

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

Realism in Synthetic Data Generation

A thesis presented in fulfilment of the requirements for the degree of:

Master of Philosophy
in
Science

Scott McLachlan

(MCSE, MCT, DipSysEng, GradDipInfSc, GradDipLaw, GradDipBus, MIITP, MBCS)
School of Engineering and Advanced Technology
Massey University
Palmerston North, New Zealand

Supervised by:

Dr. Kudakwashe Dube
School of Engineering and Advanced
Technology
Massey University
Palmerston North, New Zealand

Prof. Thomas Gallagher
Applied Computing and Engineering Technology
Missoula College
University of Montana
Missoula, USA

2017

Copyright is owned by the Author. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. This thesis may not be reproduced or disseminated elsewhere without the express written permission of the Author.

Abstract

There are many situations where researchers cannot make use of real data because either the data does not exist in the required format or privacy and confidentiality concerns prevent release of the data. The work presented in this thesis has been undertaken in the context of security and privacy for the Electronic Healthcare Record (EHR). In these situations, synthetic data generation (SDG) methods are sought to create a replacement for real data. In order to be a proper replacement, that synthetic data must be *realistic* yet no method currently exists to develop and validate realism in a unified way. This thesis investigates the problem of characterising, achieving and validating realism in synthetic data generation. A comprehensive domain analysis provides the basis for new characterisation and classification methods for synthetic data, as well as a previously undescribed but consistently applied generic SDG approach. In order to achieve realism, an existing knowledge discovery in databases approach is extended to discover realistic elements inherent to real data. This approach is validated through a case study. The case study demonstrates the realism characterisation and validation approaches as well as establishes whether or not the synthetic data is a realistic replacement. This thesis presents the ATEN framework which incorporates three primary contributions: (1) the THOTH approach to SDG; (2) the RA approach to characterise the elements and qualities of realism for use in SDG, and finally; (3) the HORUS approach for validating realism in synthetic data. The ATEN framework presented is significant in that it allows researchers to substantiate claims of success and realism in their synthetic data generation projects. The THOTH approach is significant in providing a new structured way for engaging in SDG. The RA approach is significant in enabling a researcher to discover and specify realism characteristics that must be achieved synthetically. The HORUS approach is significant in providing a new practical and systematic validation method for substantiating and justifying claims of success and realism in SDG works. Future efforts will focus on further validation of the ATEN framework through a controlled multi-stream synthetic data generation process.

Publications related to this thesis:

McLachlan, S., Dube, K., & Gallagher, T. (2017). Managing Realism in Synthetic Data Generation. *Manuscript submitted to JAMIA*.

McLachlan, S., Dube, K., & Gallagher, T. (2017). THOTH: The generic approach to and characterisation of Synthetic Data. *Manuscript submitted to JAMIA*.

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., & McLachlan, S. (2017). Synthea: An approach, method and software mechanism for generating synthetic patients and the synthetic electronic healthcare record. *Manuscript submitted to JAMIA*.

McLachlan, S., Dube, K., & Gallagher, T. (2017). The Realistic Synthetic Electronic Health Record: Challenges, rationale and future directions. *Manuscript submitted to JAMIA*.

McLachlan, S., Dube, K., & Gallagher, T. (2016). Using the CareMap with health incidence statistics for generating the realistic synthetic electronic health record. *IEEE International Conference on Healthcare Informatics, ICHI'16*.

Glossary

ATEN	The ATEN framework is an SDG lifecycle incorporating the THOTH, RA and HORUS approaches.
AU DoH	Australian Department of Health
CPG	Clinical Practice Guideline
HiKER Group	Health Informatics and Knowledge Engineering Research Group
HIS	Health Incidence Statistics
HORUS	Uses the knowledge developed by RA as the basis for validating realism in synthetic data and justifying success in SDG.
NZ MoH	New Zealand Ministry of Health
RA	A systematic approach used to discover realistic elements, characteristics and rules necessary to the creation of realistic synthetic data.
PK	Primary Key
SDG	Synthetic Data Generation
THOTH	The generic approach for SDG

Dedicated for Danika, Thomas, Liam and James.

