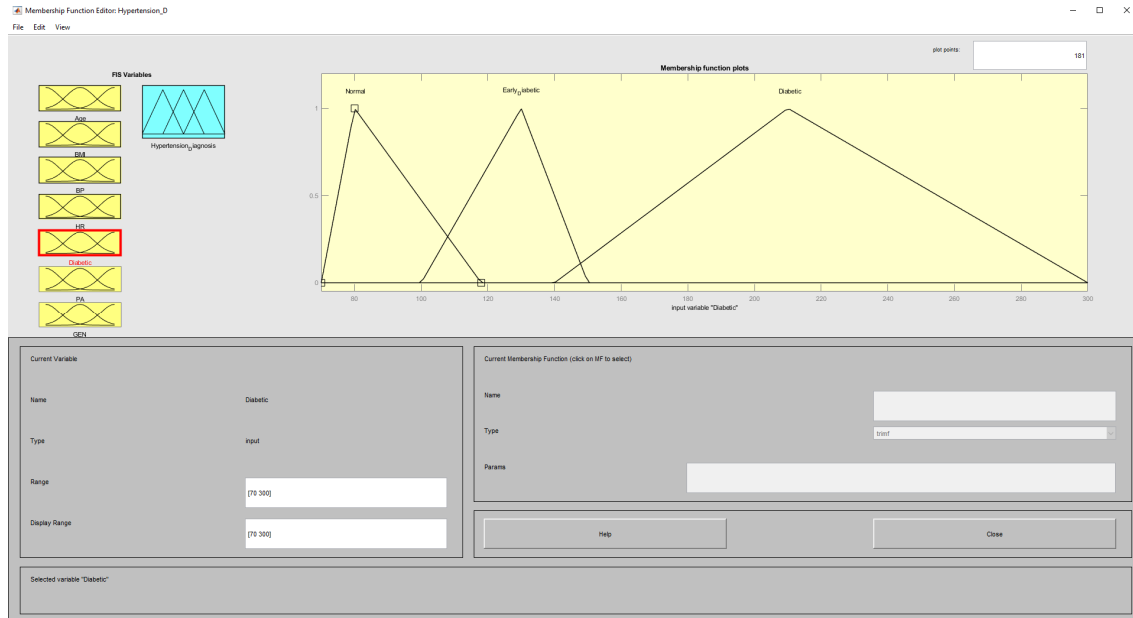


Figure A6: Fuzzy Diabetic input details for Hypertension Diagnosis.

The figure A7 describes Fuzzy PA input details for Hypertension Diagnosis, as follows:

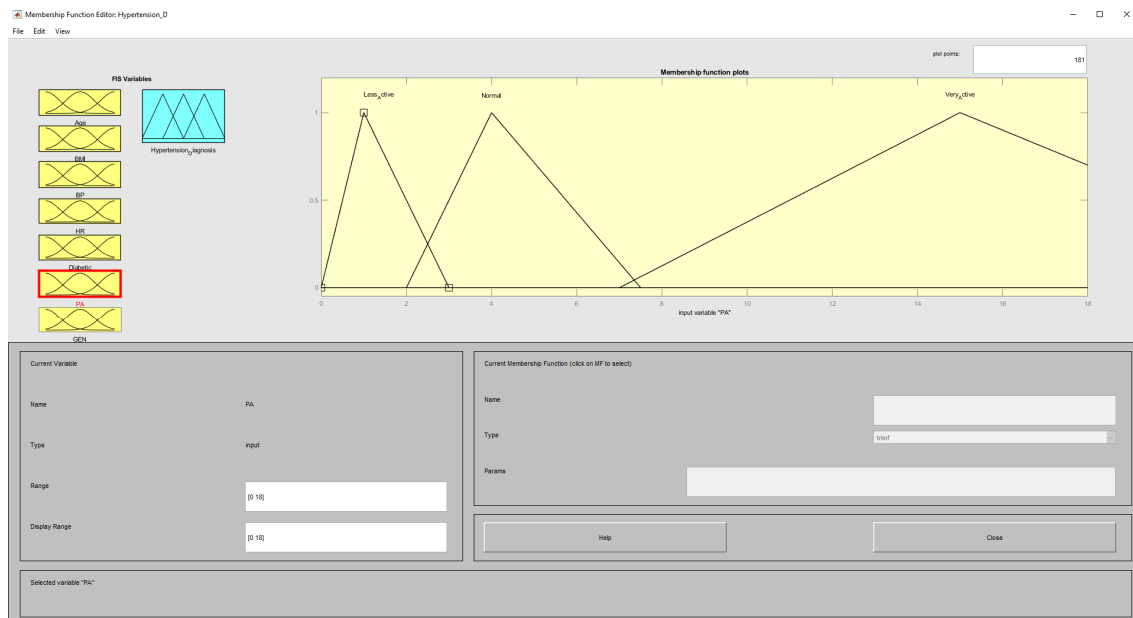


Figure A7: Fuzzy PA input details for Hypertension Diagnosis.

The figure A8 describes Fuzzy GEN input details for Hypertension Diagnosis, as follows:

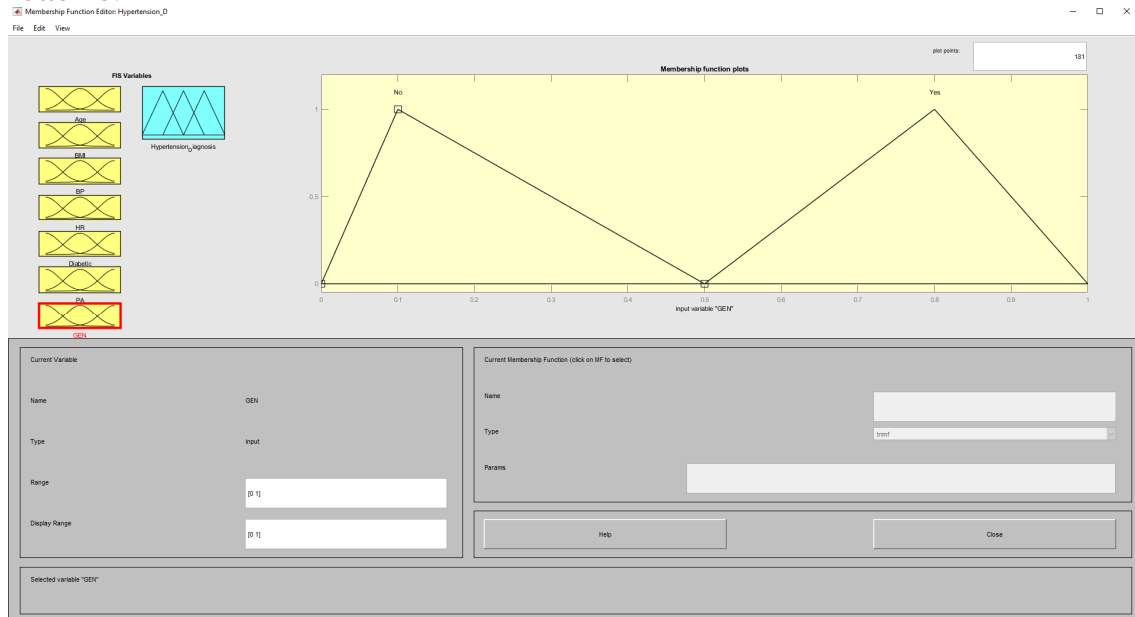


Figure A8: Fuzzy GEN input details for Hypertension Diagnosis.

The figure A9 describes Fuzzy result input details for Hypertension Diagnosis, as follows:

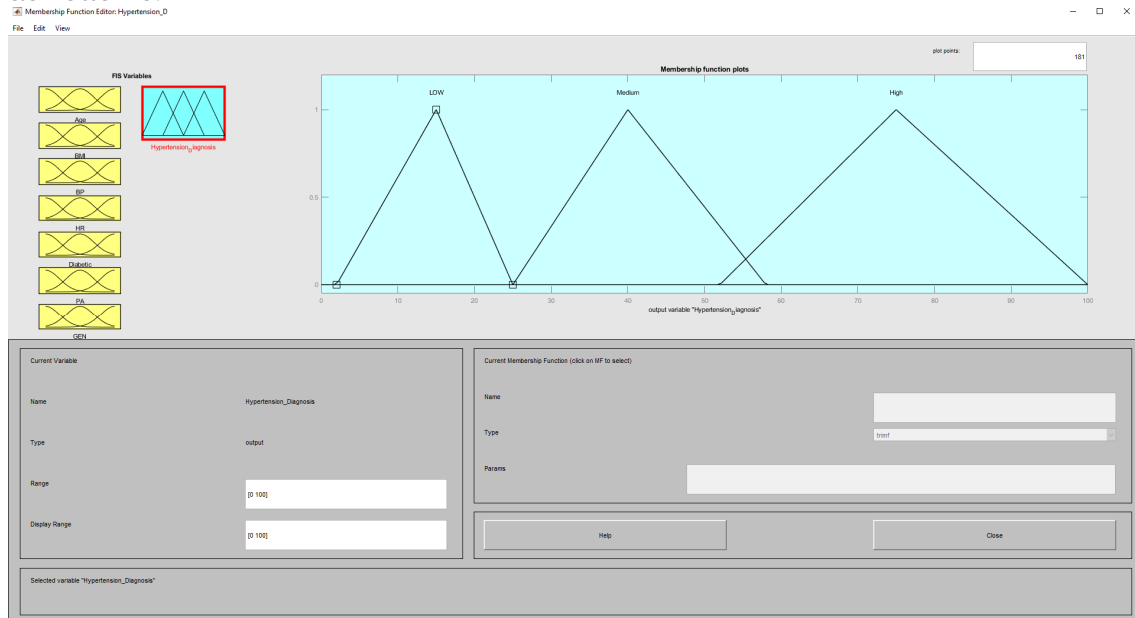


Figure A9: Fuzzy result input details for Hypertension Diagnosis.

