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

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	Heath 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

Tables

Figures

"*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)

1. Introduction

1.1 Introduction

Datasets are a common requirement of research, experimentation, software testing and systems training. Sourcing or generating experimental data can be costly and often presents an insurmountable challenge (Bozkurt & Harman, 2011; Whiting, Haack & Varley, 2008; Williams et al, 2007) The use of real data carries the risk of exposing confidential information and for this reason initial efforts in synthetic data generation focused not on any actual generation method, but on identifying and replacing these personally identifying details (Rubin, 1993).

The use of synthetic datasets has developed from a new concept not more than thirty years ago, to one that is routinely used and even seen now as the solution to training artificial intelligences (Weston et al, 2015). Current literature contains a wide array of methods and models with no definitive text to provide those considering a synthetic data generation (SDG) project with direction as to which may be the most appropriate for any given purpose. The literature includes everything from processes that replace, scramble or de-identify real data (Mouza et al, 2010), through to those that avoid the use of real records due to privacy concerns, preferring instead to use statistics and frequency distributions to generate synthetic records intended to be accurately representative of real-world incidence (McLachlan, Dube & Gallagher, 2016).

There are many reasons why we might generate synthetic data. Testing software or systems (Barse, Kvarnström & Johnson, 2003; Mouza et al, 2010; Houkjaer et al, 2006; Whiting et al, 2008), population synthesis (Gargiulo et al, 2010), testing scientific hypotheses or to generate seed data for simulations (Srikanthan & McMahon, 2001; Wan et al, 2008). While generating data might sound simple, generating good synthetic data can be extremely difficult (Whiting et al, 2008). A key reason for generating synthetic data is limiting the release of confidential or personally identifiable information inherent to the use of real data sources (Killourhy & Maxion, 2007; Mouza et al, 2010; Sperotto et al, 2009; Zanero, 2007) however some approaches still use these real sources either directly, or as seed data in the generation process (Ascoli, 2001; Barse, 2003; Bozkurt, 2011). If a poorly designed or incorrect model is used, these methods can still carry the risk of exposing that confidential or personally identifiable information.

Some claimed the use of synthetic experimental data reduced precision (Nguyen & Leung, 2009; Tsuzuki & Tanabi, 1991; Venti, 1984). Many now promote the use of synthetically generated data (Kanungo & Resnik, 1999; Lujano-Rojas et al, 2013), seeing simulated experimental data as a

mechanism that can bring greater flexibility, accuracy and validation to their models (Begue et al, 2011; Elliot et al, 2002; Mora et al, 2015). We do this under the shadow of a caution counselling us that synthetic data can never quite be a match for the real thing (Simonovic, 2012).

It isn't enough to simply generate random data. The data must be suitable to the purpose for which it will be used (Lundin et al, 2002). At the simplest level the contents of any field may be required to fall within a defined set of constraints; for example, the telephone number field should contain numbers, and not just any numbers but a collection that at least resemble the pattern of a useable phone number. Many recent projects require increasingly more intricate and complicated datasets where it isn't just the value in a single field that should be valid, but where the entire dataset is constructed to be representative of and indistinguishable from observed data (Stratigopoulos, Mir & Makis, 2009; Wu, Wang & Zheng al, 2003). It is at this point where authors draw our attention to the notion of *realism* in synthetic data.

Realism, like SDG, is another concept where a multitude of approaches and understandings exist, albeit in the absence of concrete understanding for what it *means* for data to be *realistic*. Most authors recognise the need for realism (Bozkurt et al, 2011; Killourhy et al, 2007; Jaderberg et al, 2014) however many leave the concept of realism open to generic interpretation, making it difficult to both understand how they intend to imbue their data with realistic characteristics, and how we should assess the success of their intentions after the data is created.

The common sense implication to be drawn from the term *realistic* would seem to engender synthetic data that is, as Jaderberg et al (2014) succinctly puts it, "*sufficient to replace real data"*. However, this goal is all too often an unspoken implication left wholly for the reader to infer. Realism to some is an outcome, a target that their dataset aims to possess (Speretto et al, 2009). Others describe realism as a *characteristic* we might see (Pudjijono et al, 2009). These approaches appear retrospective rather than intentional. It would seem that any approach to realism should begin by identifying what is required from the outset, the form that realism should take, providing a definition for realism and an understanding of the characteristics and goals that we could measure in order to validate that realism has been achieved in the resulting synthetic data. Such an approach would provide clarity and establish a means for post-generation assessment. A way to understand both what was wanted, and whether or not it was achieved.

This research intends to develop a conceptual approach for systematically defining, identifying and validating for realism in synthetic data. In doing so it will use a review and catalog of synthetic data generation and realism implementation approaches taken from the published literature, along with established validation approaches used in SDG or its precursor, computational modelling. In execution of the aims and objectives of this research we will come to realise that there exists a generic approach which is common to all SDG methods, and investigate how established validation methods are or may be applied to that approach. It is in comprehending how realism and its validation can be situated within

the generic approach to SDG that we can come to establish the systematic strategy and technique for ensuring realism, and the commensurate comprehensive verification model necessary for its validation.

The method employed by this research begins with a literature review and analysis of synthetic data generation; what it is, why we do it and the range of methods, models and solutions that are commonly employed. The output from this step will identify a range of SDG approaches, methods and models. The literature review and analysis will enable understanding of: (a) what methods are used in SDG; (b) what fields of research are engaging in SDG; (c) whether there are SDG methods that are preferred or used more often in a given field of study; and (d) the SDG methods and research fields where realism was or was not identified as an intended goal or feature. The realisations gleaned from this information feeds into the next part of the method. They provide a basis from which we can draw conclusions and understandings of the use of realism. From this we can generate hypotheses for a definition and validation strategy for realism in SDG methodologies. We will also be able to identify the characterisations necessary to test and validate that realism in the resulting synthetic data.

1.2 Research Problem

Documentation, testing and repeatability are key principles that underpin the scientific method (Crawford & Stucki, 1990; Creswell, 2003), or to quote Adam Savage of Mythbusters, "*The only difference between screwing around and science is writing it down".* A synthetic data generation (SDG) model that is not sufficiently robust or complete in its definition and starting data characterisation can only deliver questionable or unreliable synthetic data (Birkin, Turner & Wu, 2006). It is said that an incorrect method would be preferable to an incomplete one (Bolon-Canedo et al, 2013). For the majority of methods explored during this research an element of the SDG method, the realism, remains both undefined and lacking of testable predictions. These are necessary in order to validate that the level and characteristics of realism sought by the authors has been delivered in the resulting synthetic dataset.

This research work seeks to identify a simple, useable definition for *realism* in the context of synthetic data. To establish a simple and easily applied approach to identifying necessary realism characteristics, and a systematic process for the validating the presence of those characteristics in the synthetic dataset. These definition, identification and validation approaches will ensure future SDG methods come with a greater degree of accuracy, utility, strength and repeatability, sufficient to support the expected claims of success.

1.3 Significance

The findings of this study will provide a framework to ensure that realism can be a deliberate, systematically defined and assessable component of SDG methods. Those projects applying the

methods discussed here will be able to ensure the necessary realism characteristics are identified and incorporated and that the inclusion of a systematic approach to realism at the outset of the SDG project results in a more robust, complete SDG method. A complete SDG method facilitates the development of an assessment approach for ensuring that realism characteristics can be observed and validated in the resulting synthetically generated dataset.

1.4 Challenges

The primary challenges to developing a general definition and approach to realism comes from the breadth of methods and sheer diversity of scientific fields now employing synthetic data generation. This will require reviewing and concentrating a considerable amount of information in order to arrive at an abstract, comprehensive and domain-independent response.

There is also the almost informal way that the term realism has been attached to many of these synthetic data projects, like an afterthought, almost as if to lend them an additional air of validity in the reader's eye. This may go some way to explain the lack of a formal testing method for realistic characteristics in the resulting datasets of most models. It is difficult to assess whether this results simply from the absence of a formal or recognised structure around realism, a focus on the generation method that doesn't allow the time or capacity to look at the characteristics of that method or resulting data, or is possibly indicative of a potential issue for this piece of work, that is, a complete unwillingness to expend the extra effort of defining realism characteristics and the scope for a testing method. To some authors it is possible that this extra effort required may not provide a sufficient return for them to engage in the process.

1.5 Research Aim

The purpose of this research project is to systematically develop the concept of realism in SDG and develop strategies for its realisation and validation. Such efforts should produce a unified definition for *realism* in data and establish an approach to producing *realistic* datasets that is sufficiently robust to suit the purposes of the largest possible number of disciplines and study types.

1.6 Research Objectives and Challenges

The research objectives and challenges are:

1. *Conceptual development of realism*
    a. What is realism in synthetic data generation?
    b. What properties or characteristics are considered realistic?
    c. To what degree is realism a factor when generating synthetic data?

d. Are solutions to synthetic data generation domain specific and is realism similarly domain-specific?

2. *Development of systematic approach*
    a. How can we characterise realism as it relates to the process of SDG?
    b. How can realism be systematically incorporated into models and methods for SDG?
    c. How can we measure or verify that realism as prescribed exists within the resulting synthetic dataset?

3. *The approach to realism*
    a. What approaches can we integrate into our generation methods to deliver and validate realism in synthetic data?

1.7 Functional Goals

While the research aims and objectives focus on the overall scope and questions the research proposes to answer, the following functional goals are intended to be more practical, tightly focus and action-oriented. They describe what each element or segment, in this case chapter, will present to the reader. Functional goals are in essence a signpost that set expectations and guide the reader in their journey through each section of the knowledge this research delivers.

The functional goals of this thesis are:

Functional Goal 1. Investigate and review related works of academics and researchers and provide a detailed review of the main topics and issues relevant to the scope of this research.

Functional Goal 2. Develop and apply a classification scale and ontology to describe synthetic data along with a catalog of observed generation methods.

Functional Goal 3. Characterise SDG Methods and Synthetic Data

Functional Goal 4. Develop the generic approach to Synthetic Data Generation.

Functional Goal 5. Develop a new method for SDG validation based on established methods in CM.

Functional Goal 6. Develop a definition for realism in SDG.

Functional Goal 7. Develop a generic approach to identifying and characterising realism for SDG.

Functional Goal 8. Develop the realism validation processes and integrate them into the SDG validation method.

1.8 Thesis Structure

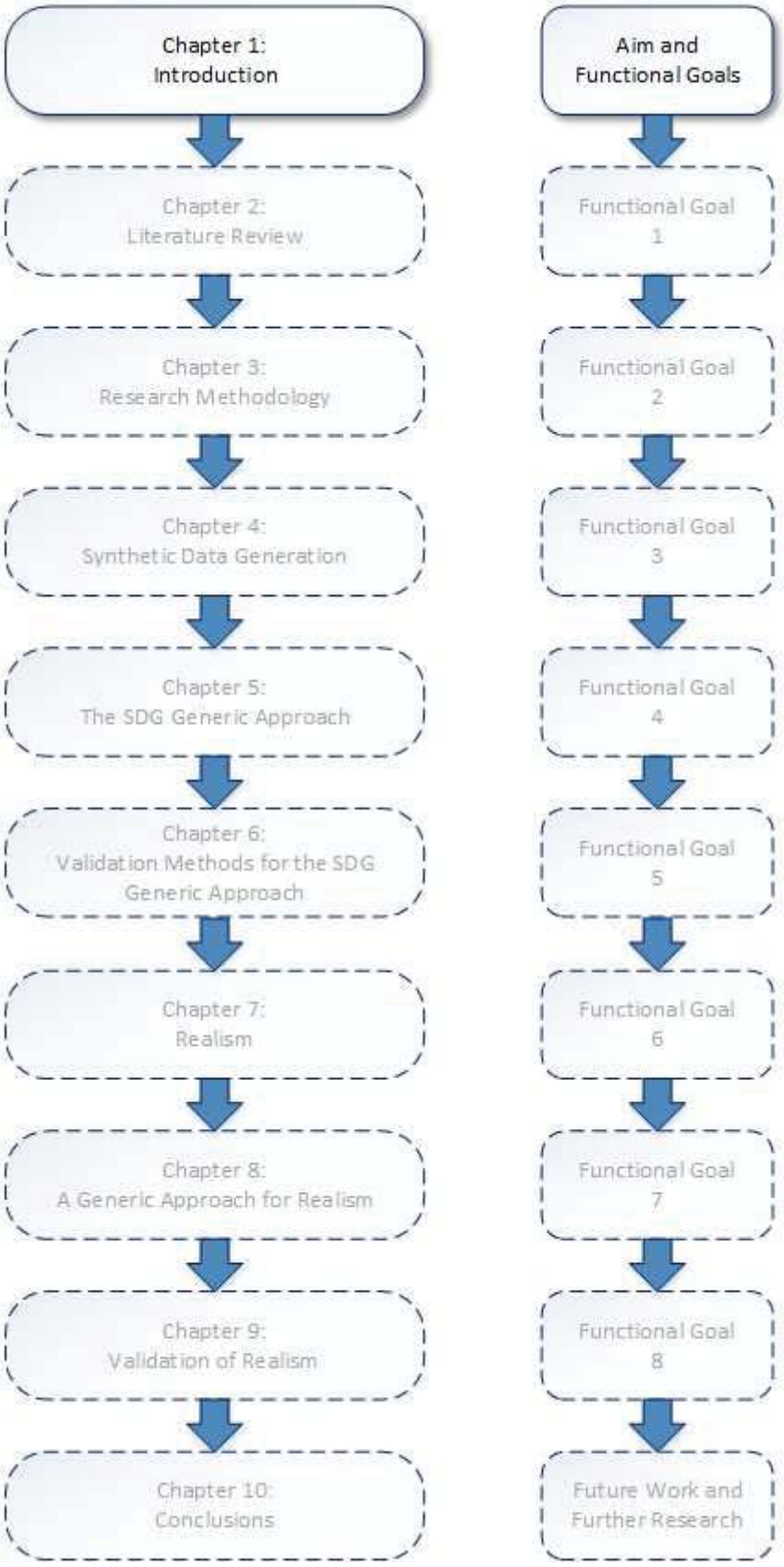


Figure 1: The Signpost Diagram used throughout this thesis

Chapter 2 investigates literature which may have reviewed or presented a model that formalises generation methods or the concept of realism in SDG (functional goal 1). It finds no prior work that presenting a complete picture of either. This chapter also describes how current SDG methods and practices can be contrasted to those of computational modelling (CM) and how CM validation methods may be applicable to SDG practices. For the benefit of the reader, literature regarding other aspects not directly covered here is considered in the relevant chapter.

Chapter 3 describes the identification and selection of SDG literature for inclusion in this research (functional goal 2). A reference list of SDG literature is introduced for use in later chapters. This literature is used to establish the range of generation methods, domains where those methods are applied, and any validation approaches employed in SDG. It is from the literature in this list that knowledge and conclusions as to definition and approach can be drawn.

Chapter 4 presents the background, exploration and analysis of SDG. It was necessary to understand SDG before we could draw our focus toward the realism aspect (functional goal 3). It characterises the observed synthetic data generation methods into five types and concludes by presenting a classification scale for the synthetic nature, or *syntheticness*, of the data created by these SDG methods.

Chapter 5 presents the generic approach to SDG. It discusses the inherent weaknesses and limitations in this approach (functional goal 4) and proposes the THOTH approach to mitigate some of those limitations.

Chapter 6 describes four established CM validation methods and applies them within the context of the THOTH approach to SDG (functional goal 5). It proposes an improved method for the conduct of the THOTH generic approach to SDG that more closely aligns with the ideals and principles of the scientific method.

Chapter 7 introduces the concept of realism, discusses the nature of realism and contextualises realism as it is seen in current SDG approaches (functional goal 6).

Chapter 8 presents an approach to identifying realism for use in SDG (functional goal 7).