Acknowledgements

I acknowledge with the greatest of appreciation the assistance of my supervisors and the wider members of the Health Informatics and Knowledge Engineering Research (HiKER) Group who supported my development as a researcher in the tradition of the scientific method. The support of my proof reader and sometimes editor who every day pointed out when my references were out of order and when what I had written didn't actually say what I thought I had said.

And I can't leave out Master 4, who recognised that I focus and work better when I have multiple streams of input and things to think about. Using this as only a four-year old can; as justification for continually distracting me with games, puzzles, stories and an insatiable need for me to join him as he played with his vast collection of toy trains. My hope is that I live to see the day when my encouragement of you culminates in my receiving a copy of your own thesis. I especially look forward to discussions about the distractions you had to deal with.

There are scores of others with whom I have interacted during the eight months spent researching and writing this thesis. But for the fact that it would take vast amounts of time and far more space than I am given on this page to single you all out, I offer my best wishes and thanks.

Scott

February, 2017.
Sydney, Australia.

To the reader;

The fact that you have chosen to pick up or download this thesis is an act that in and of itself deserves thanks. If nothing else, and in deference to the content, this single act justifies this thesis' existence.

Thank you.

This thesis is also a tribute to the late bloomers. People like Nikola Tesla, Charles Darwin, Samuel Jackson and Richard Adams. To all those who didn't even begin to realise their vast potential until later in life.

Table of Contents

ABSTRACT	3
TABLES.....	9
FIGURES.....	9
1. INTRODUCTION.....	11
1.1 INTRODUCTION	11
1.2 RESEARCH PROBLEM	13
1.3 SIGNIFICANCE	13
1.4 CHALLENGES.....	14
1.5 RESEARCH AIM	14
1.6 RESEARCH OBJECTIVES AND CHALLENGES	14
1.7 FUNCTIONAL GOALS.....	15
1.8 THESIS STRUCTURE	17
2. LITERATURE REVIEW	22
2.1 INTRODUCTION	22
2.2 IDENTIFICATION OF REVIEW LITERATURE	23
2.3 SYNTHETIC GENERATION STUDIES.....	23
2.4 SDG’S RELATIONSHIP TO COMPUTATIONAL MODELLING.....	24
2.5 VALIDATION OF THE COMPUTATIONAL MODEL.....	24
2.6 VALIDATION TECHNIQUES	26
2.6.1 <i>Grounding</i>	26
2.6.2 <i>Calibrating</i>	26
2.6.3 <i>Verification</i>	27
2.6.4 <i>Harmonising</i>	27
2.7 THE INCOMPLETENESS OF PUBLISHED SDG METHODS.....	27
2.8 SUMMARY	28
3. RESEARCH METHODOLOGY.....	32
3.1 METHOD FOR IDENTIFYING SDG LITERATURE: FUNCTIONAL GOAL 2	32
3.2 METHOD FOR CHARACTERISING SYNTHETIC DATA: FUNCTIONAL GOAL 3.....	34
3.3 METHOD FOR IDENTIFYING THE GENERIC APPROACH TO SDG: FUNCTIONAL GOAL 4.....	34
3.4 METHOD FOR APPLYING EXISTING VALIDATION METHODS TO SDG: FUNCTIONAL GOAL 5	34
3.5 METHOD FOR DEFINING REALISM IN SDG: FUNCTIONAL GOAL 6.....	35
3.6 METHOD FOR CHARACTERISING REALISM IN SDG: FUNCTIONAL GOAL 7.....	35
3.7 METHOD FOR DEFINING VALIDATION OF REALISM IN SDG: FUNCTIONAL GOAL 8	35
3.8 CASE STUDY METHODOLOGY.....	35
3.9 SUMMARY	37
4. SYNTHETIC DATA GENERATION.....	40
4.1 INTRODUCTION	40
4.2 BACKGROUND.....	40
4.2.1 <i>The attachment of pre-eminence in Fully Synthetic Data to Rubin</i>	41
4.2.2 <i>Extending the History of Synthetic Data</i>	42
4.3 APPROACHES AND METHODS FOR SYNTHETIC DATA GENERATION.....	43
4.3.1 <i>Data Masking</i>	44
4.3.2 <i>Signal and Noise</i>	44
4.3.3 <i>Network Generation</i>	45
4.3.4 <i>Music Box Method</i>	45
4.3.5 <i>Markov Chain Method</i>	45
4.3.6 <i>Monte Carlo Method</i>	45
4.3.7 <i>Walker’s Alias Method</i>	46
4.3.8 <i>Distribution of Methods and Domains in SDG</i>	46
4.4 DIFFERENTIATION FOR CLASSIFICATION	47
4.5 THE ATEN FRAMEWORK	48