The figure A10 describes Fuzzy rules details for Hypertension Diagnosis, as follows:

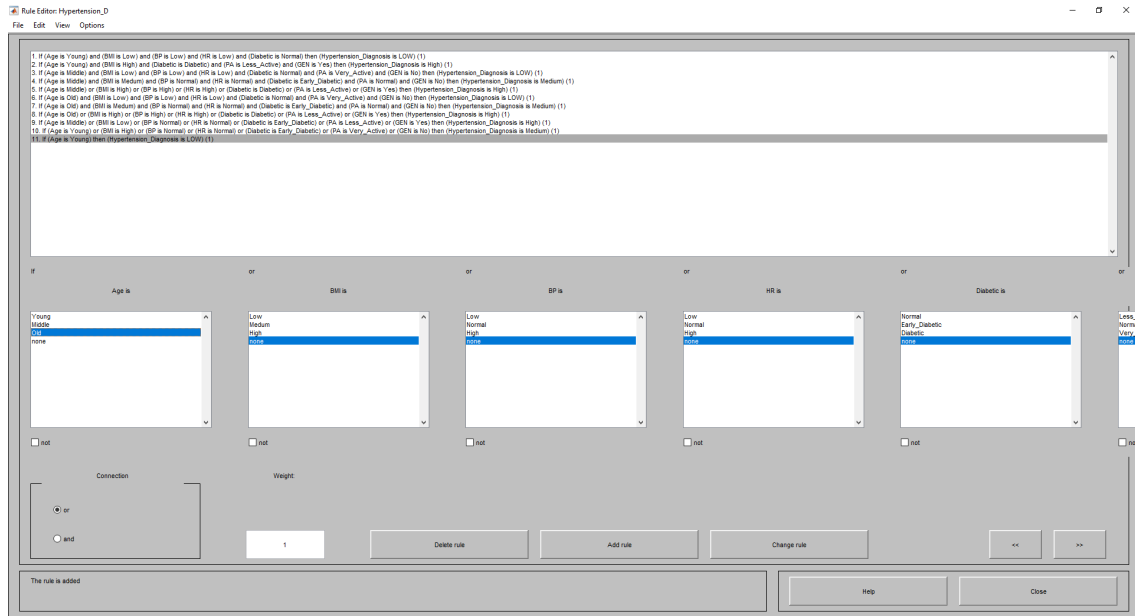


Figure A10: Fuzzy rules details for Hypertension Diagnosis.

The figure A11 describes Fuzzy result different angle view\_1 details for Hypertension Diagnosis, as follows:

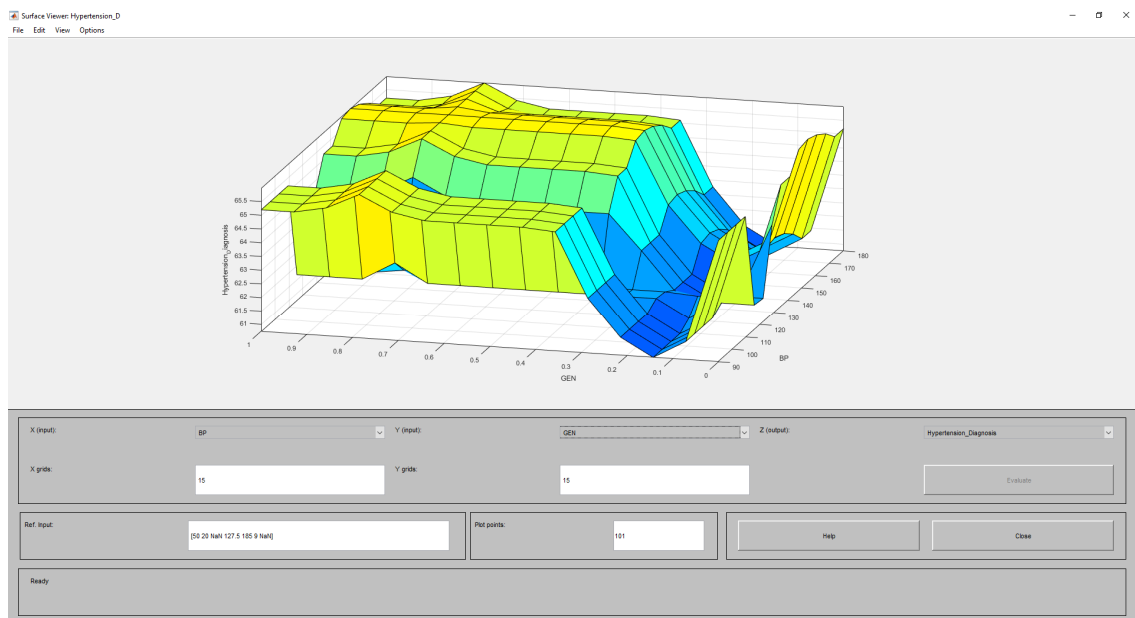


Figure A11: Fuzzy result different angle view\_1 details for Hypertension Diagnosis.

The figure A12 describes Fuzzy result different angle view\_2 details for Hypertension Diagnosis, as follows:

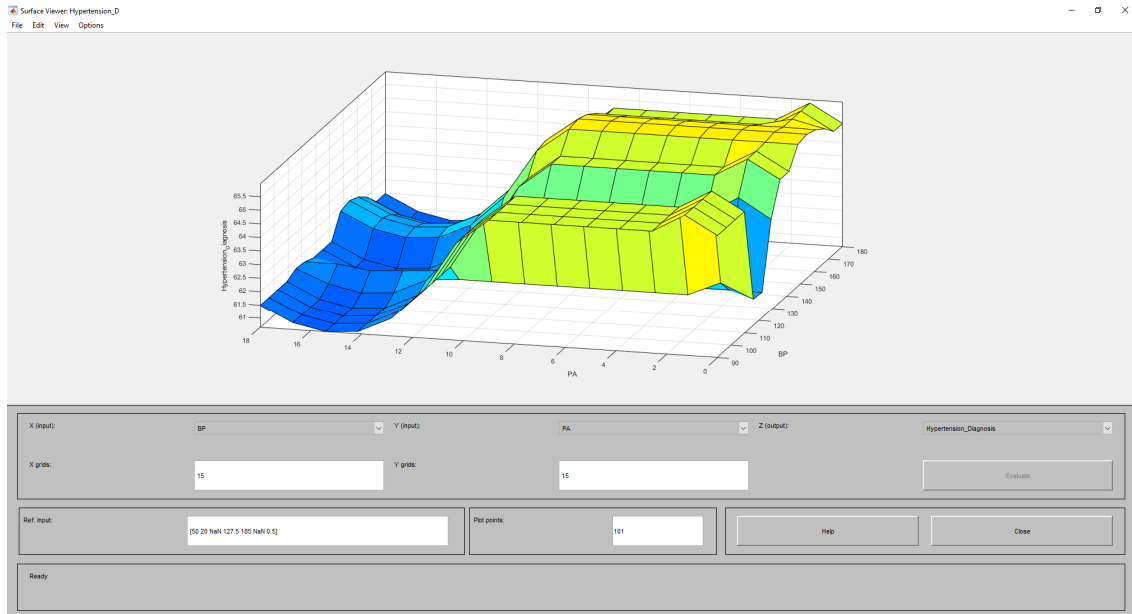


Figure A12: Fuzzy result different angle view\_2 details for Hypertension Diagnosis.