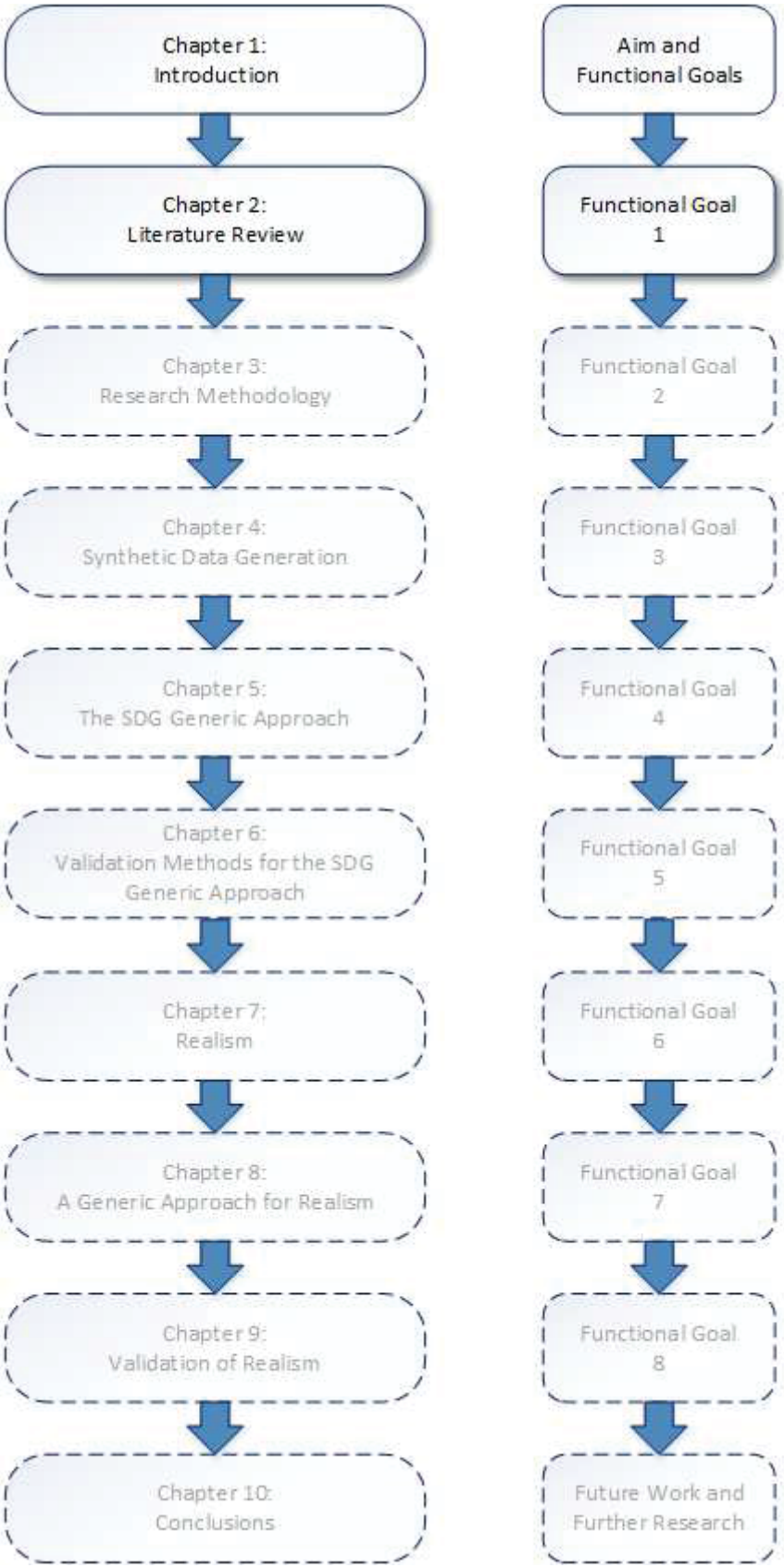
Chapter 9 presents an approach that relies on the realism identification model as a tool to providing the knowledge needed to validate realism in synthetic data (functional goal 8).

Chapter 10 concludes the thesis by describing each significant component presented in the thesis. It provides recommendations for the integration of the developed realism concept in future SDG projects and proposes future research and experimentation to further establish and extend realism validation.

This page intentionally left blank

"*The Scientific Method* is so powerful that in a mere four centuries it has taken us from Galileo's first look through a telescope at another world to leaving our footprints on the moon. It allowed us to look out across space and time to discover where, and when, we are."

(Neil deGrasse Tyson, Cosmos: A Spacetime Odyssey, 2014)


Chapter 1:
Introduction
Chapter 2:
Literature Review
Chapter 3:
Research Methodology
Chapter 4:
Synthetic Data Generation
Chapter 5:
The SDG Generic Approach
Chapter 6:
Validation Methods for the SDG
Generic Approach
Chapter 7:
Realism
Chapter 8:
A Generic Approach for Realism
Chapter 9:
Validation of Realism
Chapter 10:
Conclusions
Aim and
Functional Goals
Functional Goal
1
Functional Goal
2
Functional Goal
3
Functional Goal
4
Functional Goal
5
Functional Goal
6
Functional Goal
7
Functional Goal
8
Future Work and
Further Research

2. Literature Review

2.1 Introduction

This chapter reviews the literature relevant to the main research aim of this thesis. The review initially seeks to locate any literature that has previously reviewed SDG methods, specifically seeking to find any that have focused on the realism element. It explores literature on SDG methods and projects generally, looking carefully at those which describe an element of realism in their approach, intent on understanding how authors have defined and described this realism and any relevant citations that were used. The scope of this literature review is limited in order to provide the necessary foundation for the remainder of this thesis. For the benefit of the reader, literature specific to other topics is identified and considered within the relevant chapters. Accordingly, this literature review is more concise than would otherwise be the case for this type of thesis. The research in this chapter was carried out in order to achieve functional goal 1:

Functional Goal 1. Investigate and review the related works of academics and researchers, providing comprehensive detailed literature review on the main topics relevant to the research conducted in this thesis.

This chapter is structured as follows:

2.2 Identification of Review Literature
2.3 Synthetic Generation Studies
2.4 SDG's Relationship to Computational Modelling
2.5 Validity of the Computational Modelling
2.6 Validation Techniques
2.7 Incompleteness in Published SDG Methods

2.8 Summary

2.2 Identification of Review Literature

Very few studies have analysed, compared, contrasted or reviewed synthetic data generation approaches or models (Smith, Clarke & Harland, 2009). An initial search was conducted across Massey University's online catalog systems, which access a range of journals hosted by Elsevier, Sage, the Directory of Open Access Journals and others that are offered through a connection to the EBSCOhost research databases. Given that only one small review article was located, the search was expanded to also include the wider catalogue of Google Scholar. In total, four articles were located that discussed what were small, focused and field-specific generation methods. A fifth was found that chronologically reviewed the history of the specific issue of class imbalances in learning algorithms (Chawla, Japkowics & Kotcz, 2004). These learning algorithms feature in a number of SDG methods such that the article was, in a very indirect and minor way, of interest to this study.

Given the extremely low number of studies that have looked at SDG methods, and the contrastingly large amount of literature focused on the use of SDG methods to deliver data for a number of functional causes, it would appear that this represents a gap in the literature open to future consideration. A systematic analysis or review of SDG methods would aid researchers, providing baseline information on each SDG model and expediting method selection.

2.3 Synthetic Generation Studies

Smith et al's (2009) meta-analysed results from a variety of experiments, drawing focus on the way in which constraining variables and statistical weighting are applied during the SDG process. Their work delivers a series of solutions specific only to micro-population modelling.

Bolon-Canedo et al's (2013) work looked at how the selection of input features affects algorithms for data learning models. Being a review of input data learning methods, it was difficult to apply much of their research to SDG, or output models.

Srikanthan and McMahon (2001) review a series of rainfall generation models which mainly consisted of variations on the Markov Chain model. They highlight a number of previously unrealised shortcomings with existing models that underestimate variance, and provide recommendations as to which model variations may assist in overcoming the issue in order to arrive at more accurate scenario modelling for future climate conditions (Srikanthan & McMahon, 2001).

Burton and Obel (1995) contemplate realism as it applies within the computational models of organisational science. They assert that validity is a balance of three elements: the question or purpose, the experimental design and the computational model, with a preference for simple methods that focus solely on addressing the research problem (Burton & Obel, 1995). While the purpose of the

computational model is a standalone element for consideration in their research, realism is left as a component of the experimental design seen more specifically as something that balances a model; too little and the computational model becomes a constricting logical mathematics exercise, too much and the purpose may be lost in the quest for absolute realism (Burton & Obel, 1995). They reference their consideration of realism and purpose within the context of the four category taxonomy of computational models developed by Cohen and Cyert (1965). Burton and Obel (1995) state that certain types of computational models must occupy a degree of implicit realism in order to be relevant, but for other models it is important only to be *close* to realistic, and yet more can be less realistic and still retain their relevancy.

2.4 SDG's Relationship to Computational Modelling

Computational modelling (CM) uses mathematics and computer science to simulate a formal representation of the complex system (Galan et al, 2009). Unlike SDG with its visibility in a wider range of academic fields, CM is seen more heavily in the psychological and sociological aspects of the Business and Finance sectors, and in the environmental forecasting arena.

While CM is a method of research and experimentation in its own right, there are similarities in the processes, approaches and formalisations that the concepts of validity and realism in both CM and SDG could share a degree of similarly. The two have even been observed using similar methodological approaches, such as Bayesian networks (Friedman et al, 2000), ensemble generation (Penzotti et al, 2002) and network generation (Bate et al, 1998). It is also not unusual to find an SDG approach that employs CM constructs in its method, especially when the synthetic data will be representative of complex systems and patterns (Ponzini et al, 2012; Velasquez, 1997;), or a CM project using data produced by SDG (Barnard et al, 2002).

The most common CM types are: agent-based, system dynamic and statistical forecasting. CMs have been classified in a number of ways (Carley, 1996). Table 1 describes the eight most documented CM classifications. Each CM model has a contrasting and often more detail-oriented model, e.g.: Intellective versus Emulative, Stochastic versus Deterministic (Skalka, 2009).

2.5 Validation of the Computational Model

Meaning and use of the term *validation* is often misunderstood (Oreskes, Shrader-Frechette & Belitz, 1994). The Oxford Dictionary (2016) tells us that the act of validation is one which seeks to check or prove accuracy. Validation in operation actually describes two related but detached concepts; *validation*, which establishes that a model, method or algorithm is correct, free from defect in form or function and internally consistent; and *verification*, which tests the output or predictions rendered by the model or method are consistent with observation (Oreskes et al, 1994). A simple way to remember the

application of these two terms is to consider that validation asks the researcher to consider whether they have done something in the right way, for example; *does this model use the correct algorithm constructed and applied in the right way?* By contrast, verification asks whether something is like something else, for example; *is the data output by the model what we see in real or observation data?* (Oreskes et al, 1994).

Table 1: Established Classifications for Computational Models

Classification	Description
Intellective	An intellective model contains analogous entities, constructs, and complexities of what is being modelled rather than mimicking each specific behavior. Typically simplistic (Carley, 1996), they are used to model a small number of (<100) agents (Lee & Carley, 2004).
Emulation	Emulation models, like Intellective, are multi-agent models however the emulative model captures a much higher level of detail (Lee et al, 2004) intended to imitate the externally observable behaviour in order to match an existing real system close enough that it could be substituted for the real system (Fabrega et al, 2013).
Stochastic	Stochastic (or statistical) models implement the probabilistic analyses from approaches such as probabilistic gate models (PGMs) (Chen et al, 2010). They typically have at least one random number generation component (Skalka, 2009). Unique input can lead to different outputs on each run due to the random component (Meng et al, 2013).
Deterministic	The deterministic model has no stochastic (random) element and assumes the outcome is fixed if the input is the same (Meng et al, 2013). It makes definite predictions for quantities without any associated probability (Bregt, 1997). They are considered less accurate than stochastic models (Bierkens & Geer, 2008).
Parameterised	A parameterised model runs many simulations, varying one or more variables over which the operator has control (Thiele, Kurth & Grimm, 2014).
Heuristic	Heuristics are those approaches not guaranteed to be optimal or perfect, but considered sufficient for the immediate goal of solving the problem more expediently. In some models the heuristics may involve short-cuts and guesswork on what might produce the desired solution (Harvey, 2007), or they can take the form of a collection of rules based on prior experiential knowledge (Gigerenzer, 1991).
Enumerative	There is efficiency to an enumerative model. It works iteratively through input data seeking to identify and return structure and hierarchy in the response (Sarkar & Boyer, 1994).
Monte Carlo	A range of algorithms from Walkers to Markov Chain Monte Carlo (MCMC) imputation models that use repeated random sampling to achieve numerical results within probabilistic distributions.

The term *validation* within the context of computational modelling is used to refer to the processes and techniques for ensuring and assessing comparability between simulated, or synthetic, and real data (Carley, 1996). This process by definition is assessing the realism of the synthetic data. Validation should not be a final product quality control check however, but a continual investigative process built in from the start of the research process (Kvale, 1994).

Validity in computational modelling has been defined in terms of content, construct, and criterion (Feldman & Arnold, 1983). These break down into six validation types: conceptual, internal, external, cross-model also known as model alignment, data and security (Kneppell & Aragno, 1993). Conceptual validity concerns the adequacy of the underlying theoretical model in characterising the real world; internal validity refers to the computer code being free of coding errors, and external validity is concerned with the adequacy and accuracy with which the computer model can match real-world data (Carley, 1996; Kneppell & Aragno, 1993). Cross-model validation (Carley, 1996) or model alignment is where we assess the degree to which two different models can produce similar outcomes (Axtell et al, 1996). Data validity assesses the accuracy of the data, both real and synthetic, and gauges whether the data is adequate, and the final validity type, security, is one of assuring that the model is safe from tampering or manipulation between operations or iterations (Carley, 1996).

2.6 Validation Techniques

A range of validation techniques have been established across a number of fields and previous research has identified them as falling into four main categories: *grounding, calibrating, verifying* and *harmonising*.

2.6.1 Grounding

More often used with intellective than emulation models, *grounding* establishes the reasonableness of the model and generally only establishes its validity at the pattern level (Carley, 1996). Grounding is enhanced by demonstrating that others have used similar rules and assumptions in their models (Carley, 1996). Initialisation grounding is typically used in the probabilistic approaches of stochastic and Monte Carlo type models to set the initial or starting conditions and procedures to ensure they are based on real data conditions.

2.6.2 Calibrating

Used more often with emulation models, *calibrating* is the often iterative process of tuning a model's predictions in order to show that the model generates results that closely match detailed real data (Fabrizio & Monetti, 2015). While exceedingly appropriate for validating rules-based or interactive

methods, it is argued that any model with sufficient parameters to tweak can be massaged so that some combination or configuration generates the observed data, therefore not actually establishing or validating the researchers' model in any meaningful way (Carley, 1996).

2.6.3 Verification

While not validating the inner workings of the model, and sometimes used in conjunction with the calibration approach (Carley, 1996) *verification* techniques focus on comparing statistically or graphically the output datasets (Kleijnen, 1995a) to real world data collected under conditions comparable to those represented within the model (Kleijnen, 1995b).

2.6.4 Harmonising

The most complicated of the validation methods, the *harmonising* approach is multi-step and requires two sets of real data with the aim of illuminating whether the theoretical assumptions within the CM are grounded, or in harmony, with the real world (Carley, 1996). The multi-step approach starts with running a calibration against the model with the first set of real data. This is followed by the second step; an estimation benchmark against a linear model. The third step involves cross-validation of the second set of real data using predictions flowing from the first set, while the final step concludes the validation by statistically contrasting both the CM and linear model's predictions for the second set of real data (Carley, 1996).