4.6 CASE STUDY: INTRODUCTION	50
4.7 CONCLUSION	52
5. THOTH: THE SDG GENERIC APPROACH.....	56
5.1 INTRODUCTION TO THOTH.....	56
5.2 THE STEPS TO SDG.....	57
5.3 DISCUSSION OF THE GENERIC APPROACH	58
5.4 IMPROVING THE GENERIC APPROACH WITH THOTH	59
5.5 CONCLUSION	59
6. VALIDATION METHODS FOR THE SDG GENERIC APPROACH.....	62
6.1 INTRODUCTION TO SDG VALIDATION	62
6.2 SIMPLIFIED GENERALISED NARRATIVE OF PUBLISHED SDG ARTICLES	62
6.3 IMPROVING THE SDG GENERIC APPROACH WITH VALIDATION	64
6.4 VALIDATION APPROACHES IN THE DOMAIN OF COMPUTATIONAL MODELLING.....	65
6.4.1 <i>Grounding</i>	66
6.4.2 <i>Calibration</i>	67
6.4.3 <i>Verification</i>	67
6.4.4 <i>Harmonising</i>	68
6.5 CASE STUDY	70
6.5.1 <i>Grounding Validation</i>	70
6.5.2 <i>Calibration Validation</i>	70
6.5.3 <i>Verification Validation</i>	70
6.5.4 <i>Harmonising Validation</i>	70
6.5.5 <i>The Improved Generic Approach</i>	71
6.6 CONCLUSION	71
7. REALISM	74
7.1 INTRODUCTION TO REALISM	74
7.2 THE REALISM COMPONENT OF CURRENT SDG LITERATURE	74
7.3 DEFINING REALISM FROM THE LITERATURE	75
7.3.1 <i>Understanding Realism</i>	75
7.4 REALISM AND THE SCIENTIFIC METHOD	76
7.5 CONCLUSION	77
8. RA: A GENERIC APPROACH FOR REALISM.....	80
8.1 INTRODUCTION	80
8.2 IDENTIFYING REALISTIC ELEMENTS FROM THE REAL DATA.....	81
8.3 DIFFERENTIATING THE SUBSTANCE OF DATA	81
8.3.1 <i>Quantitative Characteristics</i>	82
8.3.2 <i>Qualitative Characteristics</i>	82
8.4 KNOWLEDGE DISCOVERY IN DATABASES (KDD).....	82
8.4.1 <i>HCI-KDD</i>	84
8.5 RA: THE ENHANCED KDD APPROACH	85
8.5.1 <i>Concept Hierarchies</i>	85
8.5.2 <i>Formal Concept Analysis</i>	86
8.5.3 <i>Characteristic and Classification Rules</i>	86
8.6 CASE STUDY: VALIDATION OF THE RA APPROACH	88
8.6.1 <i>Quantitative Aspects</i>	88
8.6.2 <i>Qualitative Aspects</i>	89
8.6.3 <i>Applying KDD</i>	89
8.6.4 <i>Concept Hierarchy</i>	90
8.6.5 <i>Formal Concept Analysis</i>	90
8.6.6 <i>Characteristic Rule</i>	92
8.6.7 <i>Classification Rule</i>	94
8.6.8 <i>Case Study: Discussion</i>	95
8.7 CONCLUSION	96
9. THE HORUS APPROACH TO VALIDATION OF REALISM.....	98