The figure A13 describes Fuzzy result different angle view\_3 details for Hypertension Diagnosis, as follows:

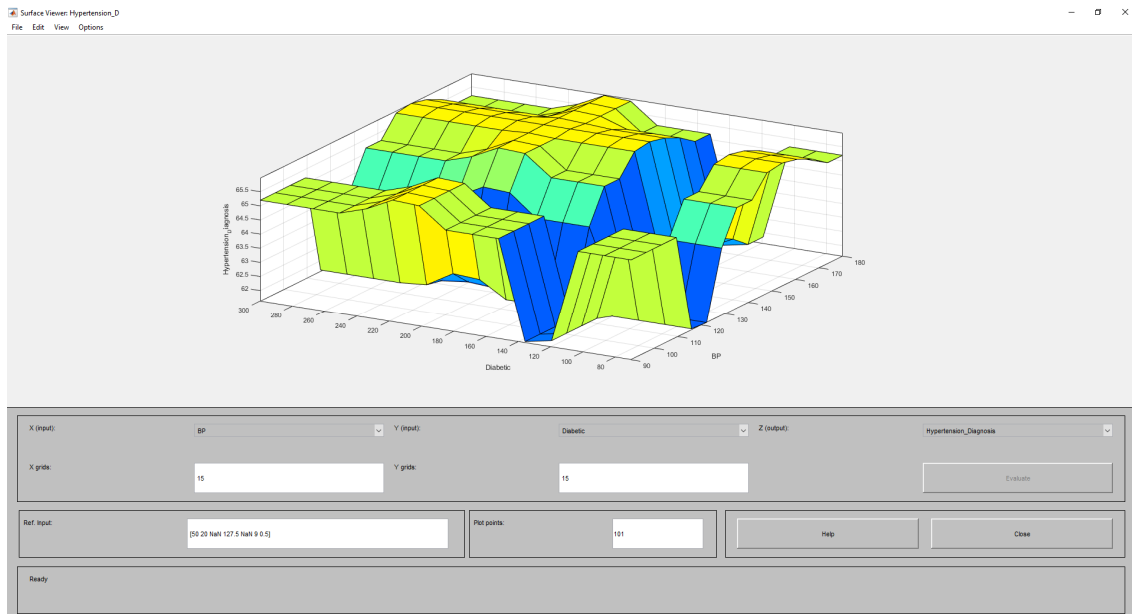


Figure A13: Fuzzy result different angle view\_3 details for Hypertension Diagnosis.

The figure A14 describes Fuzzy result different angle view\_4 details for Hypertension Diagnosis, as follows:

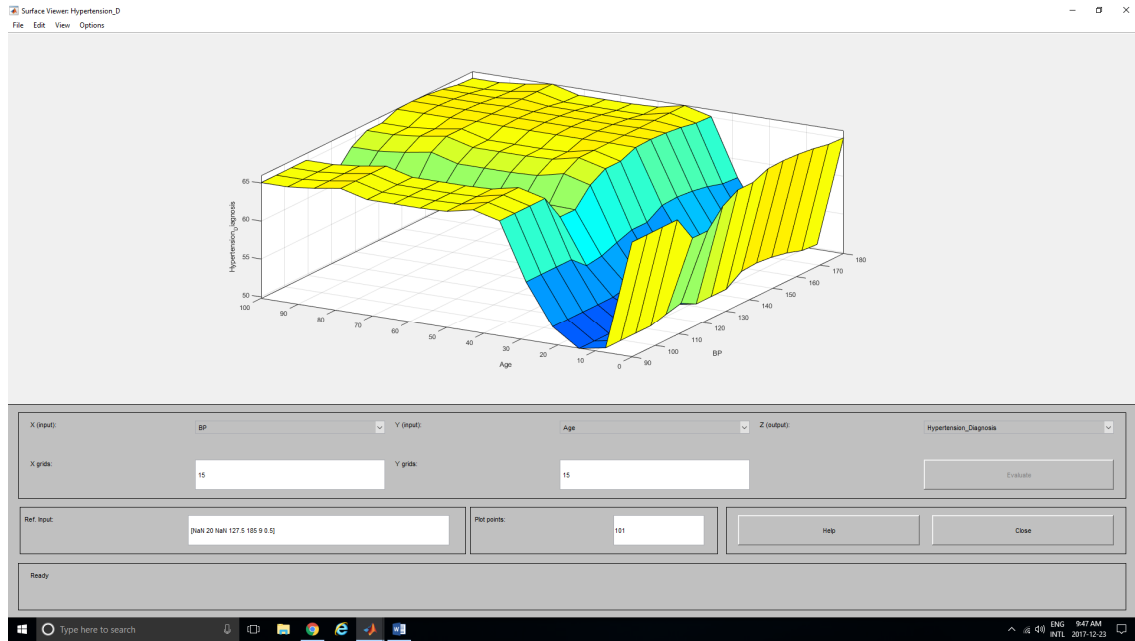


Figure A14: Fuzzy result different angle view\_4 details for Hypertension Diagnosis.

The figure A15 describes Fuzzy result different angle view\_5 details for Hypertension Diagnosis, as follows:

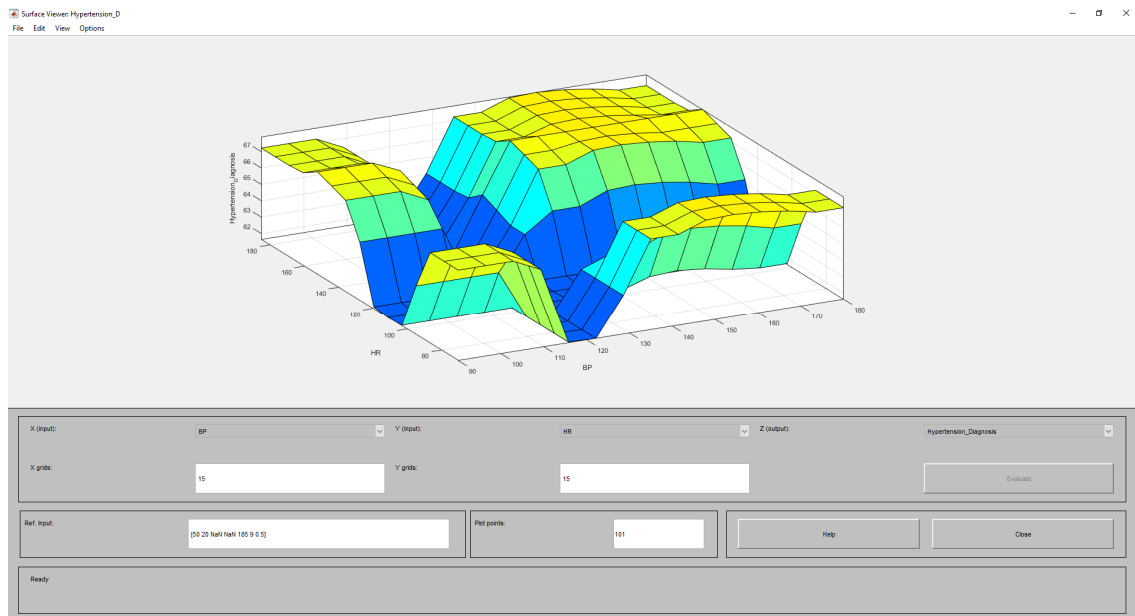


Figure A15: Fuzzy result different angle view\_5 details for Hypertension Diagnosis.

The figure A16 describes Fuzzy result different angle view\_6 details for Hypertension Diagnosis, as follows:

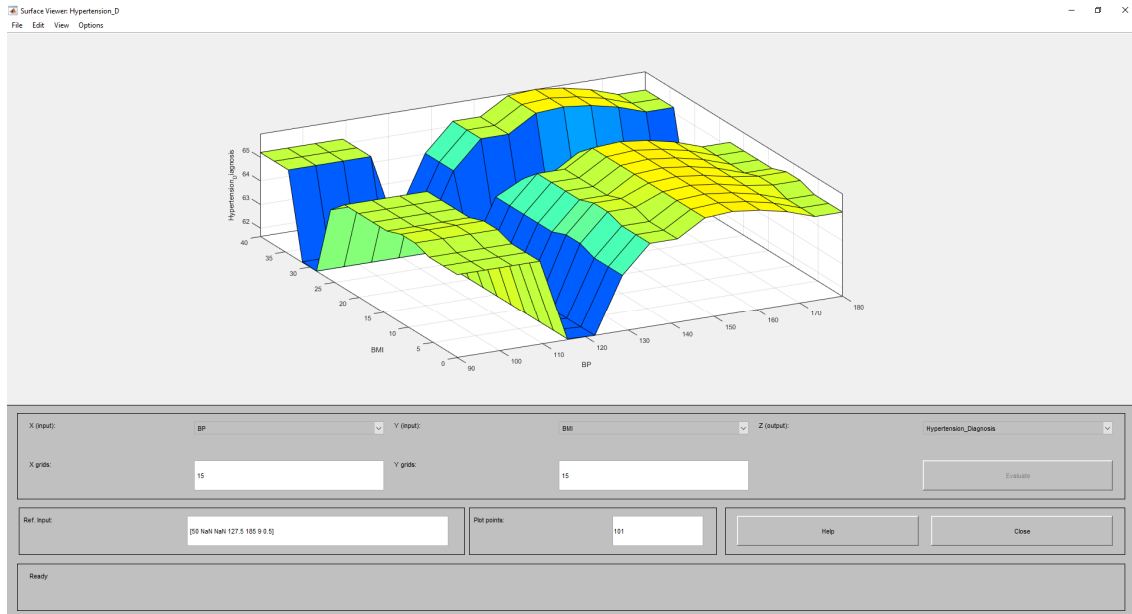


Figure A16: Fuzzy result different angle view\_6 details for Hypertension Diagnosis.