2.7 The incompleteness of Published SDG Methods

Review of SDG projects like those in the subset listed in the appendices has seen that the vast majority of these describe their experiments in an incomplete and unrepeatable manner. Repeatability has been described as the *Supreme Court of the scientific method*; incomplete experimental detail effectively defeats the ability of others to repeat the experiment and verify the claimed results for themselves (Collins, 1992; Stodden, 2010). While it may be the result of inattention, excitement or accidental and unintended omissions that led to the lapse in scientific methodology resulting in the absence of documentation for these elements of their experiments, there is a belief that many computer scientists deliberately withhold information from their peers (Frohlich, 1998 as cited in Stodden, 2010).

There are a number of reasons why scientists might either deliberately or strategically withhold data; the perceived lack of opportunities for advancement or other reward, an unwillingness to invest the level of effort required to provide complete documentation, concerns over control of resources and results, an intention to publish as little information as possible so as prevent giving other researchers a competitive advantage or due to issues with intellectual property rights (Borgman, 2007; Frohlich, 1998;

both as cited in Stodden, 2010). Scientific articles replete with omissions significantly reduce utility and diminish the reputation and rewards an individual may see from engaging in the act of further research in that domain (Stodden, 2010). Whatever the reason, whether deliberate or simply resulting from a need for brevity due to limited publication space, in the absence of complete knowledge and data for any of these SDG experiments it would be impossible for this research to faithfully recreate their entire project.

2.8 Summary

This chapter has reviewed literature to understand synthetic data generation, identifying that while there is a large collection of literature that presents individual SDG methods, there has been very little literature that has presented a collective knowledge of SDG, including its uses and methods. It is also resolved that no literature could be located that has discussed the realism aspect that many projects claim.

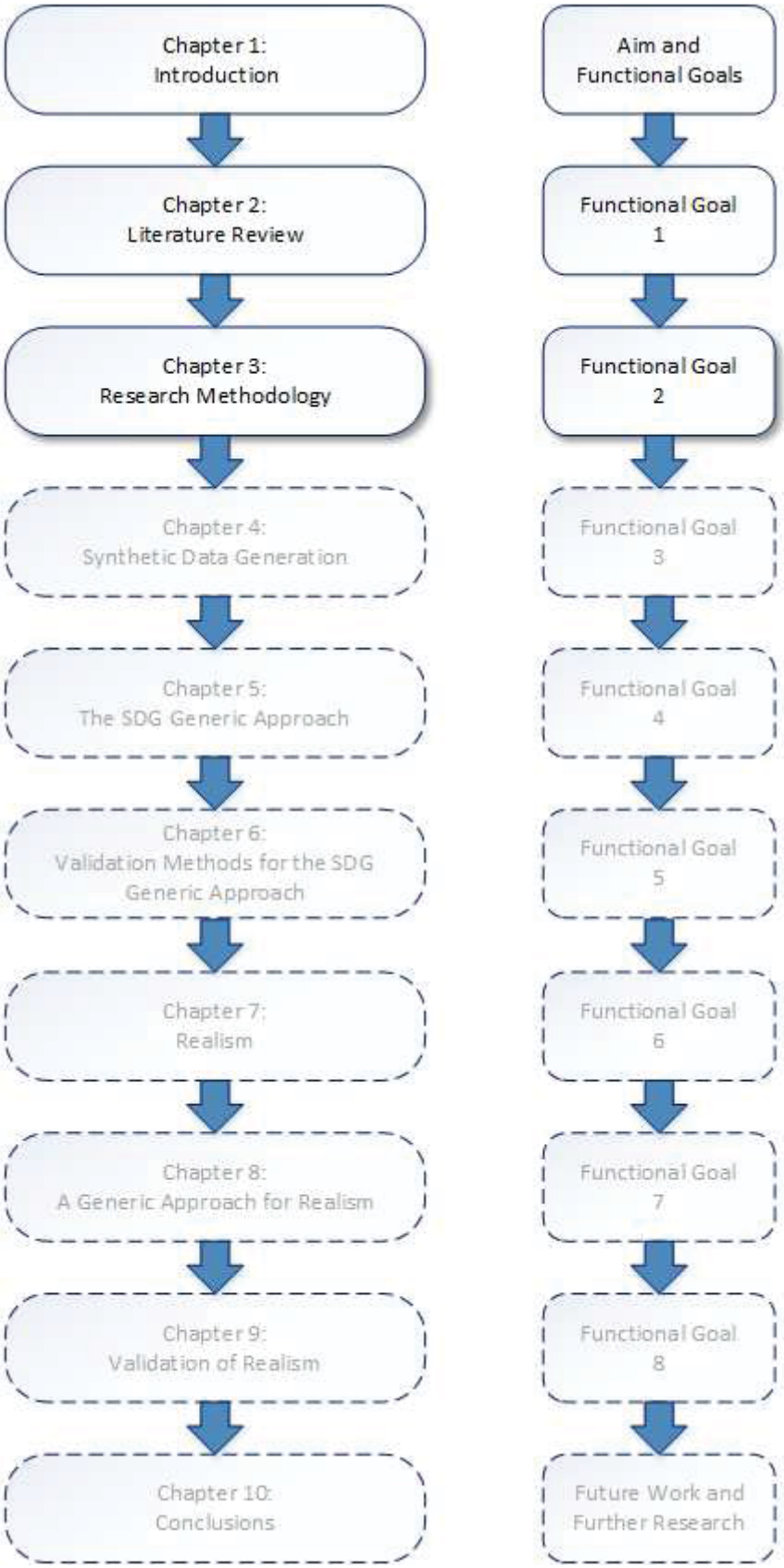
We have seen that SDG shares a strong resemblance to computational modelling, and that the two share many of the same methodological approaches. While CM has a strong history of validating the modelled, or synthetic data, SDG literature demonstrates a lack of similar scientific rigidity. A number of developed CM validation techniques were presented and explained; each of these will be analysed and applied to an SDG case study in the following chapter.

The penultimate section of this chapter discussed the incomplete nature of many published SDG articles and how being incomplete prevents the repeatability which the scientific method itself mandates for independent claim validation. Incomplete presentation of the approach, method and validation can be intended or accidental. The inability to test another's claims weakens the strength and faith readers have in scientific journals because without repeatability there can be no assessment of the strength and accuracy of the methods presented and results claimed. Readers are left unable to assess whether the proposed method is applicable to their research domain and unable to assess whether the approach used by the authors truly works. A journal article that does not document completely is of limited practical use, reduces the reputation of that scientific domain and inhibits the knowledge-sharing nature on which the scientific method was founded. Unrepeatable claims of success made in incomplete scientific publications actively limit the processes and incentives that help to motivate others to conduct further research in a given domain, delaying the development of new knowledge in that area for all.

This page intentionally left blank

“If we knew what it was we were doing, it would not be called research, would it?”

(Albert Einstein)


Chapter 1:
Introduction
Chapter 2:
Literature Review
Chapter 3:
Research Methodology
Chapter 4:
Synthetic Data Generation
Chapter 5:
The SDG Generic Approach
Chapter 6:
Validation Methods for the SDG
Generic Approach
Chapter 7:
Realism
Chapter 8:
A Generic Approach for Realism
Chapter 9:
Validation of Realism
Chapter 10:
Conclusions
Aim and
Functional Goals
Functional Goal
1
Functional Goal
2
Functional Goal
3
Functional Goal
4
Functional Goal
5
Functional Goal
6
Functional Goal
7
Functional Goal
8
Future Work and
Further Research

3. Research Methodology

This chapter describes the methodology used to achieve each of the functional goals identified for this research at section 1.7.

The research in this chapter was carried out in order to achieve functional goal 2:

Functional Goal 2. Develop classification scales to describe synthetic data and generation methods. Analyse and categorise SDG literature using these scales. Describe the methodologies to be utilised to resolve the research aims and remaining functional goals.

This chapter is structured as follows:

3.1 Method for Identifying SDG Literature: Functional Goal 2

A journal search was conducted using Massey University's Electronic Library search engine. The keywords used were chosen to provide the widest selection of articles that discussed SDG. As shown in Figure 1, the query search terms were: "*synthetic*" + "*data*" + "*generation*" + ("*realism*" or "*realistic*").

The articles collected were then evaluated for use across two streams, or scopes, of SDG research. The primary scope seeks to understand the history and development of data synthesis, while the second is intended to identify a broad range of SDG projects that propose or identify *realism* as a necessary component, evaluating the methods they employed to generate their *realistic* datasets. A selection of articles that were to be used during this research is shown in Appendix A. This evaluation will assess the context with which each project defined *realism*, the method and/or algorithms used to produce the synthetic data, the approaches used to assess or ensure realism, the *syntheticness* of the resulting dataset and finally how and to what degree the authors evaluated their success or failure at

producing *realism* as defined by their project. To keep the number of responses reasonable, the search results were constrained to those articles published during or after the year 2000.

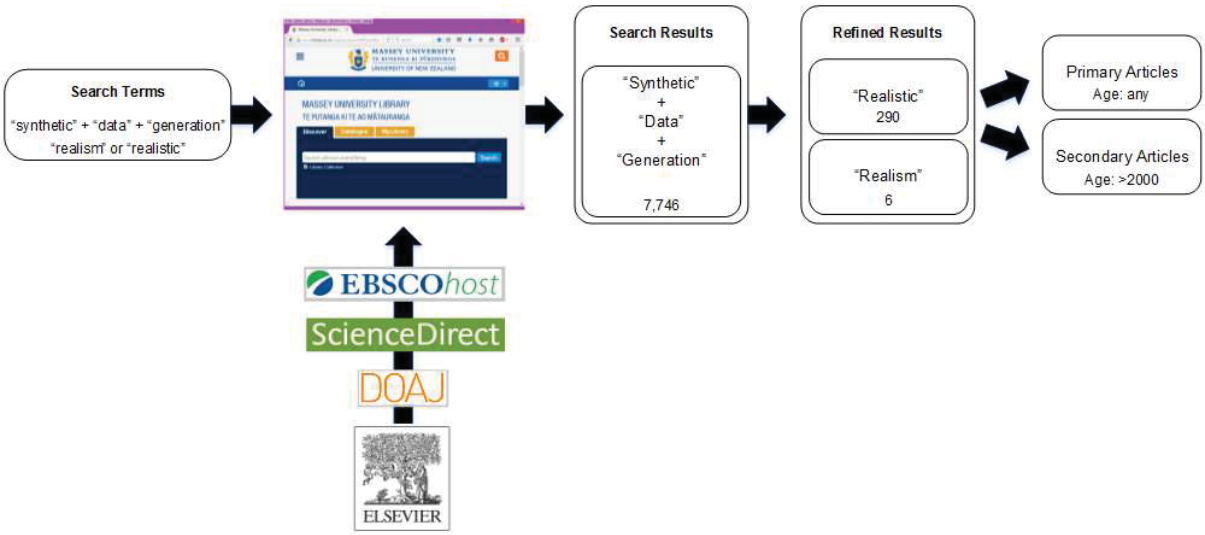


Figure 2: SDG Literature Search and Categorisation

The collection of SDG projects was narrowed to those demonstrating a number of elements. The first being an intention towards realism. This intention needed to be generally mentioned in some way within the abstract or introduction. Those where it wasn't directly stated remained if the clear inference could be drawn from the generation process or solution. The second element necessary for inclusion was discussion of the problem being simulated along with clear identification of the generation algorithm, method and process. The final key element was full source disclosure for the real records, observational data or statistics that formed the seed data.

Regard was given to any discussion of validation processes employed by the authors. Some examples found included a variety of comparisons to values or structures evident in the source or real data.

Given that a number of authors credit the genesis of fully synthetic data generation to an article published in 1993 by Rubin (Drechsler, 2011; Drechsler, 2014; Kuiper et al, 2015; Sakshaug, 2011), and partially synthetic data generation approaches to an article in the same year and journal authored by Little (Drechsler, 2014; Reiter & Kinney, 2012), the *after year 2000* constraint was removed on a second search for articles to be used in the primary SDG research scope. This ensured the widest range of material was available for the historical review and theoretical discussions, and that the collection included Rubin's allegedly inceptive paper.

Accordingly; the first scope provides an understanding of how synthetic data generation evolved, guiding understanding of how realism came to be a key focus for those researchers and engineers who create and use synthetic data in the real world. The second provides scope data from the widest possible range of fields of study, supporting assessments and validation of the contextual basis for understanding, definition and eventual formalisation of the concept of *realism*.

Together, the two streams provide a solid foundation from which the development of a taxonomy and generalised systematic method for the production and assessment of *realism* in synthetic data can be achieved.

3.2 Method for Characterising Synthetic Data: Functional Goal 3

Achieving this functional goal requires a combination of primary and secondary qualitative research. The primary research consists of systematic historical research into SDG, paying particular attention to any definitions or descriptions given that contextualise SDG and how these may or may not apply within the domain of the present research. The secondary research involves comparative analysis of the common approaches to SDG in order to build a classification of the characterisations of synthetic data.

3.3 Method for Identifying the Generic Approach to SDG: Functional Goal 4

Identifying a generic approach to SDG requires secondary research of a wide range of SDG articles in order to populate a matrix. This matrix conforms to that of a Formal Concept Analysis context table as discussed in section 8.6, listing each SDG article on the *x* axis and a generic title for each methodological component on the *y* axis, with binary identification of relationships between the two. Those components common to all approaches will be identified and their application as a methodology described. The concluding component is evaluation and discussion of potential strengths and weaknesses that may be inherent to the identified generic approach.

3.4 Method for Applying Existing Validation Methods to SDG: Functional Goal 5

In order to propose suitable validation methodologies for the generic approach to SDG an historical research methodology was required to identify current and prior validation methods that may have been used in the domain of SDG, along with comparative and descriptive methodologies to express and relate approaches drawn from the comparative domain of Computational Modelling. Through the application of a case study each validation method will be described and later contrasted with the proposed validation model for the generic SDG approach.

3.5 Method for Defining Realism in SDG: Functional Goal 6

Utilising a similar approach to that of goals 3 and 5 this goal utilises an historic methodology to identify those works that make claim to realism in their SDG approach, collecting and integrating any descriptions or definitions located with primary research looking at the concept of realism.

3.6 Method for Characterising Realism in SDG: Functional Goal 7

Identifying the characteristics of realism necessary in a synthetic dataset so that it may be a suitable replacement for real data necessitated exploration into the domain known as Knowledge Discovery in Databases (KDD). The overall characterisation method devised uses a case study approach to demonstrate the application of KDD using established data mining methods such as Formal Concept Analysis, Concept Lattices, Classification and Characteristic Rule Mining and Concept Hierarchies imbued with statistics.

3.7 Method for Defining Validation of Realism in SDG: Functional Goal 8

Describe the approach of using the output of each component of the enhanced and expanded HCI-KDD model as the basis for a validation solution. The approach continues and extends the case study in chapter 8 demonstrating how each applies to validation of the synthetic dataset.

3.8 Case Study Methodology

The *case study* is an ideal methodology for inquiry that comes with a well-developed history and documented robust qualitative procedures for investigation and process validation (Eisenhardt, 1989; Tellis, 1997). Authors across many fields have detailed and developed the case study as a grounded method of comparative research (Eisenhardt, 1989) and considered it to be a method that spans the project lifecycle, providing a *real-life* perspective on observed interactions (Cockburn, 2003). Case study methodologies are frequently used in the information sciences (Lee, 1989) as a way of conducting and presenting research into subject areas that include qualitative and mixed-mode information science inquiry (Cable, 1994), assessment of the practice and consequences observed implementing different geographical information systems (Robey & Sahay, 1996), information retrieval (Fidel et al, 2004; Smithson, 1994), investigation of management information systems (Lee, 1989) and even in demonstrating an information technology alignment planning process (Peak, Guynes & Kroon, 2005). Case studies are considered as well developed and tested as any other scientific method, can be used to generate and test theory. Case studies are a valid modality where the rigid approach of experimental research cannot or does not apply (Eisenhardt, 1989; Tellis, 1997; Yin, 2011). For these reasons this thesis adopts the case study methodology over the more usual experimental approach seen in computer science research.