9.1 INTRODUCTION	98
9.2 APPLICATION OF THE HORUS APPROACH	99
9.2.1 <i>Input Validation</i>	99
9.2.2 <i>Realism Validation 1</i>	100
9.2.3 <i>Method Validation</i>	100
9.2.4 <i>Output Validation</i>	101
9.2.5 <i>Realism Validation 2</i>	101
9.2.6 <i>Validation: Discussion</i>	101
9.3 CASE STUDY: APPLICATION OF THE VALIDATION APPROACH	103
9.3.1 <i>Input Validation</i>	103
9.3.2 <i>Realism Validation 1</i>	104
9.3.3 <i>Method Validation</i>	104
9.3.4 <i>Output Validation</i>	105
9.3.5 <i>Realism Validation 2</i>	107
9.4 DISCUSSION AND SUMMARY	107
10. CONCLUSION.....	112
REFERENCES.....	117
APPENDIX A: SYNTHETIC DATA GENERATION LITERATURE	128
APPENDIX B: REALISM IN SDG APPROACHES	132
APPENDIX C: A REVIEW OF THE KARTOUN SDG METHOD	138
APPENDIX D: A REVIEW OF THE SYNTHEA SDG METHOD	141

Tables

Table 1: Established Classifications for Computational Models	25
Table 2: Comparison of Rubin (1993) to Birkin & Clark (1987)	43
Table 3: Characterisation of Synthetic Data Generation Methods.....	46
Table 4: Classification of Synthetic Data	48
Table 5: Simplified Generalised Narrative of SDG Articles.....	63
Table 6: Justification Examples for Part 1 of the Simplified Generalised Narrative.....	63
Table 7: Operational examples for Part 2 of the Simplified Generalised Narrative	63
Table 8: Result examples for the Simplified Generalised Narrative	64
Table 9: Ethnicity Statistics for births at CMDHB in 2012 (expressed as percentages).....	88
Table 10: Age Statistics for births at CMDHB in 2012 (expressed as percentages)	88
Table 11: Midwifery Patient Database Patient Relational Table Schema extract.....	89
Table 12: Formal Concept Analysis for 10 Random Labour and Birth Patients	92
Table 13: Generalised Relation Table	94
Table 14: The qualitative classification rule for Caesarean based on previous mode/s of delivery	94
Table 15: Realism Validation Questions	100
Table 16: CoMSER Input Validation Case Study.....	104
Table 17: Demographic Analysis Table from CoMSER CoMENGINE	106
Table 18: Ethnicity Statistics Comparison.....	106
Table 19: Age Statistics Comparison.....	106
Table 20: Synthetic Data Generation Literature	128
Table 21: Realism in SDG Approaches	132
Table 22: Sample gender-specific conditions from the Kartoun (2016) EMR dataset.....	139
Table 23: Ten Random Patients from Kartoun (2016).....	139
Table 24: Documents provided by the Synthea Team.....	141
Table 25: Additional Sources for Type2 Diabetes Validation Data.....	142

Figures

Figure 1: The Signpost Diagram used throughout this thesis.....	17
Figure 2: SDG Literature Search and Categorisation	33
Figure 3: Distribution of SDG Methods	47
Figure 4: Distribution of SDG Domains.....	47
Figure 5: The ATEN Framework	49
Figure 6: Context Diagram for the CoMSER Method (from McLachlan et al, 2016)	51
Figure 7: CoMSER UML Activity Diagram (from McLachlan et al, 2016)	51
Figure 8: The Generic Approach to Synthetic Data Generation	57
Figure 9: The three-step THOTH approach.....	59
Figure 10: The Improved Generic Approach to Validation for Synthetic Data Generation.....	65
Figure 11: Grounding Validation of the Generic Approach.....	66
Figure 12: Calibration Validation of the Generic Approach.....	67
Figure 13: Verification Validation of the Generic Approach.....	68
Figure 14: Harmonising Validation of the Generic Approach.....	69
Figure 15: The KDD Process	84
Figure 16: Midwifery Patient Database Relational Schema extract.....	89
Figure 17: Concept Hierarchy for Child Birth	91
Figure 18: Concept Hierarchy for Child Birth with Statistics	91
Figure 19: Concept Lattice example	93
Figure 20: Characteristic Rule from the domain of Midwifery.....	94
Figure 21: Classification Rule from the domain of Midwifery.....	95
Figure 22: The HORUS approach embedded into THOTH	102
Figure 24: Synthea Validation Review: Diabetes Prevalence.....	143
Figure 25: Age at Diagnosis of Type-2 Diabetes Mellitus.....	145

“Behind every algorithm there is always a person. A person with a set of personal beliefs that no code can ever completely eradicate. You must identify your own personal bias. You need to understand that you are human and take responsibility accordingly.”

(Ekstrom, 2015)