The figure A17 describes Fuzzy result different angle view\_7 details for Hypertension Diagnosis, as follows:

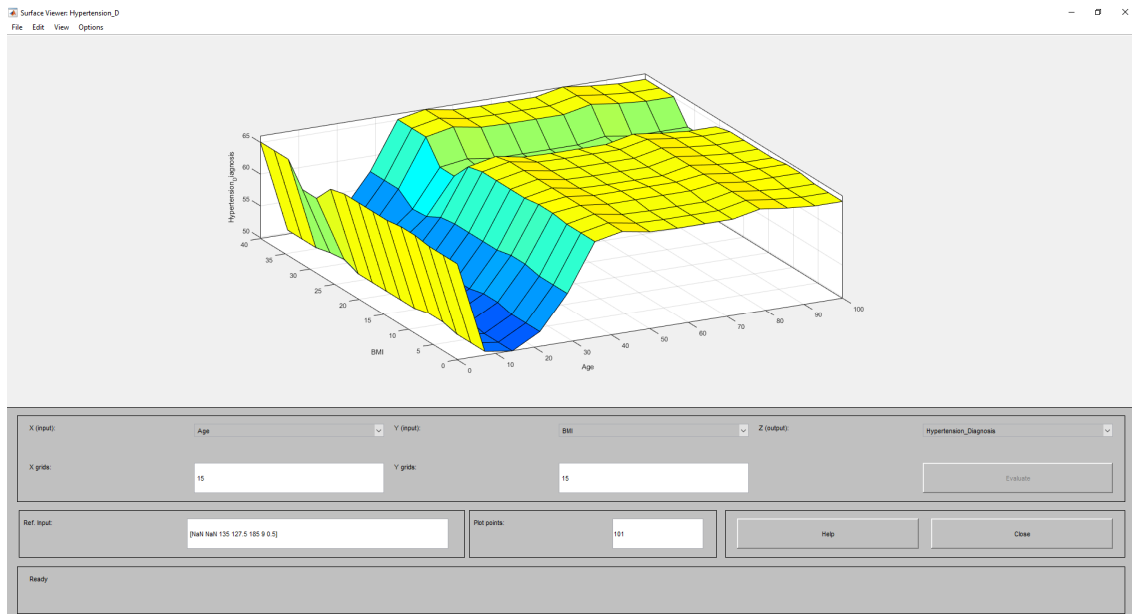


Figure A17: Fuzzy result different angle view\_7 details for Hypertension Diagnosis.



## **APPENDIX B: Mathematical Simulation for Hypertension Diagnosis Based on Markov Chain Probability Model**

---

Appendix B contains a copy of Mathematical Simulation for Hypertension Diagnosis Based on Markov Chain Probability Model

The figure B1 describes the Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.35” and “BMI to HR = 0.30”, as follows:

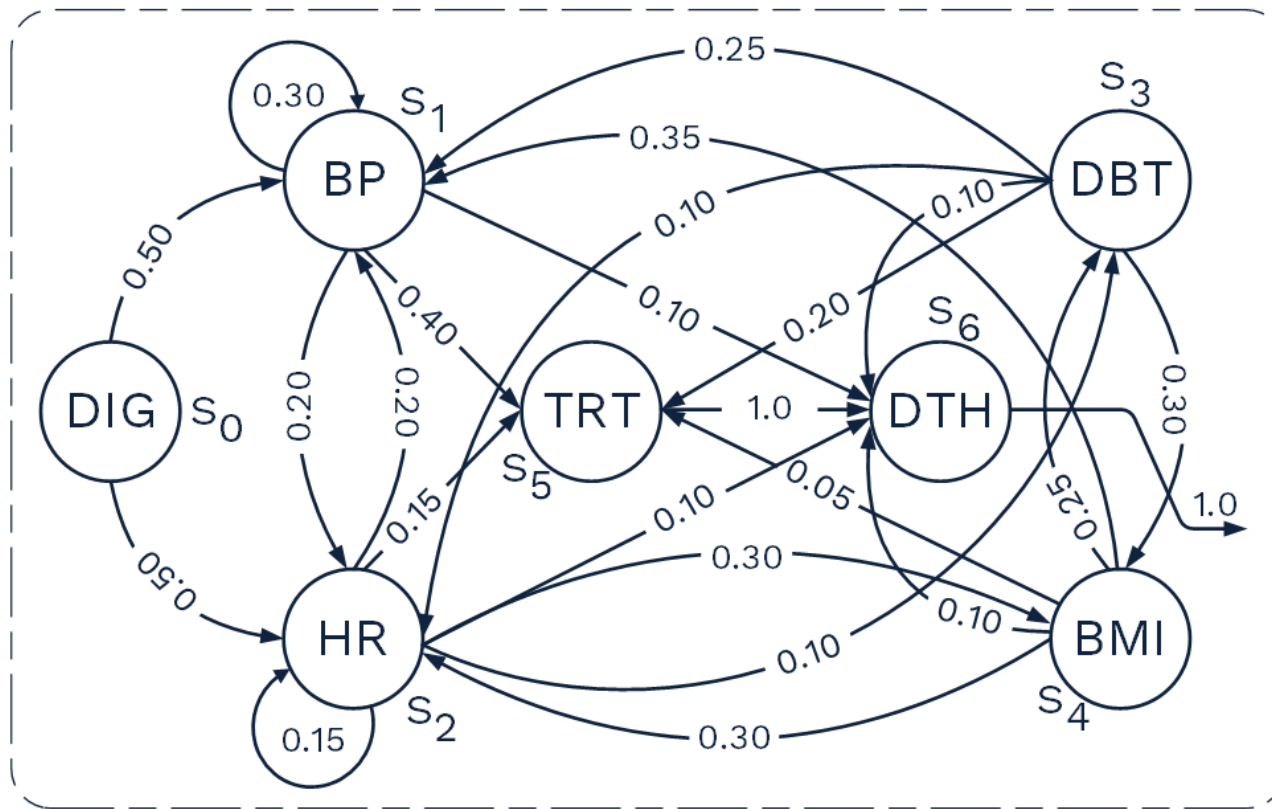


Figure B1: the Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.35” and “BMI to HR = 0.30”.

The figure B2 describes the graphical representation of Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.35” and “BMI to HR = 0.30”, as follows:

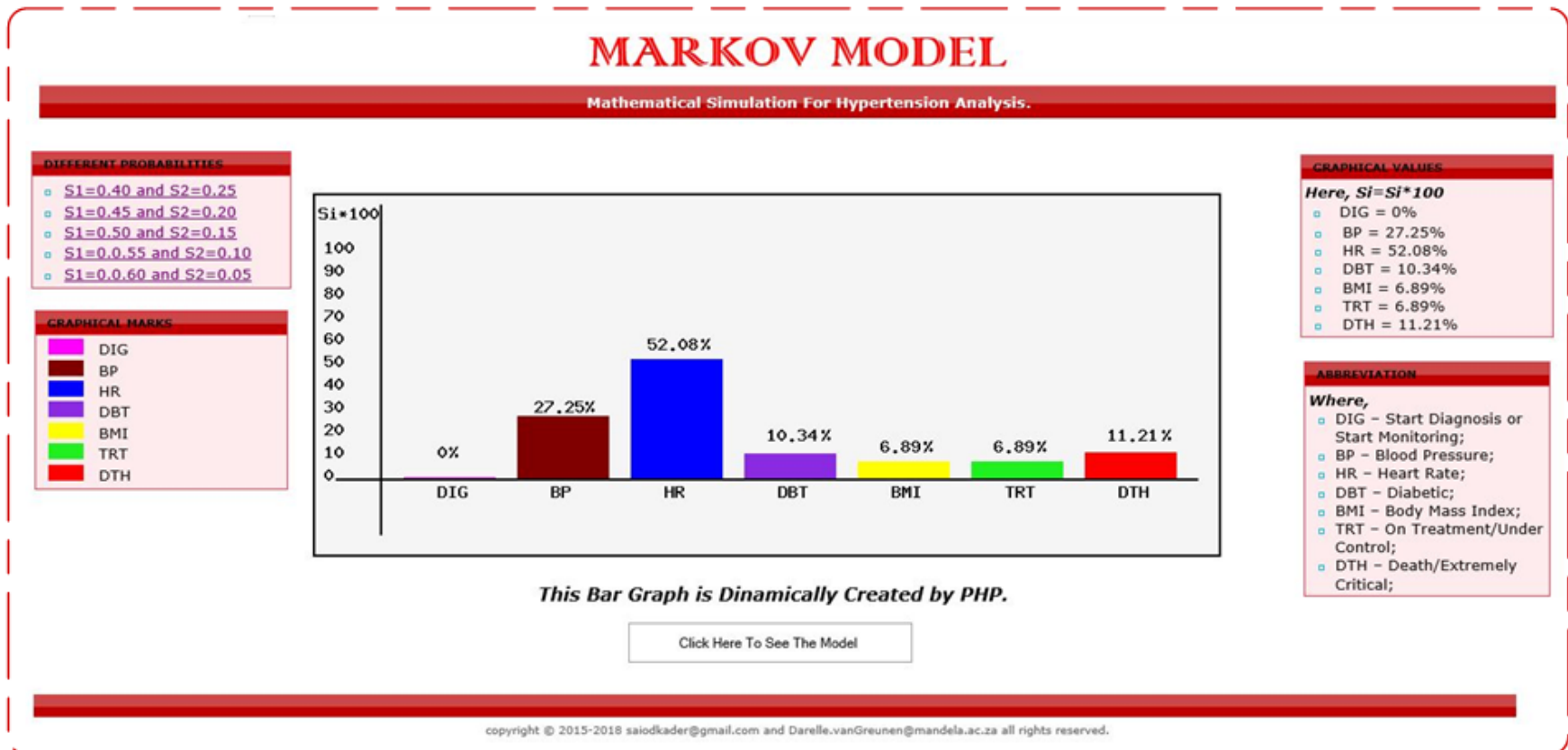


Figure B2: the graphical representation of Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.35” and “BMI to HR = 0.30”.