One criticism of the case study methodology is that it does not present with a fixed reporting format due to the unique nature of each use case (Trellis, 1997). Another is that the case study approach is not able to systematically handle data (Trellis, 1997). A number of case study utilisation methods are formally described, including; exploratory, explanatory, descriptive, intrinsic, instrumental and collective types (Pickard, 2013; Trellis, 1997; Yin, 2011; Zainal, 2007). Exploratory case studies explore and examine phenomena in collected data (Zainal, 2007). The explanatory case study approach presents an example and through both high-level and deep-level exploration of the data, highlighting differences between the model described and the example; inferring and explaining relationships in those differences (Yin, 2011; Zainal, 2007). The descriptive case study is one which begins by recounting a theory, articulation of what is known about phenomena in the data and followed by assessment and evaluation (Zainal, 2007). The descriptive case study is one which may present as a narrative (McDonough & McDonough, 1997) such as was seen in the example of journalistic description of the Watergate scandal by two journalists (Yin, 1984). The intrinsic case study is one which is documented to primarily provide a better understanding of the case, the instrumental case study provides more in-depth examination of a particular phenomenon or theory and finally the collective case study is one which uses multiple instrumental cases to explore elements of the problem being researched.

Proper respect for the scientific method requires repeatability of any experiment. In the case of each SDG method reviewed and included in the appendices there are elements missing or not discussed in the article that would be necessary to effect replication of the author's research, whether it be access to an equitable dataset to that used, or in many cases, some evidence of the method used to identify and define elements and attributes in the real dataset that were considered necessary to the SDG output. Finally, discussion and documentation that demonstrates validation of how the authors came to the realisation that their method *worked* was absent. In the absence of an ability to accurately replicate and represent one of these published studies, it would be difficult to then directly operate the realism extraction and validation methods discussed in the present work and draw accurate and unbiased conclusions. In any event, such experimentation is outside the scope of this thesis and would be better served in a future research project where a new SDG problem could be defined and conducted in a manner consistent with existing established practice, that is, sans realism validation, and then repeated with the approaches presented in this thesis. The case study approach sufficiently allows for demonstration of how the theory and approaches presented apply to a real-world SDG case without the potential of misrepresenting any case or presenting potentially invalid experimental results. This thesis relies on a combination of descriptive and instrumental case study approaches to analyse and apply contributions.

The case studies in this thesis utilise the published CoMSER SDG model of *McLachlan, Dube and Gallagher* (2016) for illustration, contrast and validation of the theories and concepts presented. While the CoMSER example is limited in its single SDG method, it is at least one example where all the motivations, input data, methodology, code and generated records are available to this author.

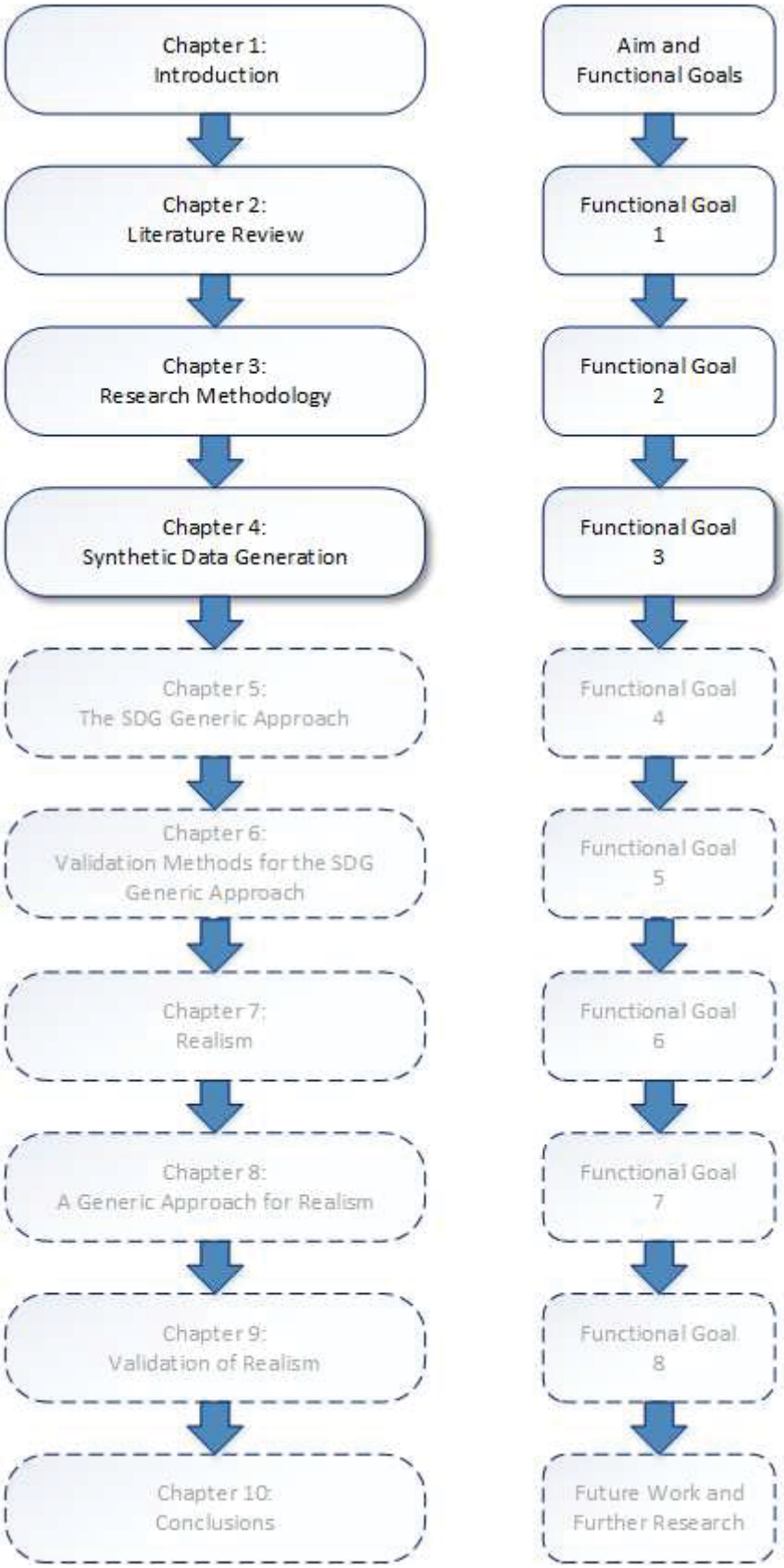
The CoMSER application uses clinical practice guidelines and specialist clinical knowledge to prepare CareMaps. A CareMap is a unidirectional clinical pathway of the progression of a health condition and treatment steps that a patient may undergo. The CareMap is imbued with statistics that are used by CoMSER to establish probabilities for each synthetic patient who enters the pathway. The output CoMSER generates are *Realistic Synthetic Electronic Health Records* (RS-EHR); synthetic patient records describing the labour and birth event that are statistically consistent to the incidence and treatment of real patients at one large tertiary hospital birthing unit in Auckland, New Zealand.

3.9 Summary

This chapter began by discussing how literature was located for the primary research that was to be conducted, and how this search was extended to draw additional articles to provide the foundation and history necessary for complex review and discussion in the literature review. The chapter went on to address the methodology to be used to resolve research goals in each of the proceeding chapters. It also concluded that the more usual experimental approach in computer science research could be better replaced in this instance by the established case study methodology. Finally, it explained how case studies utilising the CoMSER model that performed SDG in the domain of Midwifery would be applied to each theoretical or practical example throughout this thesis.

“At the heart of science is an essential balance between two seemingly contradictory attitudes – an openness to new ideas, no matter how bizarre or counterintuitive they may be, and the most ruthless sceptical scrutiny of all ideas, old and new. This is how deep truths are winnowed from deep nonsense.”

(Carl Sagan, The Demon-Haunted World, p304. 1997)


Chapter 1:
Introduction
Chapter 2:
Literature Review
Chapter 3:
Research Methodology
Chapter 4:
Synthetic Data Generation
Chapter 5:
The SDG Generic Approach
Chapter 6:
Validation Methods for the SDG
Generic Approach
Chapter 7:
Realism
Chapter 8:
A Generic Approach for Realism
Chapter 9:
Validation of Realism
Chapter 10:
Conclusions
Aim and
Functional Goals
Functional Goal
1
Functional Goal
2
Functional Goal
3
Functional Goal
4
Functional Goal
5
Functional Goal
6
Functional Goal
7
Functional Goal
8
Future Work and
Further Research

4. Synthetic Data Generation

This chapter seeks to identify the generic characterisation for synthetic data. It begins to develop characteristics by reviewing the work of Rubin and others who have sought to define key concepts in the synthetic data generation domain. It continues by examining the approaches and methods used before arriving at a classification matrix that structurally describes the character of each class of synthetic data that SDG methods are creating.

The research in this chapter was carried out in order to achieve functional goal 3:

Functional Goal 3. Characterise SDG Methods and Synthetic Data

This chapter is structured as follows:

4.1 Introduction

In order to characterise a concept, we must first look at how other researchers have attempted to define and use it within the scope of their own work (Dey, 2001). This section will review the history and applications found for synthetic data, along with the methods used in its creation. From this collection of reference material, a characterisation and classification matrix will be developed that can be used to understand the level of synthesis employed measured by the potential risk for exposure of the real data that we seek to synthesise.

4.2 Background

It was claimed that SDG resulted from the need to disseminate data safely (Drechsler & Reiter, 2011; Rubin, 1993), however many early SDG methods were not actually focused on the protection of sensitive information. These initial methods spanned a number of fields including the environmental, material and chemical sciences, generating and using synthetic data in testing hypotheses and early software applications, predictive modelling, or for simulations (Geweke & Porter-Hudak, 1983; Leggett & McBryde, 1975; Stedinger & Taylor, 1982; Van Holde & Weischet, 1978). It is more recent projects, at least those since the mid-to-late 1980's, that have sought to isolate and replace personally identifiable

data, with the concomitant goal of maintaining the integrity of data that an organisation may wish, or be required, to publish (Drechsler et al, 2011).

Over the last ten years an increasing number of authors have credited the concept of *fully synthetic data* to the commentary contained within an article published in 1993 by statistician Dr. Donald B. Rubin (Drechsler, 2011; Drechsler, 2014; Kuiper et al, 2015; Sakshaug, 2011). Also, those authors usually credit *partially synthetic data* to an article by Little (1993) published contemporaneously in the same journal (Drechsler, 2014; Reiter & Kinney, 2012;). In the 1993 article Rubin introduces the idea that the information in datasets can contain confidential elements that cannot be made public (Rubin, 1993). He proposes using the variability and characteristics inherent to the confidential dataset to seed the creation of synthetic data using his particular multiply-imputed generation method (Rubin, 1993). In other papers he discussed using a similar approach to infer and populate the value of empty fields within a dataset (Rubin, 1976; Rubin, 1987; Rubin, 1996). While the idea of using real data to create synthetic datasets may have appeared novel in his chosen field of statistics, it was established and commonplace to many other fields of science.

4.2.1 The attachment of pre-eminence in *Fully Synthetic Data* to Rubin

Prior to Rubin's 1993 paper the field of statistics, and particularly his focus area of population statistics, had primarily used the term *synthetic data* to describe the larger dataset that results from the merger of two or more smaller datasets (Alter, 1974; Wolff, 1980). Rubin's (1993) raises a number of primary and secondary points that are listed in Table 2. Rubin's 1993 work became well known within his field and is regularly cited as the first work to describe fully synthetic data. This was started by one of his early Ph.D. students, Jerome Reiter, who later propagated the claim with students he went on to supervise.

The first references we see to Rubin's 1993 work simply report it as an alternative approach to releasing public use data, and focus on the suggested method rather than any claim of pre-eminence (Raghunathan, Reiter and Rubin, 2003; Reiter, 2004). Rubin himself does not appear to engender the claim that he fathered fully synthetic data, rather, he focuses attention on his multiple imputation method (Raghunathan, Reiter and Rubin, 2003). This has also been the position of many other authors, most of whom more correctly focused on Rubin's invention of the method of multiple imputation in the generation process rather than shorthand claims of his creating the overall concept of *synthetic data* (Abowd & Woodcock, 2001; Alfons et al, 2010; Klein & Sinha, 2015; Mateo-Sanz, Marinez-Balleste & Domingo-Ferrer, 2004; Yuan, 2005;).

The first time we see text eluding to Rubin being first to describe the overall concept of *synthetic data* comes in a paper published by Reiter in 2004. Under the heading *Synthetic Data* he writes:

> *"If data disseminators are not willing or not allowed to release genuine microdata, another approach is to release synthetic, or simulated, microdata that look like the genuine data. This was first proposed by Rubin (1993)."* (Reiter, 2004a)

The next reference continues to develop the claim of Rubin's pre-eminence in the dissertation of Christine Kohnen, a Duke University Ph.D. student supervised by Reiter in 2005 (Kohnen, 2005). At page 9 of her dissertation she writes at the start of a new paragraph:

> *"Synthetic data methods were first proposed by Rubin (1993) as a means of creating public-use data sets that honor the confidentiality of respondent values while maintaining simplicity in their analysis."* (Kohnen, 2005)

The impression drawn from these claims has resulted in its expression in a more public forum, Wikipedia. On the 6th of December 2009 the generic Synthetic Data Wikipedia page was edited to report that "*the history of the generation of synthetic data dates back to 1993*" (Wikipedia, 2009a) without supporting citation. Two days later another edit was made to extend this formative statement, providing that the original idea for *fully synthetic data* came from Rubin and citing the 1993 paper (Wikipedia, 2009b). This overlooks approaches already established in the common literature before the articles of either Rubin or Little were published, ignoring entirely those that had proposed approaches resulting in fully or partially synthetic data.

4.2.2 Extending the History of Synthetic Data

The earliest reference to synthetic data located during this literature search was in a 1971 article that describes the creation of tables of synthetic data for use in testing and modifying marketing data (Green & Rao, 1971). We see methods that pre-date Rubin and Little for the creation of fully synthetic data based on observed statistics (Birkin & Clarke, 1988; Stedinger & Taylor, 1982), used to predict and test observational outcomes (Geweke et al, 1983) and generation driven by probability models for use in simulations (Graham et al, 1988) and forecasting (Delleur & Kavvas, 1978).

Birkin and Clarke's 1988 paper described an approach to generating synthetic populations based on modelling from statistical reference data such as census tables (Birkin & Clarke, 1988). While Rubin's 1993 article may appear to some to go further in the preamble in describing why we would need or use SDG, Birkin and Clarke's paper appears to be a good place to begin refuting the assertion that Rubin was first to describe synthetic data. Shown in Table 3, comparison of the problem and reasoning of both papers demonstrates significant similarity, in that both discuss issues that arise when we need to disseminate data. Both expound on the absence of industry and occupation details in published census data, presenting methods aimed at populating these fields from alternate sources. Birkin and Clarke's method, published almost six years before Rubin's 1993 article, uses Iterative Proportional Fitting (IPF) (Birkin & Clarke, 1988). The IPF method itself has been around in literature since 1940 (Deming & Stephen, 1940) and has been widely researched and validated (Pukelsheim & Simeone, 2009). Rubin's approach concerns a novel type of SDG, that being his own *multiple imputation method* which is used both to generate the synthetic data (Rubin, 1993) as well as to infer plausible replacements for the

missing values in empty or sanitised data fields (Rubin, 1987; Yuan, 2010). Rubin acknowledges that his method can result in a larger number of errors when contrasted to actual data (Rubin, 1993). These errors reflect the level of uncertainty that results from incomplete data with missing values (Yuan, 2010). When we consider that the IPF method is more rigorously justified through decades of effort expended by numerous research groups to improve and validate the model, we can understand why Birkin & Clarke's (1988) method claims a higher percentage of accuracy when compared to Rubin's.

Given that there were models with greater accuracy available in the years prior to Rubin's (1993) article and that Rubin's paper is more correctly reported by others as being the *first model of a multiple imputation method* for fully synthetic data, it is difficult to understand why so many authors continue to misreport the contention that Rubin was the first person to describe *synthetic data*, fully synthetic or otherwise.

Table 2: Comparison of Rubin (1993) to Birkin & Clark (1987)

	Rubin (1993)	Birkin and Clarke (1987)
(a) Asserts that there is an increased demand to release data, especially publically funded population-based data.	Yes	Yes
(b) Discusses confidentiality as an issue	Yes	Speaks of the barriers Government have placed on the dissemination of data.
(c) Discusses that methods should be put into place to measure and mitigate risks associated with release of data.	Yes	
(d) Claims a method that solves the issue by delivering synthetic data that contains no real individual's identifiable information.	Yes – Rubin's own multiple imputation method.	Yes – using the established methods of IPF and contingency table analysis.
(e) Discusses that the method must in some way be expedient and efficient	Yes	Yes
(f) Discusses that information may be lost when creating the synthetic data	Yes	Discusses requirement for future work to validate output.
(g) Uses census data as an example discussion point and dataset	Yes	Yes

Over time, the use of synthetic data generation has spread into almost every field of study and the range of methods and solutions appears inexhaustible. Table 3 lists articles that propose SDG projects which possess all of the elements described in section 3.1 as being necessary for inclusion in the second scope of this research.

4.3 Approaches and Methods for Synthetic Data Generation

A variety of SDG methods are used across a wide range of domains. In the literature, **five categories** of synthetic generation methods are evident. In the **first category**, there are *data masking* models that

replace personally identifiable data fields with synthetic data (Mouza et al, 2010; Domingo-Ferrer, 2012; Winkler, 2004). In the **second category**, embedding synthetic target data into recorded user data in a method known as *Signal and Noise* (Barse et al, 2003; Killourhy et al, 2007; Whiting et al, 2008). In the **third category**, one finds *Network Generation* approaches that deliver relational or structured data (Ascoli et al, 2001; Tsvetovat & Carley, 2005; Van den Bulcke, 2006). In the **fourth category**, one finds truly random data generation approaches like the *Music Box Model* (Mwogi, Blondich & Grannis, 2014). In the **fifth category**, one finds the probability weighted random generation models like the *Monte Carlo* (Houkjaer et al, 2006) *Markov chain* (Mwogi et al, 2014) and *Walkers Alias* methods (McLachlan et al, 2016).

This section describes a number of the synthetic generation methods frequently seen in the literature. Table 4 then follows providing a summary for each method and an example of one cited application from those reviewed and listed in the appendices.

4.3.1 Data Masking

Data Masking is also referred to as data scrambling, data binding, data anonymization, data sanitization or data encoding. Data masking obfuscates (Radhakrishnan et al, 2005) or replaces the sensitive data (Mouza et al, 2010). Typical established methods of data masking include;

- **Randomisation:** where real data is replaced with randomly generated data that may be governed by rules to limit its scope to fall with a given range, variance or percentage of the original value.
- **Blocking:** Also known as substitution. This is where the original data is replaced entirely or partially with an artificial record, usually from a lookup table.
- **Masking:** Where the original data is fully or partially replaced with a masking character, such as the asterisk that is often used to replace digits in a credit card number or password.
- **Scrambling:** The data type and size is preserved, however the value is entirely replaced.
- **Shuffling:** The substitution data is derived entirely from the value in the column itself. The data in the column is randomly moved or shifted between rows.

(Adapted from: www.etl-tools.inf)

4.3.2 Signal and Noise

The signal and noise method involves the collection or creation of a large dataset of generally normal *noise* data, that is, data that would be seen in the target system when the issue, or signal, is not evident. This might consist of several days' worth of normal traffic seen on an interface that connects to a web server, application service or console (Killourhy et al, 2007), network or system firewall (Sperotto et al, 2009), or collected from a database engine such as that which might contain or process messages such as emails (McHugh, 2000). This normal traffic may be used as it was captured, or may constitute the seed data pool for some form of randomised or constrained generation method.

The *signal* is that element that the method is mostly concerned with. It is that which is sought out in the eventual synthetic dataset, often being used to train other applications, systems or analysts (Whiting et al, 2008). The *signal* may be artificially or manually created by the researchers (Whiting et al, 2008), or may be drawn from a set of data such as breach data that has been captured in a manner similar to that of the *noise* data (Sperotto et al, 2009).

4.3.3 Network Generation

The Merriam-Webster dictionary defines a network as a group or system of intersecting or connected people or things (MerriamWebster, nd). Network generation is concerned with the creation of a dataset that describes a network of objects, be it a social network of people similar to that which might be seen in a terrorist organisation (Tsevotat et al, 2005), or a variety of network types generated in medicine, such as that for gene expression (Bulcke et al, 2006) and the study of nerve tissue structure (Ascoli et al, 2001). The network generation dataset generally describes the nodes of the network, as well as the paths that interconnect or intersect each node.

4.3.4 Music Box Method

Mwogi et al (2014) describe the music box method as one where they de-identified real source Health Level 7 (HL7) health record data, broke the complete source data down into its respective HL7 components and used these components to generate new random event records based on the pre-calculated collecting together of random HL7 segments that match the event type. They claim such a method is similar to the random appearance of a music box plucking the teeth on the song drum (Mwogi et al, 2014).

4.3.5 Markov Chain Method

The Markov Chain method is a process by which each component of a record is generated by a random process constrained or dependant only on the value of the current state and the conditional state probability of the next step (Markov, 1971). Markov Chain models use probabilities, in this case, the probability that the system will transition from the current state to any one of a number of random next-states (Mwogi et al, 2014). Mowgi et al (2014) analysed existing HL7 records, building a dataset of HL7 record segments and the probabilities that a given segment would be followed by another given segment. This was then used to generate synthetic records where each had a random HL7 segment starting point, and probabilities were used to randomly select each HL7 segment selected until the new synthetic record was complete (Mwogi et al, 2014).

4.3.6 Monte Carlo Method

A Monte Carlo method is any method that solves the problem of generating suitable random numbers and observing a fraction of the numbers obeying some property or properties (Hoffman, 1998). The Monte Carlo method for synthetic data generation is one of probabilistic nature, in that it generates data with similar probability properties as those which have been observed in real data (Manno, 1999).

4.3.7 Walker's Alias Method

Walker's Alias method is an efficient two-step pseudo-random approach that uses defined frequency distributions with a finite number of outcomes to generate synthetic data (Walker, 1974; Walker, 1977; Davis, 1993). Over thousands or tens-of-thousands of records the alias method has been found to produce synthetic data with a high degree of accuracy when compared to statistical distributions taken from observed data (Ahrens et al, 1989; McLachlan et al, 2016).

4.3.8 Distribution of Methods and Domains in SDG

Those SDG methods collectively described as using probabilistic weighted models were the most common observed during the course of this study, including; Walkers Alias, Markov and Monte Carlo. The second most common observed were a variety of data masking approaches. Figure 3 shows the distribution of common SDG model types as seen in the reviewed literature listed in appendices A and B.

Figure 4 shows the distribution of domains using SDG in the same collection of papers. This data was drawn by referring to the domain of the lead author, cross referenced to the purpose or domain the data was intended for. Computer sciences were observed most frequently, followed by the energy and environmental sciences that in many cases could have been collected together as they were most often concerned with modelling of wind, solar or ocean currents either for the prediction of weather forecasts (environmental science) or renewable energy outputs (energy science).

Table 3: Characterisation of Synthetic Data Generation Methods

Generation Model	Example	Example Context
Data Masking	Mouza et al, 2010	Rule based identification, labelling and replacement of sensitive data.
Signal and Noise	Whiting et al, 2008	The collection of a large dataset of general and random traffic (the *noise*) and the creation of a threat that analysts are being trained to find (the *signal*). The eventual dataset is created by interweaving the *signal* within a synthetically generated dataset of randomly selected *noise*.
Network Generation	Ascoli et al, 2001	A method that is able to generate neuronal dendrite structures using lessons learned from a limited sample of real dendrite structural data
Random Models (Music Box)	Mwogi et al, 2014	Generation of complete synthetic records using randomly selected events and randomly selected HL7 segments linked to that event.
Probability Weighted Random Models (Monte Carlo, Markov Chain and Walkers Alias)	Mwogi et al, 2014; Houkjaer, 2006; McLachlan et al, 2016	A Monte Carlo method is used to populate a range of fields constrained by the definition of a database structure. Walkers Alias method is used to generate paths based on frequency distribution, node to node, through a health-based state transition machine that are consistent with the frequency seen in the treatment statistics of a group of 8100 real patients.

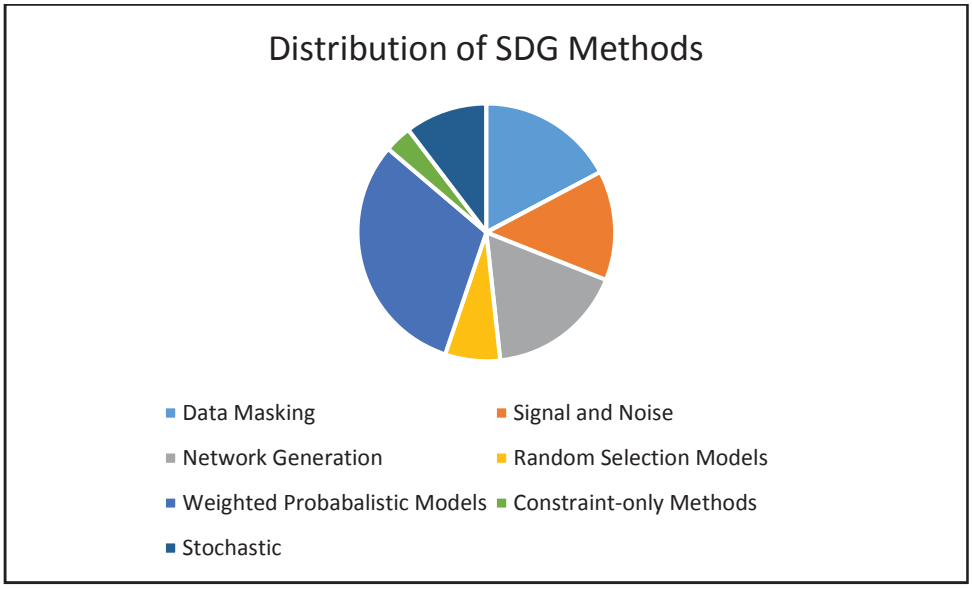


Figure 3: Distribution of SDG Methods

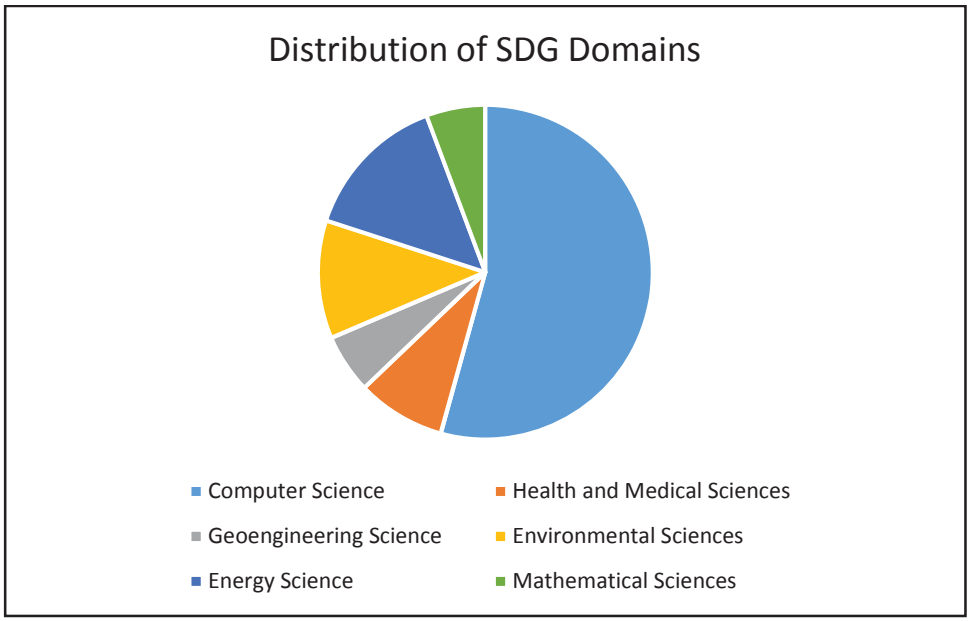


Figure 4: Distribution of SDG Domains

4.4 Differentiation for Classification

The McGraw-Hill Dictionary of Scientific and Technical Terms defines synthetic data as "*any production data applicable to a given situation that are not obtained by direct measurement*" (Parker, 2003). Synthetic data are intended to be representative of real data (Barse et al, 2003). Machanavajjhala et al (2008) describes synthetic data generation as a data anonymisation technique. A technique used for

ensuring privacy when the use of real-world sensitive data is required (Machanavajjhala et al, 2008). It is clear from the literature however, that defining synthetic datasets as merely a type of anonymised data does not account for the scope we see in how synthetic data is both created and used, nor does such a definition allow for data which can truly be described as synthetic, that is, data which was not seeded by real-world or sensitive seed data.

This paper proposes a set of qualifying definitions that would allow us to correctly identify and classify the type of synthetic data that a given method has produced. The ability to classify the synthetic data will assist in the validation processes to be discussed later in this thesis.

The classification types and their definitions are explained in Table 4:

Table 4: Classification of Synthetic Data

True Synthetic Data	Data which has been generated where no confidential or sensitive data has been directly used. Exemplar generation may rely on algorithms that use models or frameworks to populate a dataset with generic seed data based on statistical probability or acute randomness. An example of True Synthetic Data can be seen in CoMSER (McLachlan, Dube & Gallagher, 2016)
Fully Synthetic Data	Synthetic Data is that where no real-world data is contained within the output dataset. Some Synthetic Data approaches still use real or aggregate data in the input phase, however none of the real data is maintained across the generation method. Common methods for ensuring that the data is synthetic involve capturing and breaking up real-world data into much smaller components, rebuilding these components into new rows of data. Another uses the real data to construct a database architecture, populating the new database architecture with synthetic data based on the observed real data (Houkjaer et al, 2006).
Partially Synthetic Data	The partially synthetic dataset consists of some form of simulated or synthetic data intermixed or aggregated with unaltered real data. An example would be the Outbreak-Detection system which used simulated signals injected or superimposed on real background noise (Cassa, Olsen & Mandl, 2004).
Anonymised-Only Data	Projects which operate to identify and replace or scramble sensitive fields within a dataset, leaving the rest of the dataset largely unchanged.
Real Data	Real or observed data in which no attempt has been made to anonymise, conceal or synthesise the values of any sensitive or confidential fields.

4.5 The ATEN Framework

In scientific papers it is not uncommon to see methodologies that possess either multiple separate, combined or sequential components presented as a framework (Green, Caracelli & Graham, 1989). In keeping with this approach, this thesis presents the ATEN framework shown in Figure 5, a synthesis of three components named THOTH, RA and HORUS, that when used collectively breathe realism into synthetic data.

In the thirteenth century BC the Egyptian king Amenhotep IV changed his name to Akhenaten (Freed, D'Auria & Markowitz, 1999; Gore, 2001). He did this in reverence to the syncretized deity Aten whom he worshiped daily, and who he intended to elevate as the single central god of all Egyptian religion (Gore, 2001). Aten, whose name was a shortened version of *Ra-Horakhty*, or *Ra-Horus-Aten* (meaning *Ra, who is Horus of the two horizons*) was the synthesis of as many as a dozen gods from Egyptian mythology and was often represented as a stream of rays, the rays of light that breathe life, emanating from the sun's disk (Gunn, 1923; Wilkinson, 2008).