The figure B3 describes the Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.40” and “BMI to HR = 0.25”, as follows:

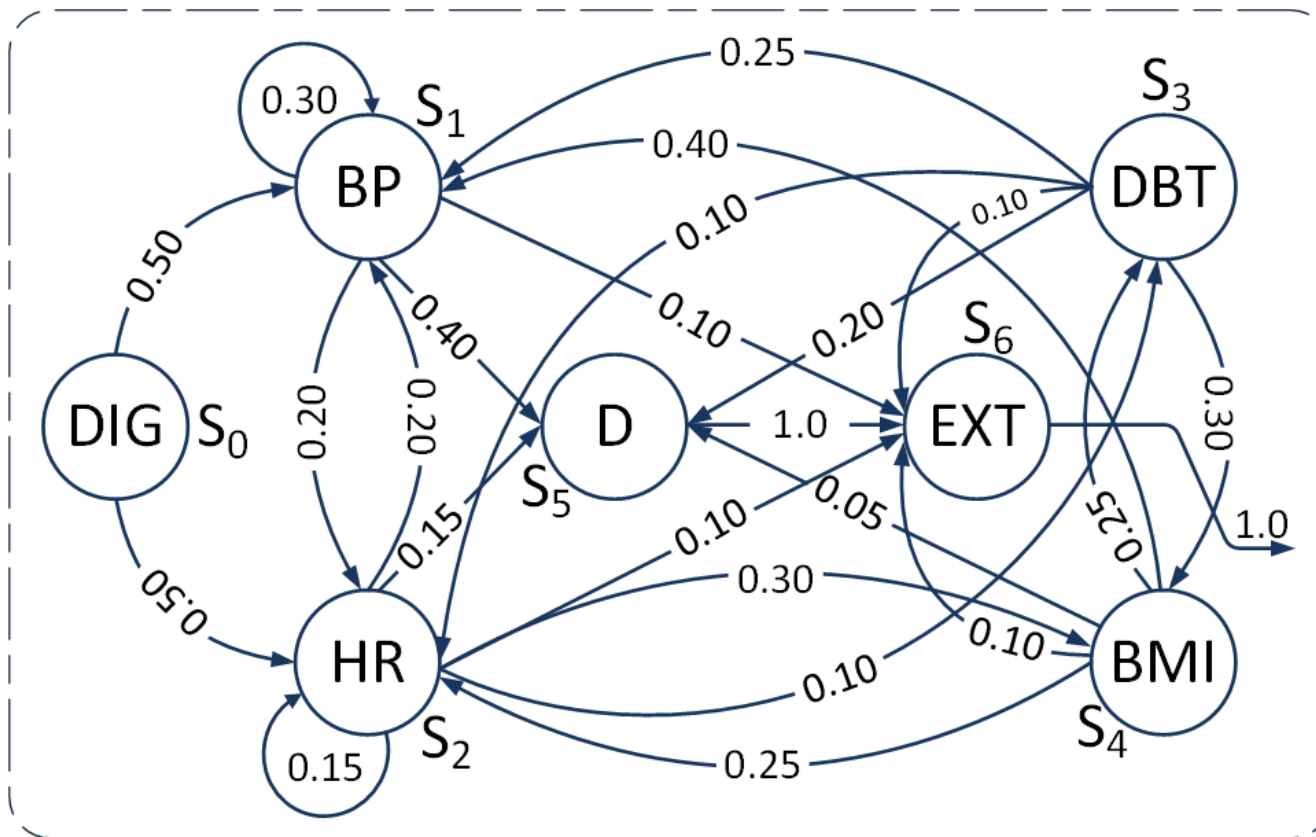


Figure B3: the Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.40” and “BMI to HR = 0.25”.

The figure B4 describes the graphical representation of Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.40” and “BMI to HR = 0.25”, as follows:

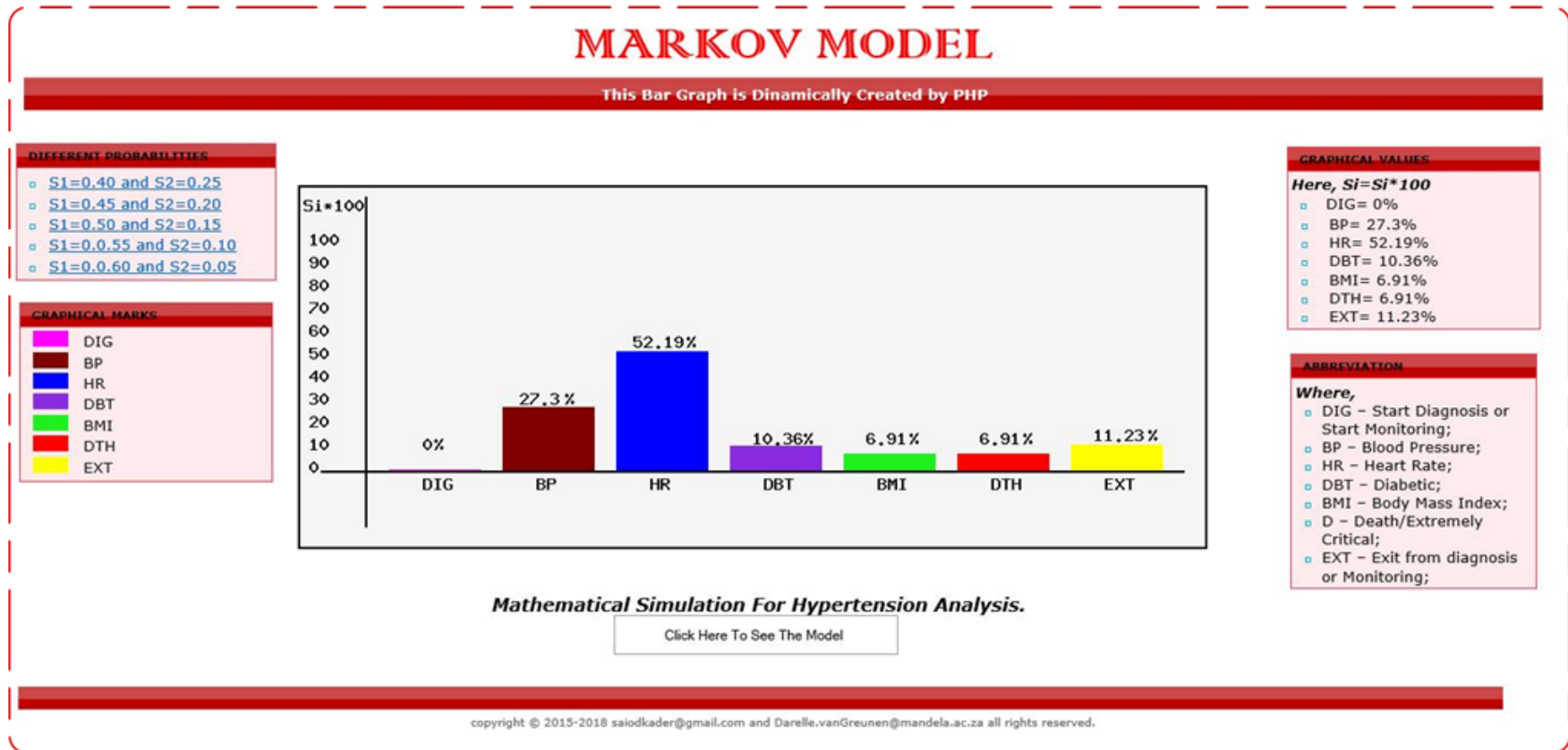


Figure B4: the graphical representation of Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.40” and “BMI to HR = 0.25”.