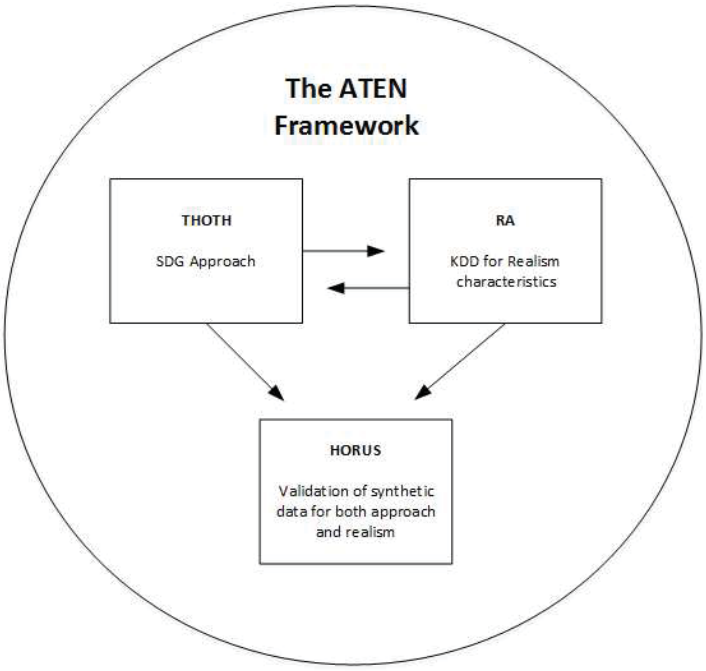


Figure 5: The ATEN Framework

The questions that each component of the ATEN framework seeks to resolve are:

THOTH	How can we generate synthetic data?
RA	What knowledge is necessary to achieve realism in generated synthetic data?
HORUS	If THOTH operates using RA's knowledge, can we say realism was achieved in the resulting synthetic data?

There are a number of ways that the ATEN framework exceeds the methodologies followed and documented in the articles listed in the appendices. Firstly, it is a complete SDG lifecycle framework that considers every element before, during and after the data generation procedure has occurred. Secondly, it encourages and necessitates a more complete level of documentation than the majority of current literature presents. The third improvement results from the first two, in that when applied completely during an SDG project the ATEN framework provides the necessary knowledge and verification to support claims of success and enable repeatability that is fundamental to the scientific method. It is likely most researchers are following sound procedures that conform to the ATEN

framework and that these are simply not being reflected in their publications. It is also possible that many are not considering much of the framework's elements and that this second group of researchers may see the additional knowledge discovery and documentation processes as an onerous and time consuming delay to their generation projects. Some of the knowledge discovery processes can be automated, however many will be dependent on the type of data being analysed, the generation method, the synthetic data that is sought and the use to which that data will be applied. In any event, a little extra time expended in service of greater accuracy and in support of claims of success is time that no researcher should consider wasted.

4.6 Case Study: Introduction

The CoMSER method is an example of SDG that follows the generic approach discussed later in Chapter 5. CoMSER is an SDG approach that implements the described elements of another more global method known as GRiSER, as shown in Figure 6. CoMSER identifies the need for synthetic data as a way of providing Electronic Health Records for secondary uses without exposing real patient data (McLachlan et al, 2016). The knowledge gathering phase collects incidence and treatment statistics, clinical practice guidelines and a clinical vocabulary which is grounded through input from expert clinicians. This knowledge becomes the input used to develop and populate a clinical CareMap which provides a clinician-readable context for a state transition machine (STM). The CareMap is reviewed and where necessary and in consultation with the clinician is amended in the first part of the process shown in the UML diagram at Figure 7. The clinician also provides context-specific patient notes for each stage of the CareMap. The STM, along with generation constraints and clinicians' notes are utilised by an algorithm built on the foundation of a Walker frequency distribution engine (McLachlan et al, 2016).

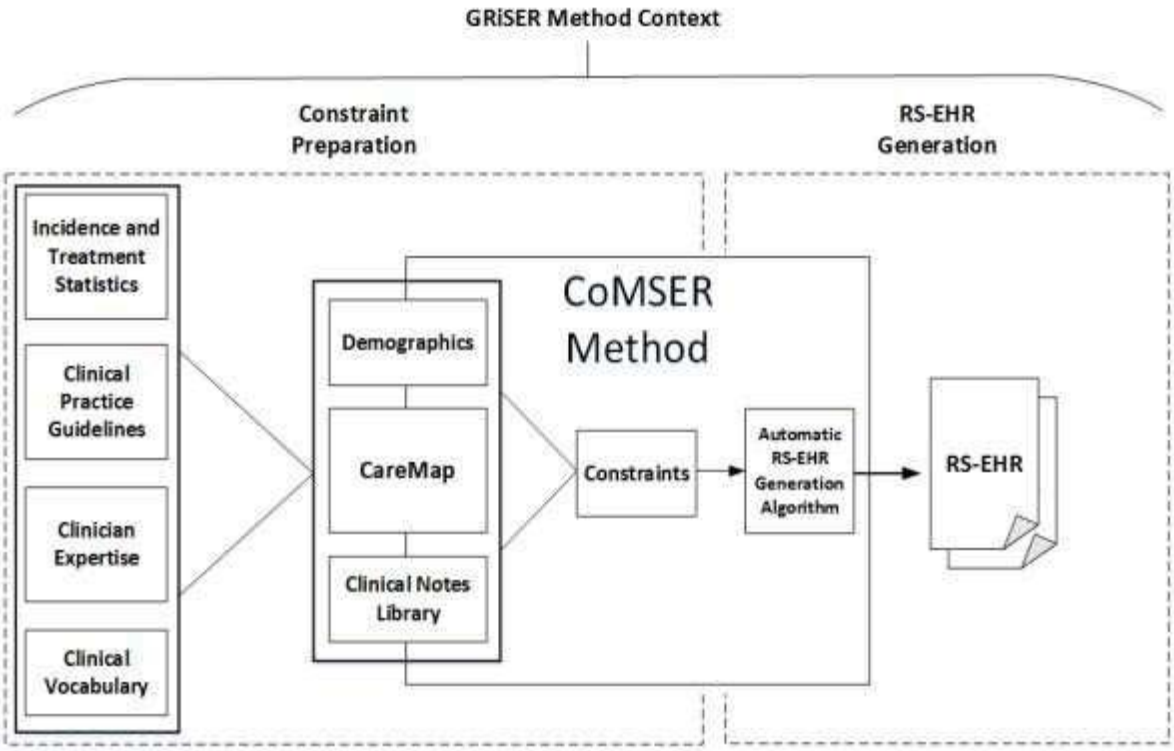


Figure 6: Context Diagram for the CoMSER Method (from McLachlan et al, 2016)

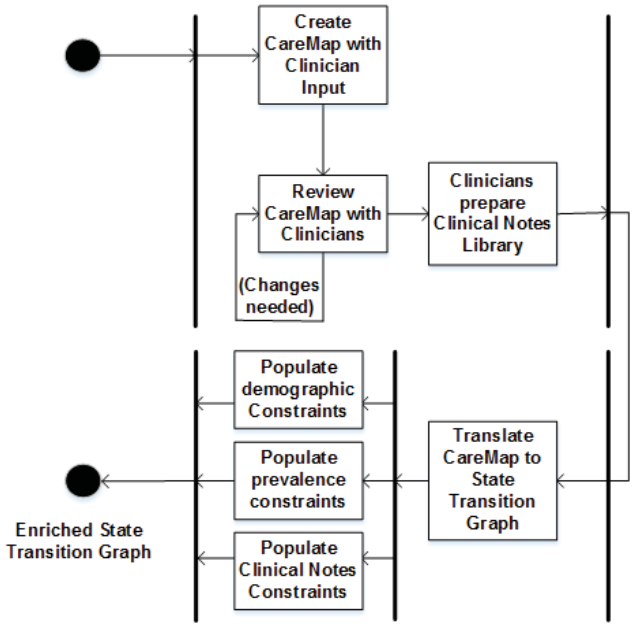


Figure 7: CoMSER UML Activity Diagram (from McLachlan et al, 2016)

4.7 Conclusion

It should be noted that the focus in SDG should not simply be on producing the data; it should equally be on the intended use which that synthetic dataset is to be applied. This could be a simple or direct use such as releasing synthetic data that is statistically accurate to test predictive models. It could be a way of releasing patient health records that, while still accurately reflecting the true range of disease incidence or treatment outcomes seen within a given population, are entirely synthetic and therefore completely without the risk of harm to the privacy of any real patient.

We have seen that synthetic data can take a number of forms, each described by their character and the context that real data plays both within the input and generation method as well as whether it is in some way extant in the resulting synthetic dataset. At the lower end of the spectrum comes that data which while having some personally identifying fields removed or anonymised, remains largely unchanged from source. If the purpose of generating synthetic data is to reduce the risk of exposing information in the real data, then releasing anonymised or partially synthetic data should never be entertained.

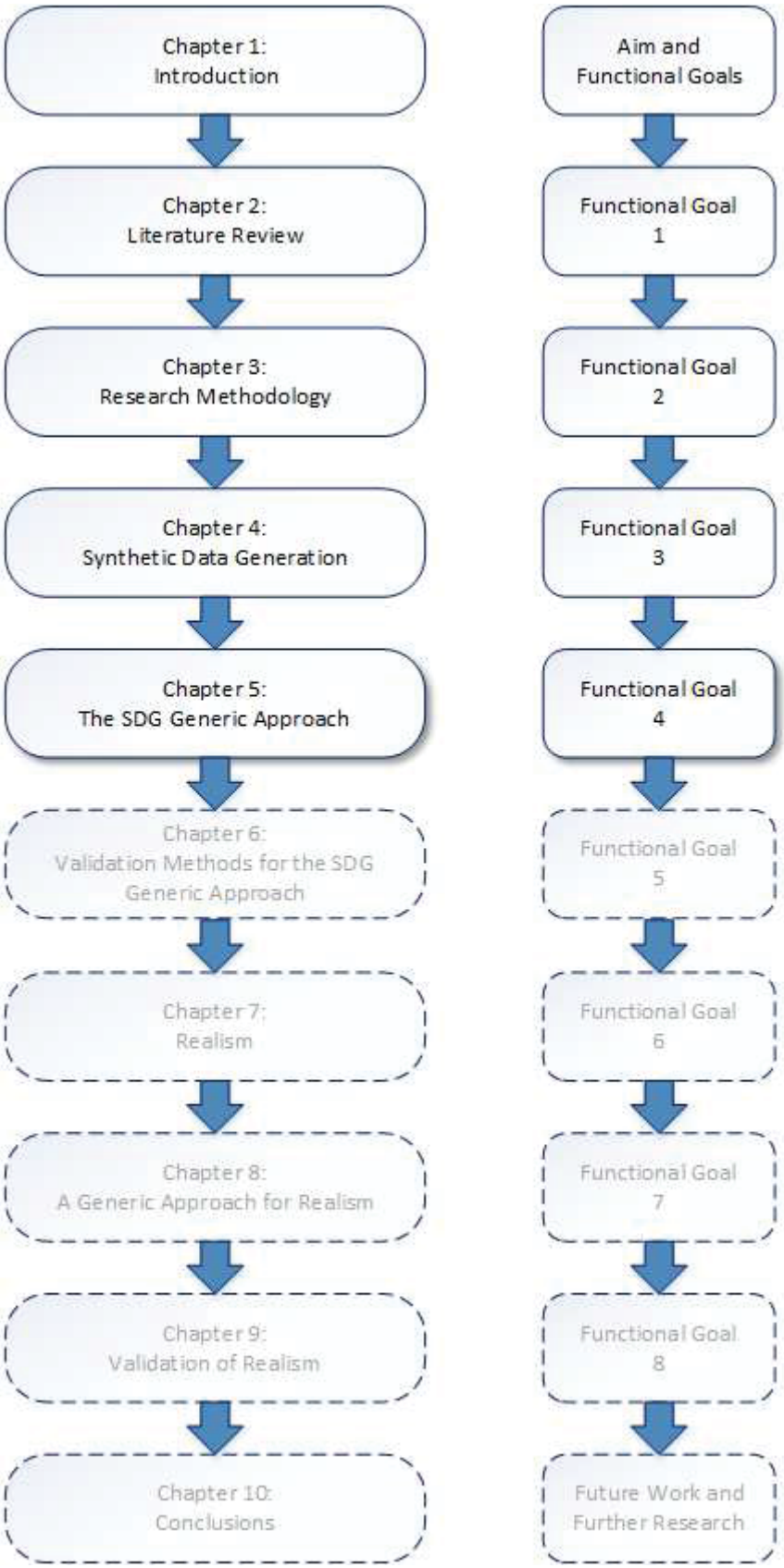
At the upper end we find the fully and truly synthetic data types. Truly synthetic data generated without the use of real data should always be the gold standard, however in some instances fully synthetic data may be a necessary compromise. Those researchers using this methodology do so in order to realise the true nature of the real data they seek to synthesise, and should do so in controlled circumstances that expressly limit the potential for data exposure, deliberate or accidental. Further, in situations where the fully synthetic data compromise is required, the Knowledge Discovery in Database (KDD) methodology discussed in chapter 8 of this thesis should be employed to ensure no record from the real dataset comes in contact with the generation process or can ever exist in the resulting synthetic dataset.

Those using real data in their generation processes need to be aware of the privacy issues and potential risks that come from releasing anonymised or poorly synthesised data. While it is often not obvious within that one dataset alone, the inherent risk often comes when others link or merge records from a number of data sources together. It is at this point that flaws in the anonymisation or synthesis process become apparent. Where the ethical and legal issues of releasing data meet the reality of risk, and individuals' private data is drawn into focus. Once that information is exposed, it is impossible to cover it back up.

This page intentionally left blank

“*Google* is not a synonym for *research*.”

(Dan Brown, The Lost Symbol)


Chapter 1:
Introduction
Chapter 2:
Literature Review
Chapter 3:
Research Methodology
Chapter 4:
Synthetic Data Generation
Chapter 5:
The SDG Generic Approach
Chapter 6:
Validation Methods for the SDG
Generic Approach
Chapter 7:
Realism
Chapter 8:
A Generic Approach for Realism
Chapter 9:
Validation of Realism
Chapter 10:
Conclusions
Aim and
Functional Goals
Functional Goal
1
Functional Goal
2
Functional Goal
3
Functional Goal
4
Functional Goal
5
Functional Goal
6
Functional Goal
7
Functional Goal
8
Future Work and
Further Research

5. THOTH: The SDG Generic Approach

This chapter presents THOTH, the generic approach followed by the vast majority of synthetic data generation methods. It discusses the component steps in order; providing critique and highlighting inherent weaknesses.

This chapter seeks to achieve functional goal 4:

Functional Goal 4. Develop the generic approach to Synthetic Data Generation.

This chapter is structured as follows:

5.1 Introduction to THOTH

In Egyptian mythology THOTH was originally worshiped as the moon god (Green, 1992). The moon provided light at night, allowing for measurement of the passage of time even in the absence of the sun. (Assmann, 2001; Green, 1992). Over time THOTH was increasingly more associated with the elements of wisdom and measurement, and the regulation of events and time, so much that he became the selected god of those that created all written records; *the Scribes* (Assmann, 2001; Green, 1992; Wilkinson, 2003). This thesis invokes both associations in titling the overall three-part generic approach after His name. The moon acted as a pseudo-sun providing the ability to make measurement when the sun was not available, in the same way that realistic synthetic (pseudo) data allows for systems and methods to be tested when real data is not available. Just as THOTH became associated with the concept of wisdom through the gifts of measurement, writing (making records) and regulation, our approach describes the previously undescribed generic approach to SDG drawn from a collection of literature, extending this with the gifts of a prescribed method to measure (characterise) and regulate (classify) the components necessary to that approach.