The figure B5 describes the Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.45” and “BMI to HR = 0.20”, as follows:

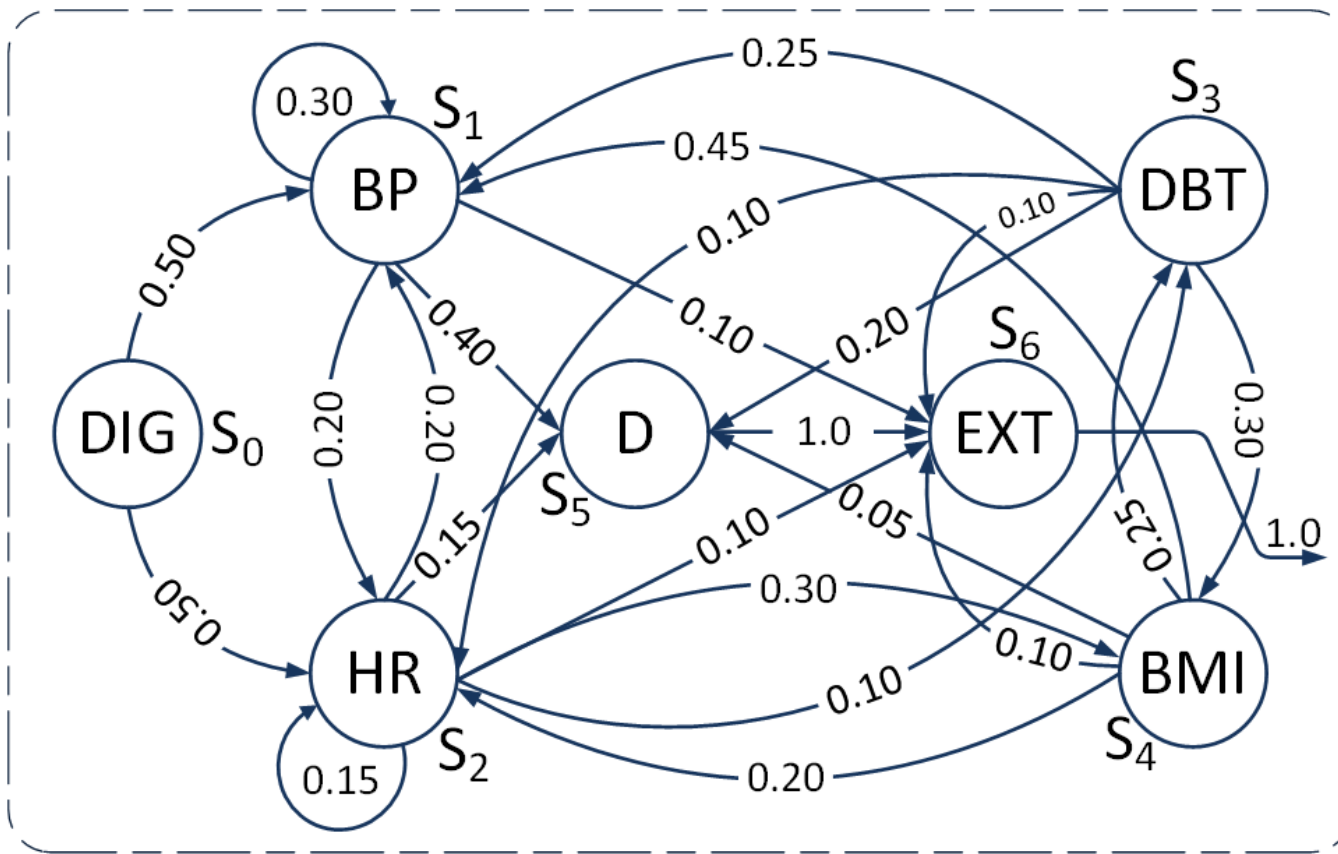


Figure B5: the Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.45” and “BMI to HR = 0.20”.

The figure B6 describes the graphical representation of Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.45” and “BMI to HR = 0.20”, as follows:

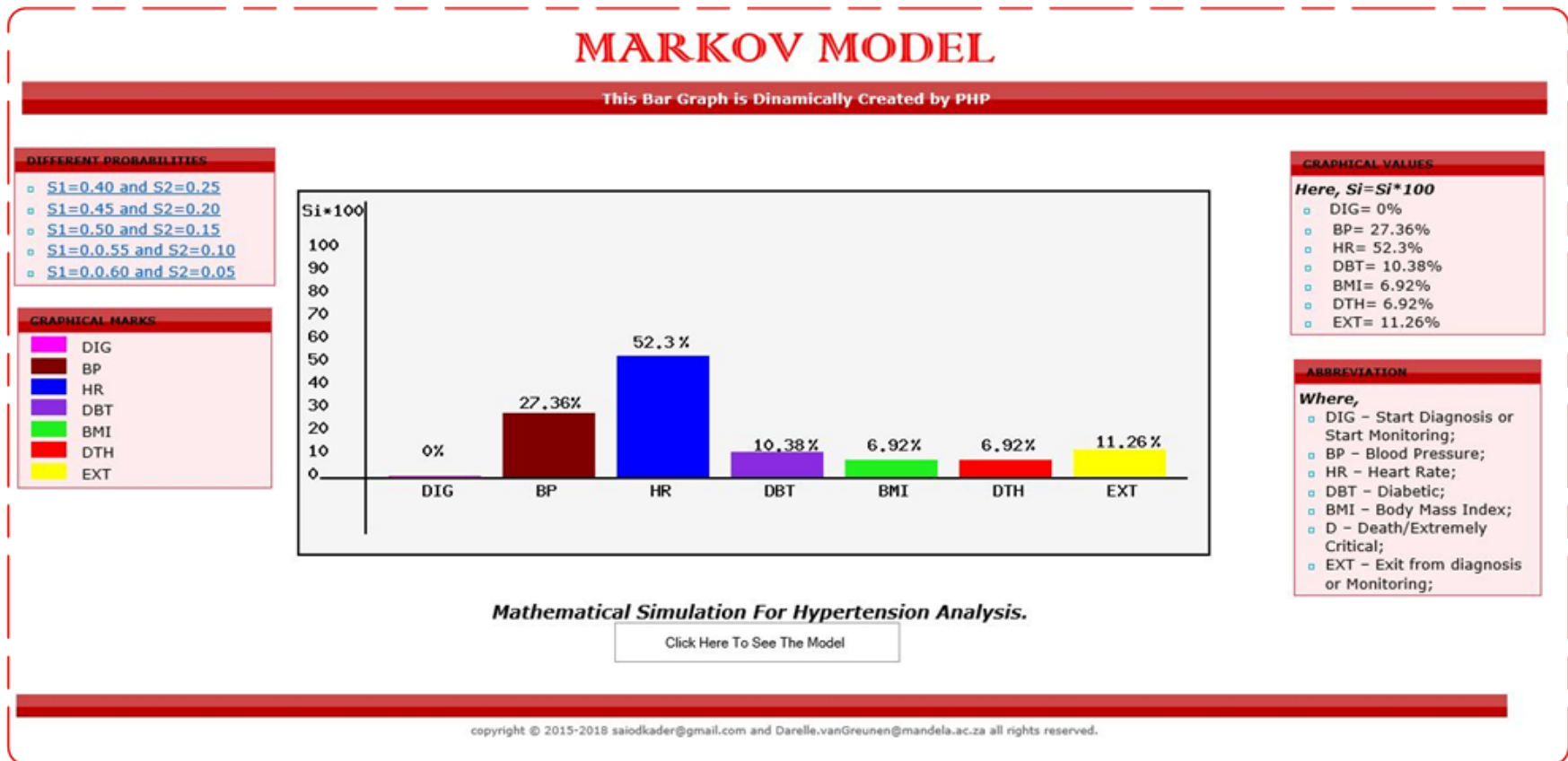


Figure B6: the graphical representation of Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.45” and “BMI to HR = 0.20”.

The figure B7 describes the Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.50” and “BMI to HR = 0.15”, as follows:

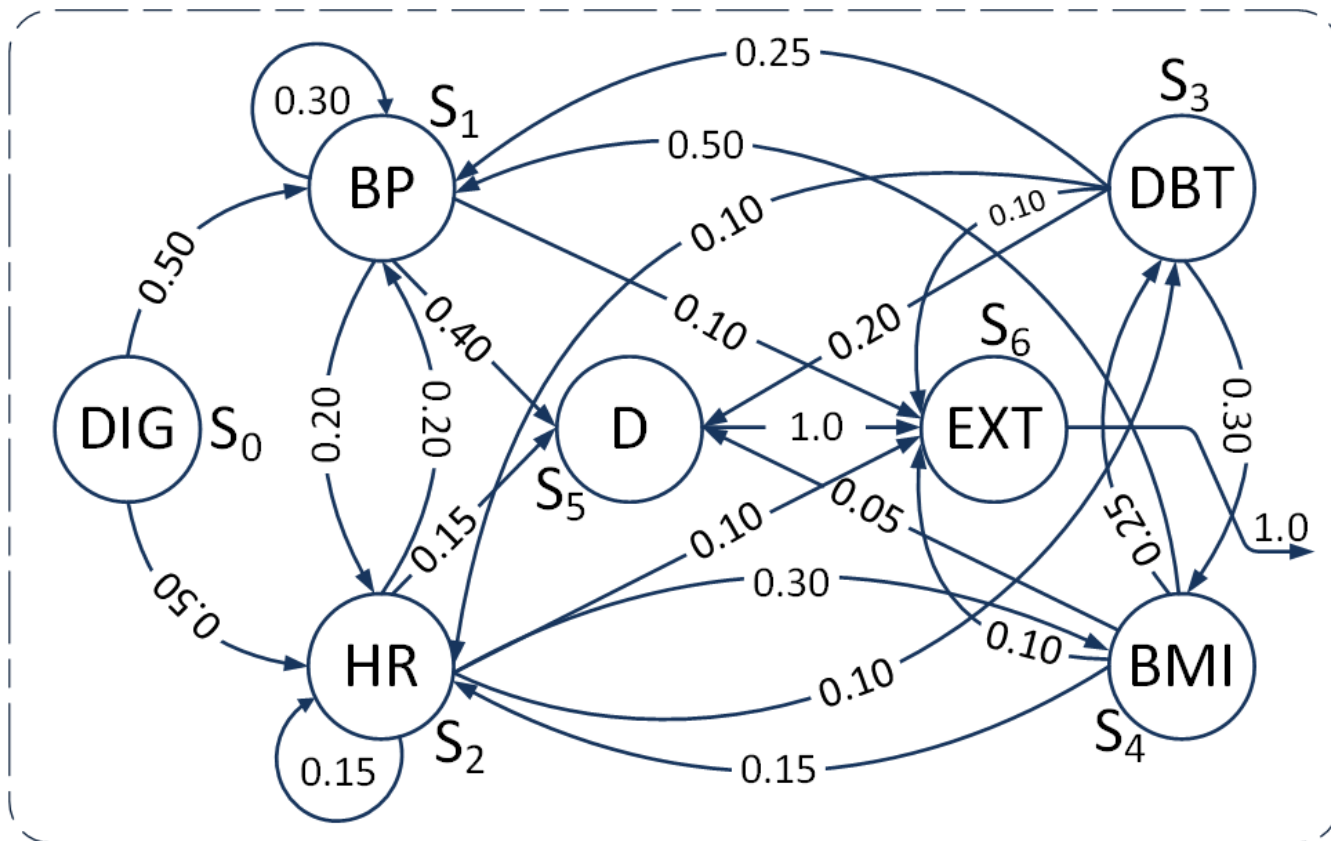


Figure B7: the Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.50” and “BMI to HR = 0.15”.



The figure B8 describes the graphical representation of Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.50” and “BMI to HR = 0.15”, as follows:

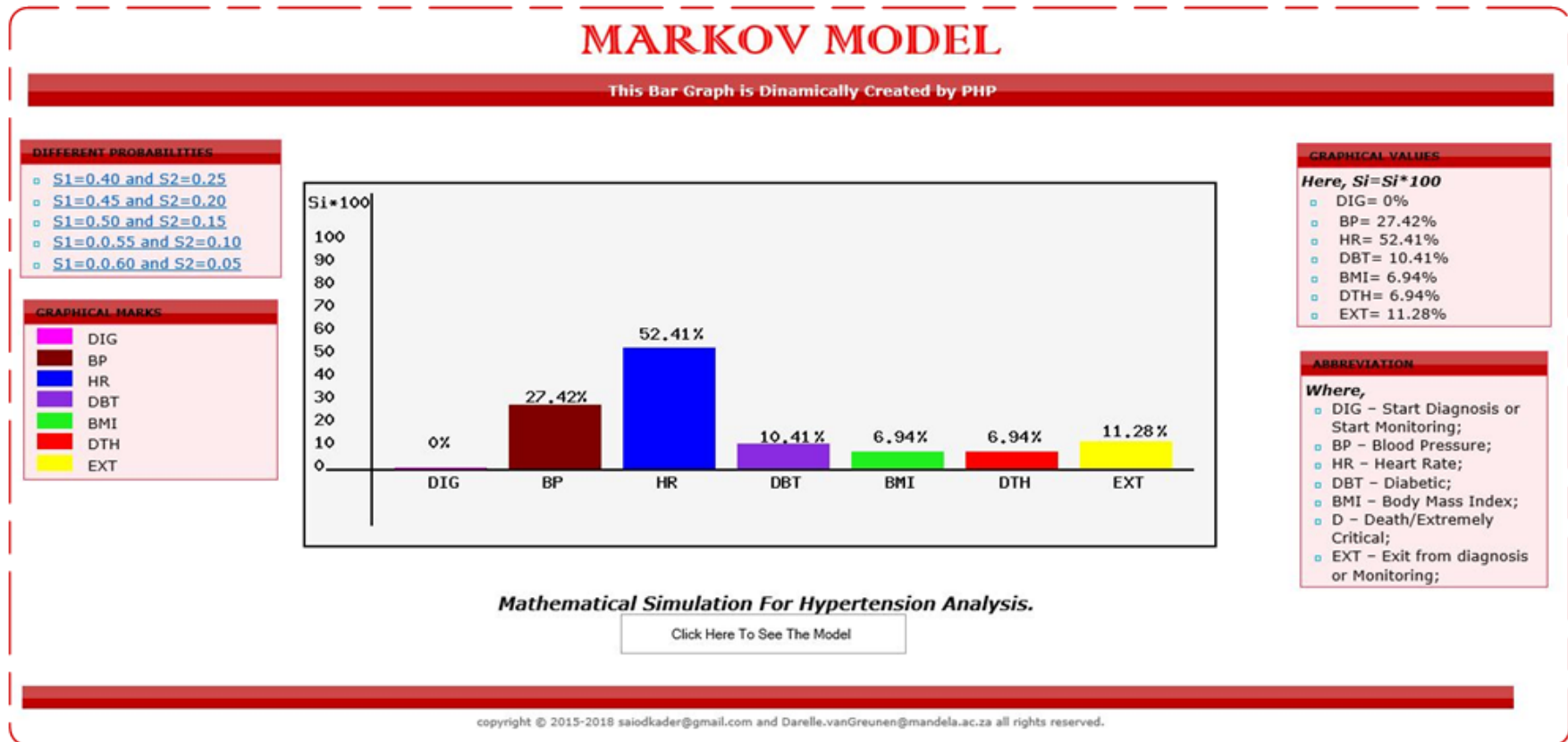


Figure B8: the graphical representation of Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.50” and “BMI to HR = 0.15”.



The figure B10 describes the graphical representation of Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.55” and “BMI to HR = 0.10”, as follows:

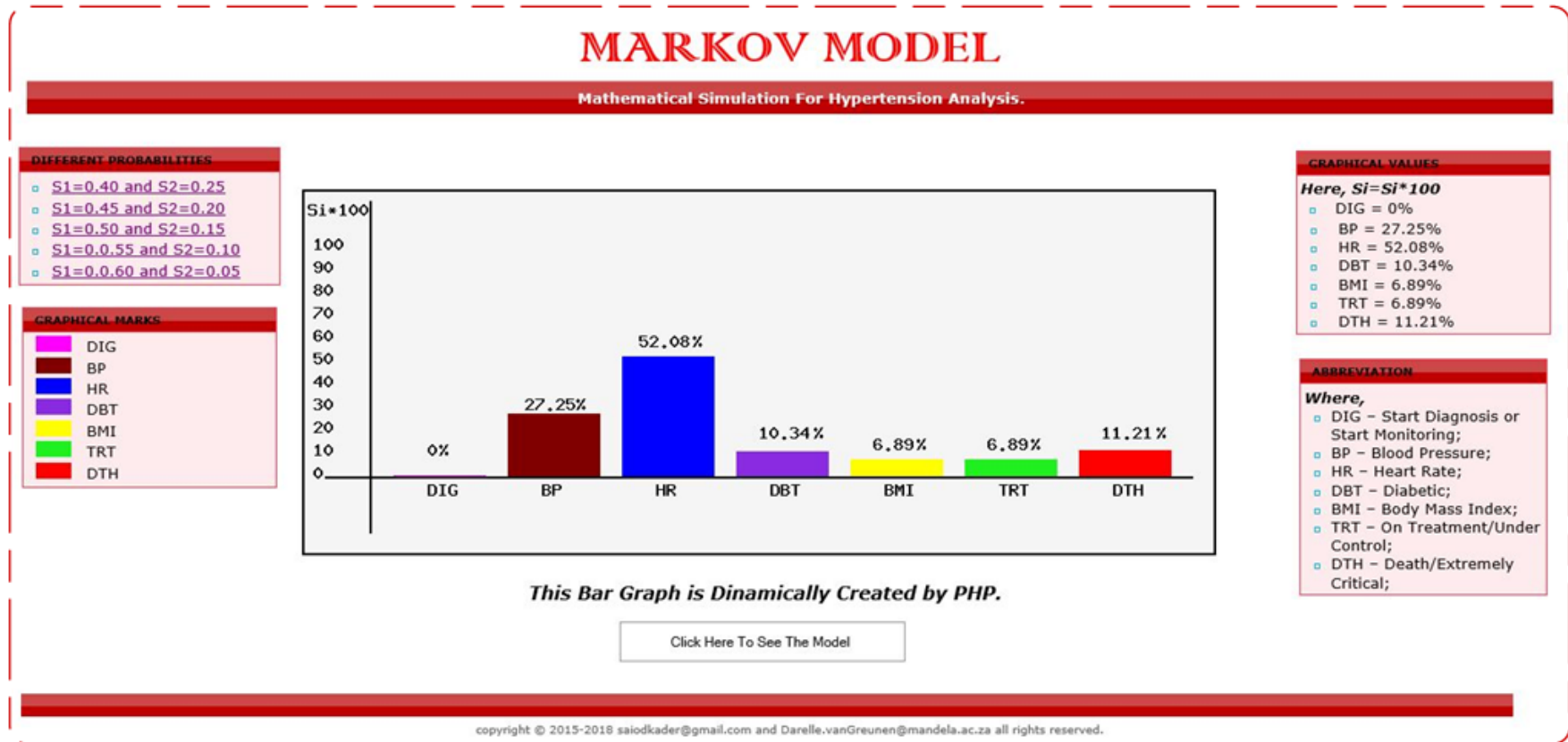


Figure B10: the graphical representation of Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.55” and “BMI to HR = 0.10”.