5.2 The Steps to SDG

A study of the literature collection identified in appendixes A and B yields the generic SDG approach identified in Figure 8. This approach incorporates the minimum and most common structural elements shared by all of the SDG methods. The generic approach best lends itself to representation as a pure waterfall model. This is primarily due to its similarity to the described structure and operation of the waterfall model; its cumulative and sequential nature where the start of the next phase is driven solely by completion of the last (Lydiard, 1992). Verification, which is almost never seen but should be a required step of any scientific endeavour, can only occur during limited opportunities at the end of each step of the method or after the SDG operation is complete (Lydiard, 1992).

1. IDENTIFY THE NEED FOR SYNTHETIC DATA: This involves not just recognising that synthetic data is required, but also establishing the justification, or reason for creating it. The most commonly expressed justification across the literature was that the synthetic data being created was necessary to replace real data containing personally identifiable, sensitive or confidential information.

2. KNOWLEDGE GATHERING: Knowledge gathering can involve a number of sub-steps assessing the requirements for the synthetic dataset being created. It usually begins with analysis of the data to be synthetically created, identifying such things as the necessary fields to be generated, the scope and constraints to be imposed.

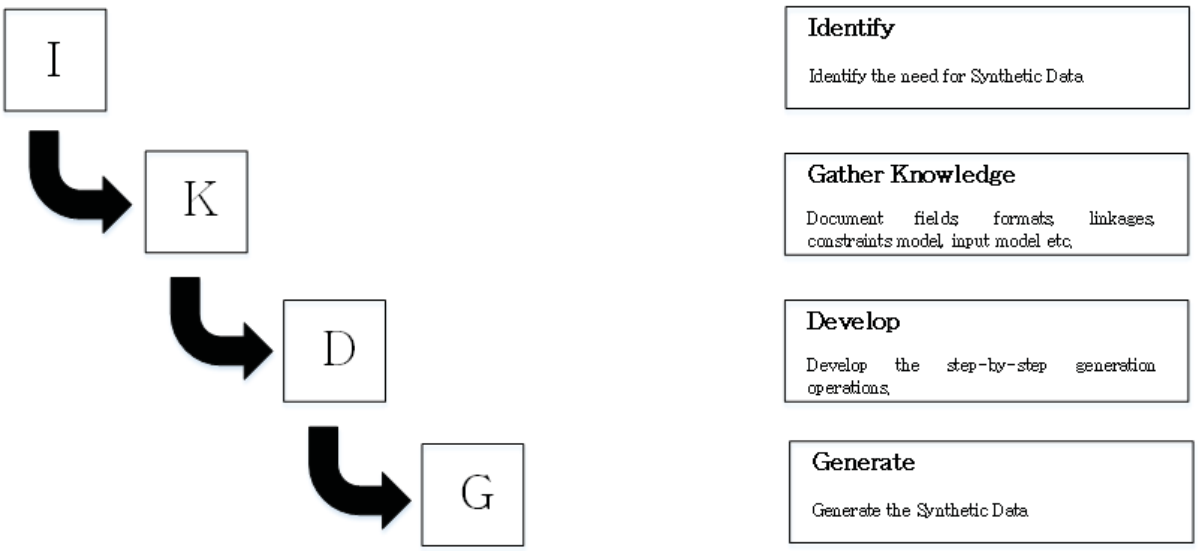


Figure 8: The Generic Approach to Synthetic Data Generation

3. DEVELOP THE METHOD OR ALGORITHM: It is not unusual for researchers to identify common solutions that have become preferred for a given research method or field; a method or algorithm that have drawn significant focused attention. Many of these algorithms have operational steps or processes which require focused attention, or for which data must be properly prepared. Developing the generation solution is as important as the need, and the level of attention paid during this step has a direct correlation to the quality of the output.

4. GENERATE THE SYNTHETIC DATA: At its simplest the process of generation involves presenting any seed data, conditional requirements and/or constraints to the generation algorithm which performs processes that output synthetic data for use or storage.

5.3 Discussion of the Generic Approach

The generic approach represents a simple method for the conduct of SDG experiments. Simple methods are sometimes considered better, favoured because of their increased usefulness, reduced complexity and experiment time, all of which can reduce cost (Ishigami et al, 2000; Mahmoud, 1984; Nicoletti et al, 1991; Rosevear, 1984). However, the generic approach suffers from many of the same issues as any pure waterfall development model; it flows only in one direction meaning that any change in the requirements being modelled or issue identified in the resulting dataset requires expensive, time consuming and often complete redevelopment and retesting (Parnas & Clements, 1986). Waterfall models entertain no flexibility and generally require complete redesign every time the use case changes.

Given that for those few solutions actually engaging in true validation it only occurs once development has ceased and the generation process is complete, it can be difficult to identify where any issue diagnosed in the synthetic dataset may have occurred (Lydiard, 1992). It becomes impossible to easily ascertain whether it has been an artefact of the seed data or statistics, a coding or algorithmic error in one or more steps of the developed generation process, an output error or a combination of all of these. It is in these instances where the waterfall approach really demonstrates why many believe it to be a dead model; one to be avoided at any cost (Pressman, 1998).

The inflexibilities inherent to the generic approach limit our ability to adapt it when inputs (seed data or statistics) or dataset requirements change. It is also a causal factor that restricts the applicability of many SDG approaches across fields. For this reason, a more adaptable and agile approach to SDG development is recommended.

Many SDG papers focus disproportionately on the performance of their method, going into great detail to report how many thousand records their model can generate in the shortest amount of time using

how few CPU cycles (Agrawal et al, 1996; Agarwal, Aggarwal & Prasad, 2000; Jain et al, 1997). In a very high percentage of SDG papers it is these metrics that form much of the verification and validation highlight, with far less attention paid to the accuracy of the model, the synthetic-ness of the generated datasets, or the level to which they attained the *realism* goal introduced in their opening paragraphs. Definition, design and validation of any generation method should be the key focus of the work, not the speed at which it generates what may prove to be potentially useless rows of data (Agrawal et al, 2015).

5.4 Improving the Generic Approach with THOTH

Pre-planning and preparation may represent one mitigation to the unidirectional linear nature and inflexibility of the SDG's waterfall presentation. This is where the THOTH approach is able to assist. THOTH encourages the SDG researcher to perform decisive steps prior to engaging in the generation process.

The three-step THOTH approach shown in Figure 9 begins with identifying the level of syntheticness desired in the generated data. The characterisation level is selected from those listed in Table 4, and once established is one of the elements that aids in the second step, the selection of a generation model classification from those listed in Table 3. Once the synthetic characterisation and model classification have been selected the researcher can engage the generic SDG approach described in section 5.2, however they begin with an additional level of wisdom that comes from knowing where they are going (the level of syntheticness required of their efforts) and a framework for how they are going to get there (the informed selection of a generation model).

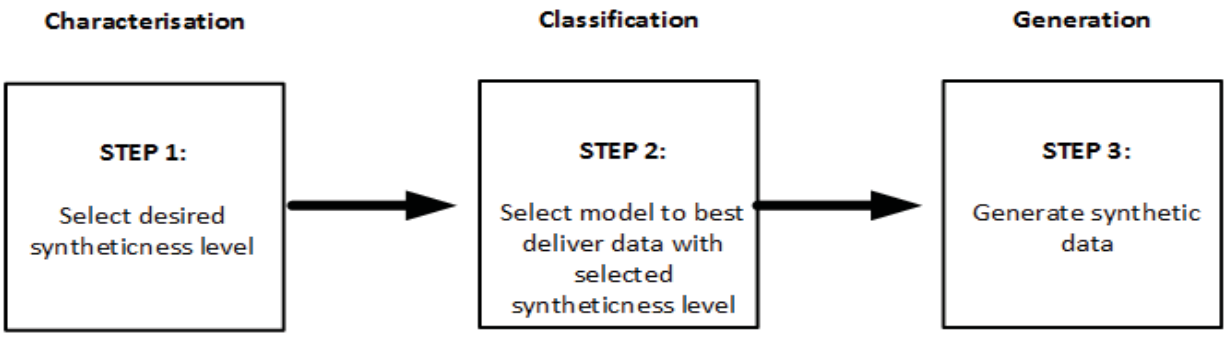


Figure 9: The three-step THOTH approach

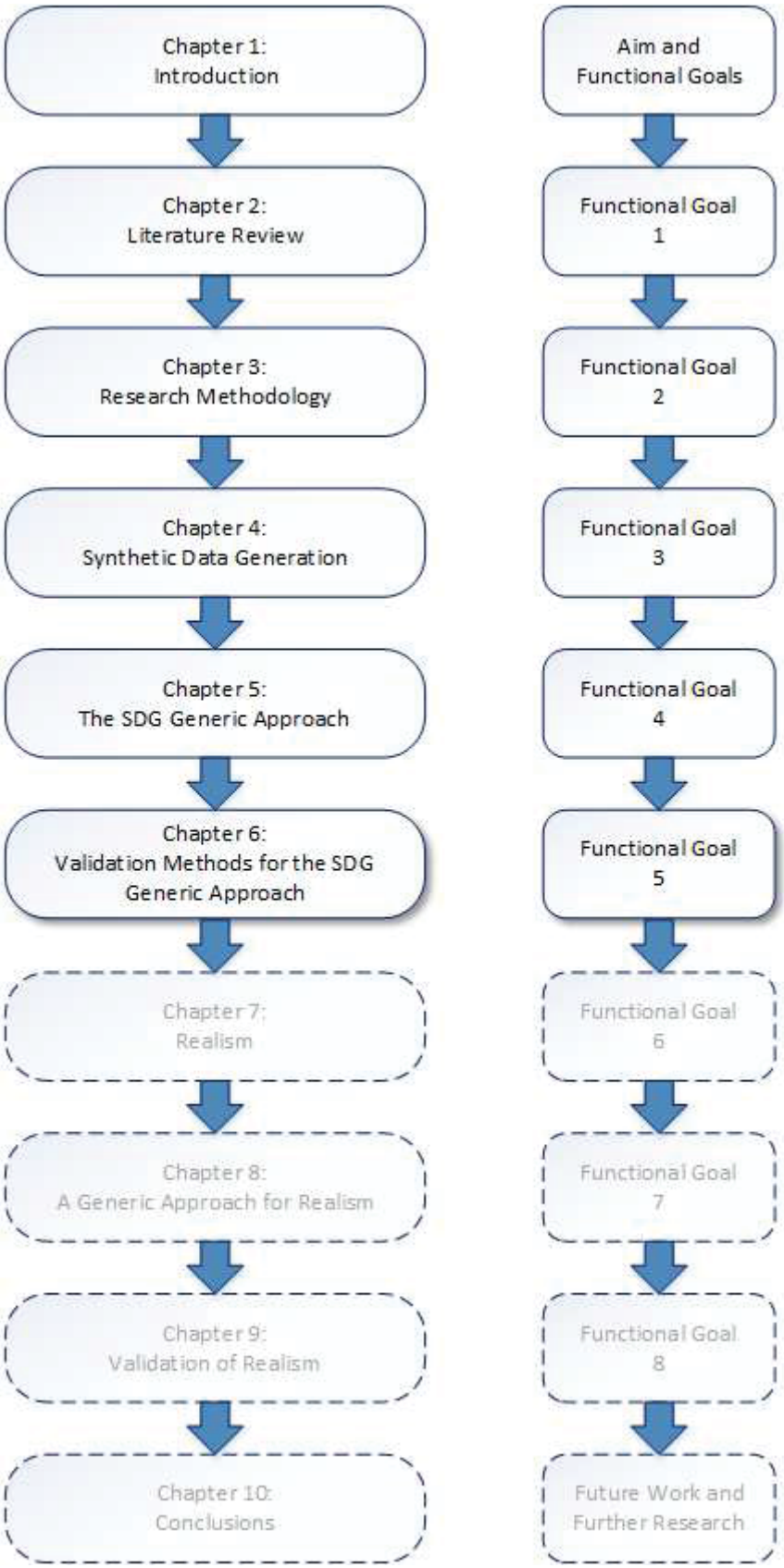
5.5 Conclusion

A generic four-step waterfall approach is common to the majority of SDG methods. This approach sees researchers move through identifying a need for synthetic data to gathering the knowledge they feel is

necessary to its generation. They then develop or more often customise a method of generation common to their field or solution needs before finally generating the synthetic data. True validation of the input, generation process and output of this waterfall approach is difficult and rarely occurs, with many reporting speed, memory usage or other system-based metrics as the core deliverables of their validation model. Validation of the generation method and resulting synthetic data is necessary if we are to demonstrate its application to the problem that necessitated generation. Without validation we certainly cannot hope to understand whether the realism goal was met.

Incorporation of the THOTH approach would benefit the researcher, providing additional knowledge of their requirements and direction in the overall synthetic data generation approach. Researchers using the classification and characterisation steps of THOTH prior to generation would have greater awareness and should see less surprise in the results they produce.

Researchers who remain unaware of the outcome they desire and who have not used a reasoned approach to the selection of a generation method can be likened to a new resident in a large city wandering the streets at night without access to a map. Further, those who do not validate their selected method, the synthetic data it creates or the realism element that was a claimed requirement of the project, and those who chose not to include any discussion of validation detail in the articles they author, can only leave the reader questioning any successes the researcher claims in the overall approach.


Chapter 1:
Introduction
Chapter 2:
Literature Review
Chapter 3:
Research Methodology
Chapter 4:
Synthetic Data Generation
Chapter 5:
The SDG Generic Approach
Chapter 6:
Validation Methods for the SDG
Generic Approach
Chapter 7:
Realism
Chapter 8:
A Generic Approach for Realism
Chapter 9:
Validation of Realism
Chapter 10:
Conclusions
Aim and
Functional Goals
Functional Goal
1
Functional Goal
2
Functional Goal
3
Functional Goal
4
Functional Goal
5
Functional Goal
6
Functional Goal
7
Functional Goal
8
Future Work and
Further Research

6. Validation Methods for the SDG Generic Approach

This chapter demonstrates application of the four established CM validation models discussed in Section 2.5 to SDG, and discusses elements that may be used to counteract the inherent weaknesses discussed in Section 5.2.

The research in this chapter was carried out in order to achieve functional goal 5:

Functional Goal 5. Develop a new method for SDG validation based on established methods in CM.

This chapter is structured as follows:

6.1 Introduction to SDG Validation

The addition of basic validation methods is seen in some literature. Basic validation operations can demonstrate a methods' consistency with established approaches for dealing with similar input or output conditions, or delivery of synthetic data comparable to real-world observation. It is important to understand how each validation method is applied to the generic approach seen in the literature so that we may understand what degree of validation each actually provides.

6.2 Simplified Generalised Narrative of Published SDG Articles

Analysis of the wider collection of SDG articles lead to the realisation that within the majority there exists a standard set of six narrative beats, two for each of three narrative themes. This simplified generalised narrative with its bracketed beats is presented in Table 5.

Table 5: Simplified Generalised Narrative of SDG Articles

1. Justification

It is hard because of [some difficulty] to get real data for [some use], so we developed a new method to generate synthetic test data for this purpose.

2. Operational

Our method uses [some input] to generate the synthetic data using [some method].

3. Result

We performed [some action] and believe that the synthetic data created by our generation method is promising for [some reason].

Tables 6 – 8 demonstrate application of this simplified generalised narrative drawn from articles selected randomly from those listed in appendix A. Each table completes the relevant narrative sentence by identifying how that author populated the bracketed components.