The figure B11 describes the Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.60” and “BMI to HR = 0.05”, as follows:

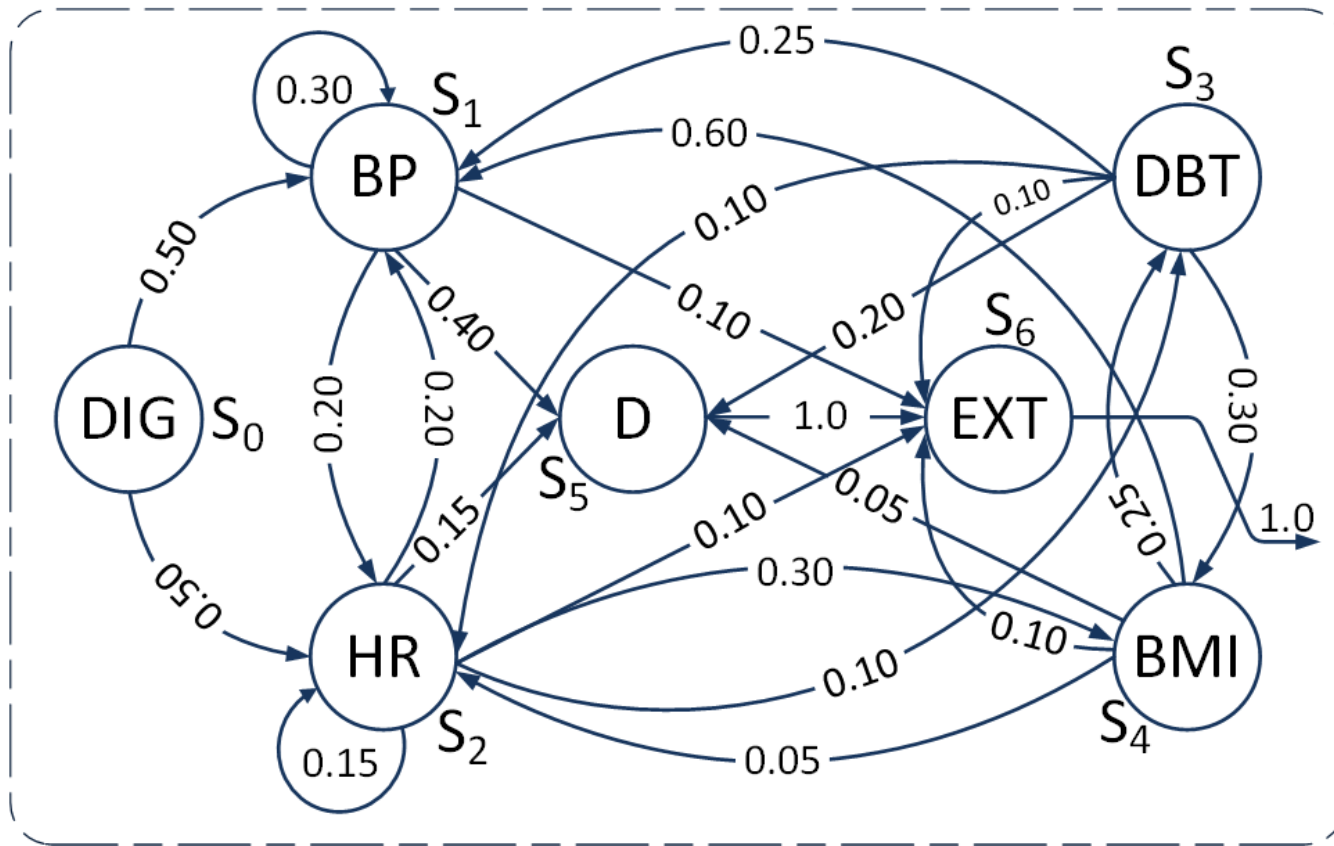


Figure B11: the Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.60” and “BMI to HR = 0.05”.

The figure B12 describes the graphical representation of Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP = 0.60” and “BMI to HR = 0.05”, as follows:

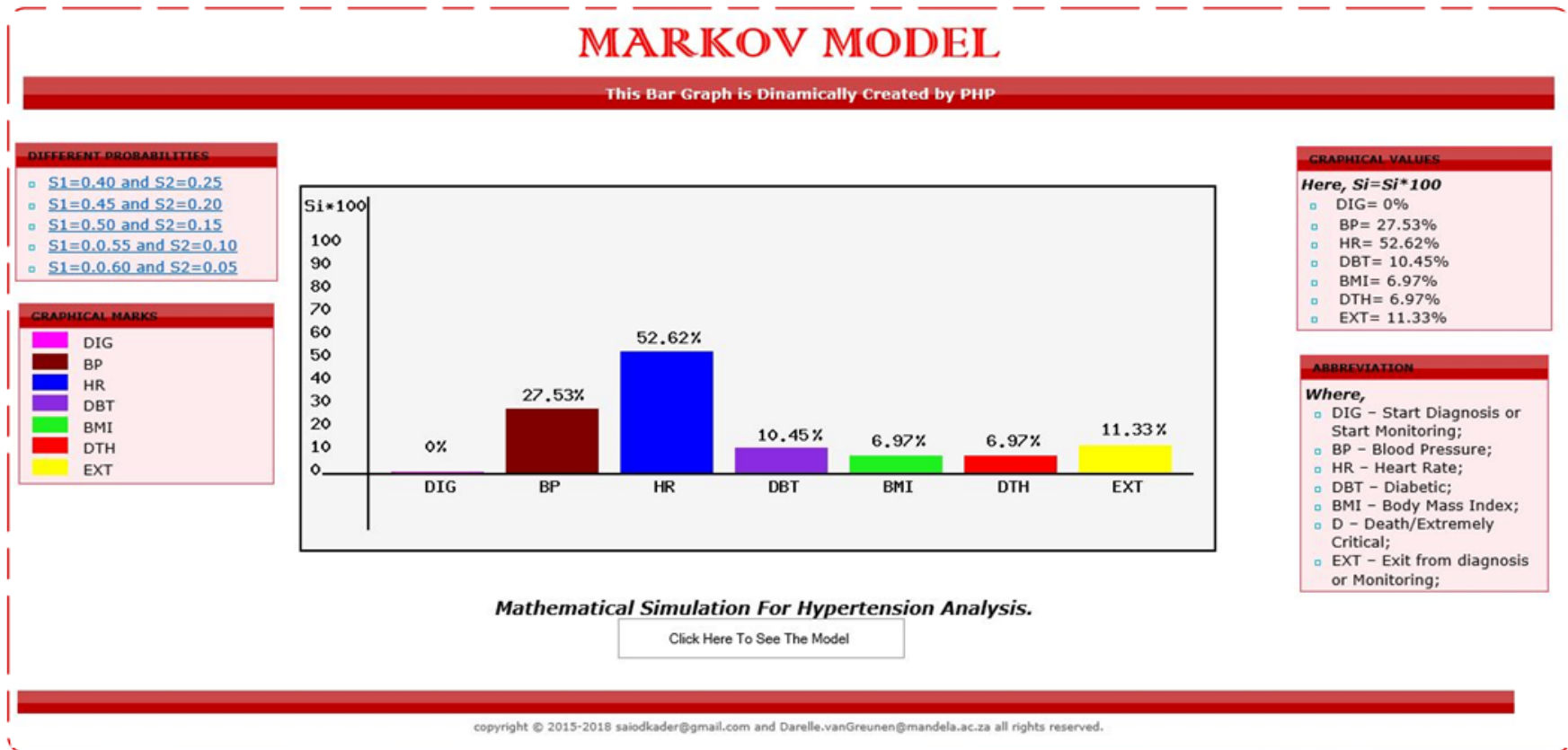


Figure B12: the graphical representation of Markov Chain Probability link structure of Hypertension progression risk model when “BMI to BP 0.60” and “BMI to HR = 0.05”.