Table 6: Justification Examples for Part 1 of the Simplified Generalised Narrative

Lead Author/Year	Difficulty	Use
Van den Bulcke, 2006	Limited available data	Testing of learning algorithms
Domingo-Ferrer, 2012	Privacy protection	Enable the release of data that is analytically useful
Efstratiadis, 2014	Limited available data	Provide sufficiently large samples or evaluate a wide range of possible outcomes
Giannotti, 2005	Privacy protection	Prediction of moving objects in a network
Mateo-Sanz, 2004	Privacy protection	Release of dataset that maintains confidentiality

Table 7: Operational examples for Part 2 of the Simplified Generalised Narrative

Lead Author/Year	Input	Method
Van den Bulcke, 2006	Known network structures	Neighbour addition and cluster addition with random selection
Domingo-Ferrer, 2012	Hierarchical nominal data	Centrality-based mapping
Efstratiadis, 2014	Historic observational statistics	Multivariate stochastic simulation and Markov models
Giannotti, 2005	Global and Group Parameter configuration files	Change behaviour modelling and log generation
Mateo-Sanz, 2004	Original microdata set	Covariance, Cholesky decomposition, matrix product and Pearson correlation matrix.

Table 8: Result examples for the Simplified Generalised Narrative

Lead Author/Year	Action	Reason
Van den Bulcke, 2006	Benchmark and performance test simulating known networks	Claimed resemblance to real transcriptional networks
Domingo-Ferrer, 2012	Demonstrational experiment with mathematical proofs	Claimed success of mathematical proofs
Efstratiadis, 2014	Comparative graphs on multiple generation runs	Claimed model advantages and software capabilities
Giannotti, 2005	Synthetic scenario comparison	Likeness of synthetic data to a synthetic scenario
Mateo-Sanz, 2004	Performance and complexity analysis	Claimed reduced disclosure risk score

6.3 Improving the SDG Generic Approach with Validation

This research found that the majority of SDG approaches, including those listed in the appendices set the overarching goal of simply producing synthetic data that the authors feel can replace real data. The focus is heavily weighted toward the outcome, the synthetic data, and the validation methods used tend to be singular or simplistic. Some consist of direct comparisons between either the entire dataset or fields within the synthetic data to observations from the real data (Bozkurt et al, 2011), or of graphical or statistical comparisons between the two (Ascoli et al, 2001; Efstratiadis et al, 2014; Gafurov et al, 2015). Many discussed no performance of validation at all (Brinkhoff et al, 2003; Brissette et al, 2007; Giannotti et al, 2004). The goal of their publications comes down to answering the question of whether the synthetic data they have created can be switched out for real data without understanding why, or even whether the data indeed possessed all of the properties or realistic elements necessary to its being a true substitute.

A necessary inclusion for any SDG model should be the requirement of robust validation; incorporation of a three-step model which validates the initial seed data, statistics or input conditions (*input validation*), the generation method (*method validation*), and finally the synthetic data that is created (*output validation*). Application of this three-step validation model to the generic approach to SDG is demonstrated in Figure 10. At any point where validation presents an issue, the researcher should follow the green path back to whichever prior step requires rework. Documentation supports the validation effort. Incorporation of the validation approach in any publication will improve and complete the presentation and allow others to adequately assess whether a project has indeed met its goals with the success the authors claim. It ensures that SDG experiments can be independently verified to the same standard as other scientific endeavours.

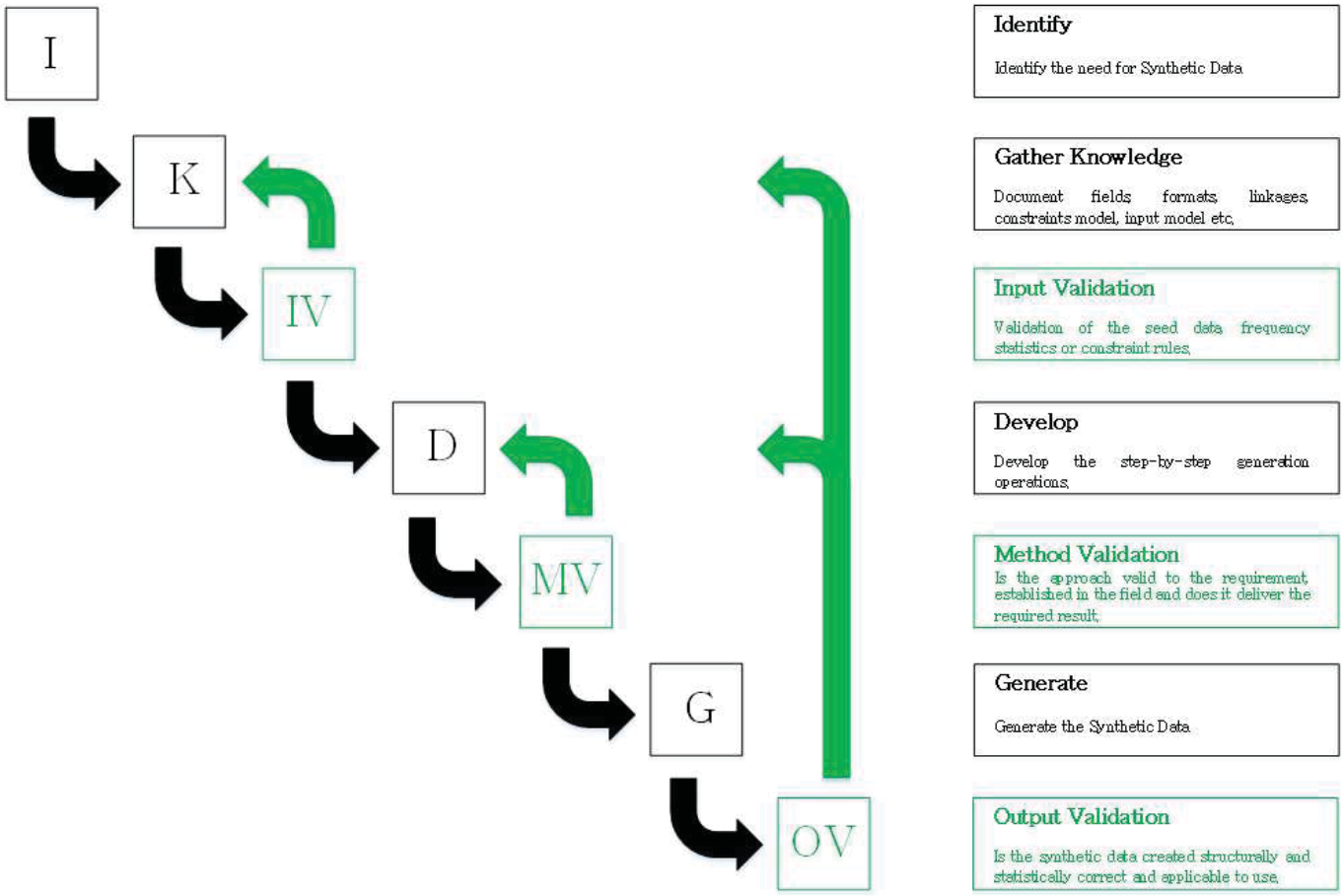


Figure 10: The Improved Generic Approach to Validation for Synthetic Data Generation

6.4 Validation Approaches in the Domain of Computational Modelling

It would be difficult for SDG authors to claim that existing validation models did not exist in the literature. Given the relationship and similarities of SDG to the domain of Computational Modelling discussed in Chapter 2, the validation approaches used there may have some application to SDG. This section reviews the four primary validation models from Computational Modelling that were introduced in section 2.6. Each validation approach has been overlaid onto the SDG generic approach diagram to highlight where and how it applies, and demonstrate how each might be used as a component in one possible validation model.

6.4.1 Grounding

The grounding validation approach demonstrated in Figure 11 utilises external comparison (Carley, 1996). It does so in one of two ways:

1. Through demonstration and claim that the approach used was consistent with approaches used by other authors in their field when generating data that is based on similar needs, inputs or requirements. This approach was the most frequent one seen in the literature listed in Table 2, or;
2. By demonstrating that the starting statistical conditions used in their approach were consistent with those observed in the real world.

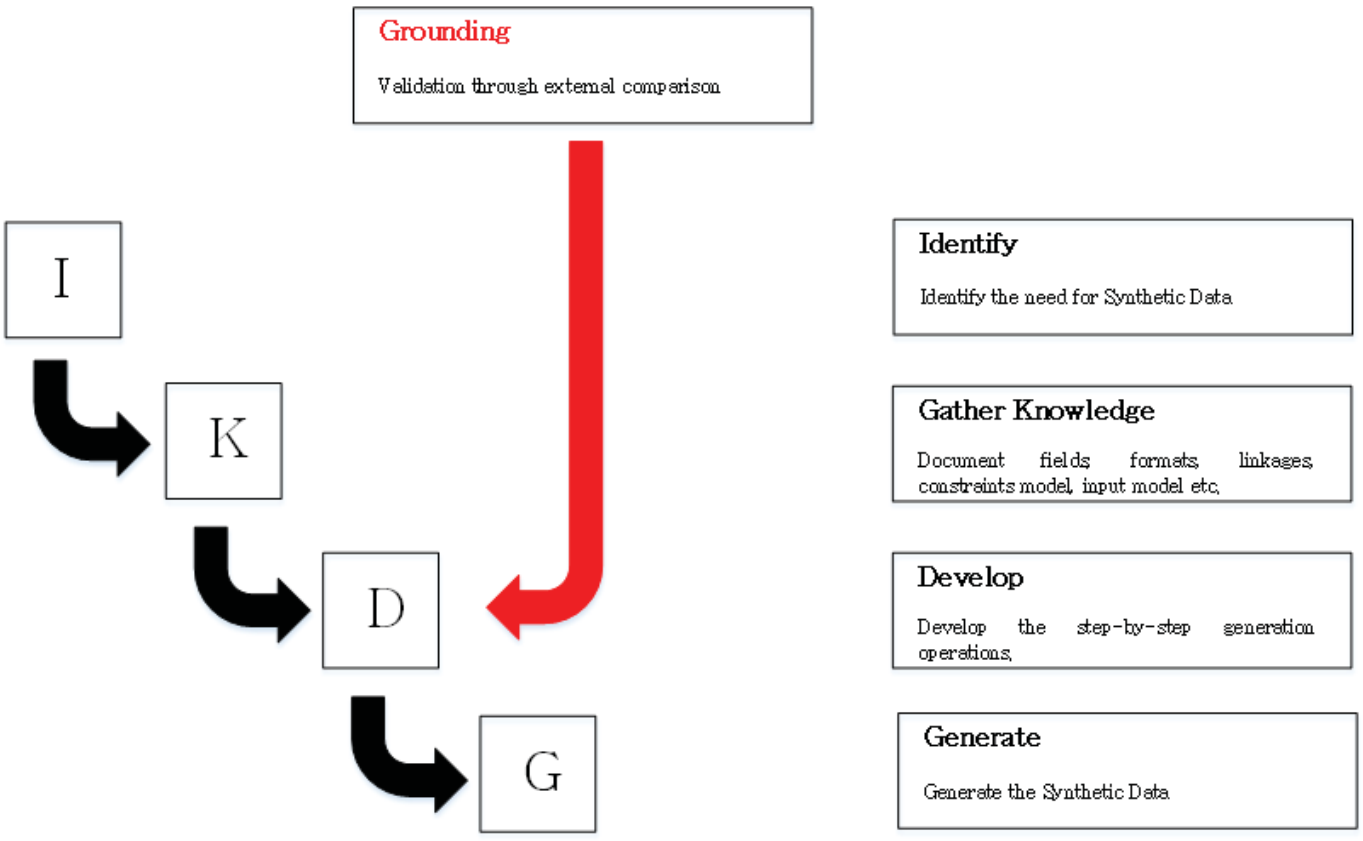


Figure 11: Grounding Validation of the Generic Approach

6.4.2 Calibration

The calibration method of validation as displayed in Figure 12 involves repeated manipulation or tweaking of the rules or algorithmic conditions over successive generation events until the output data is consistent with observations (Carley, 1996).

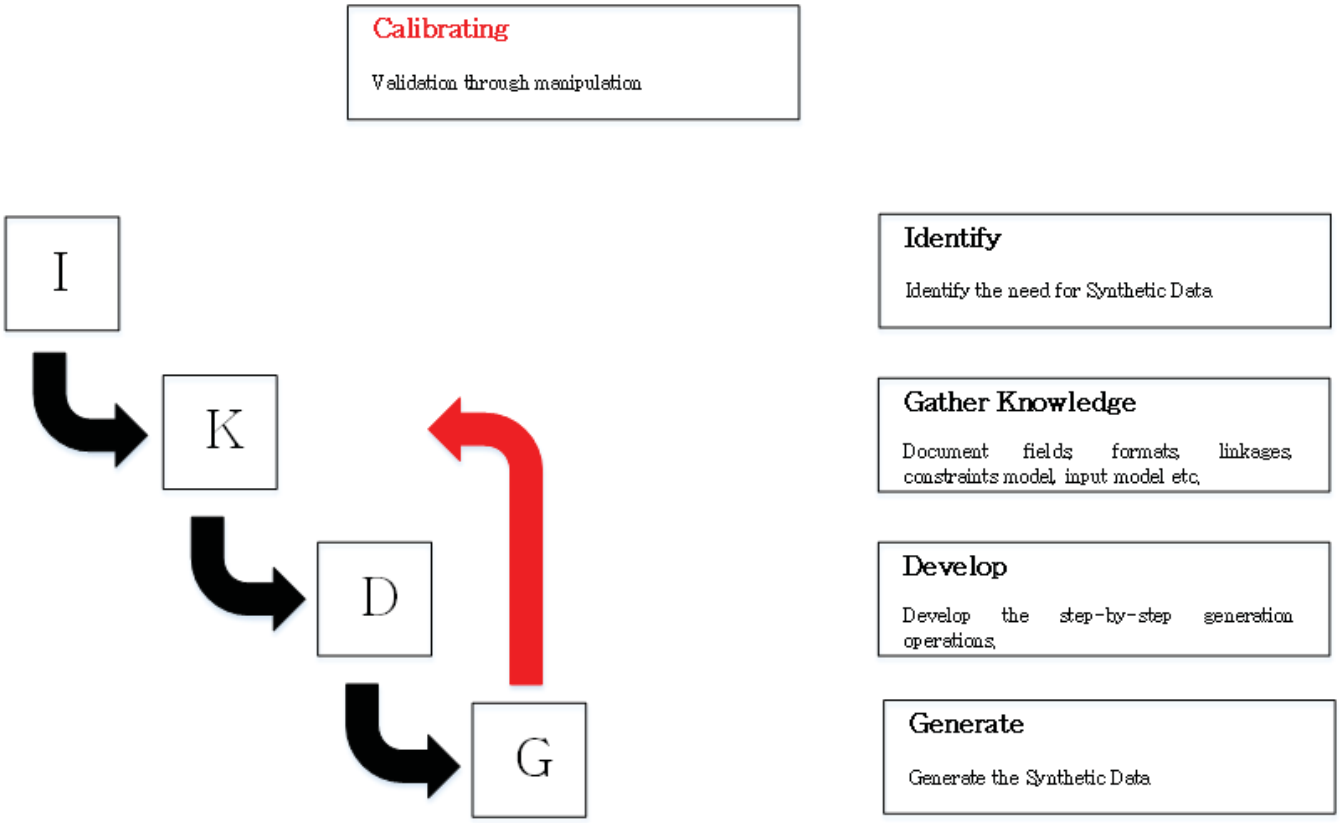


Figure 12: Calibration Validation of the Generic Approach

6.4.3 Verification

The verification validation approach merely compares statistics derived from the synthetically generated data against those derived from or contained within the input data or modelling constraints (Carley,

1996). As shown in Figure 13 the verification approach is not of itself iterative, but may be used in conjunction with the calibration validation method in order to achieve the author's required level of consistency (Carley, 1996).

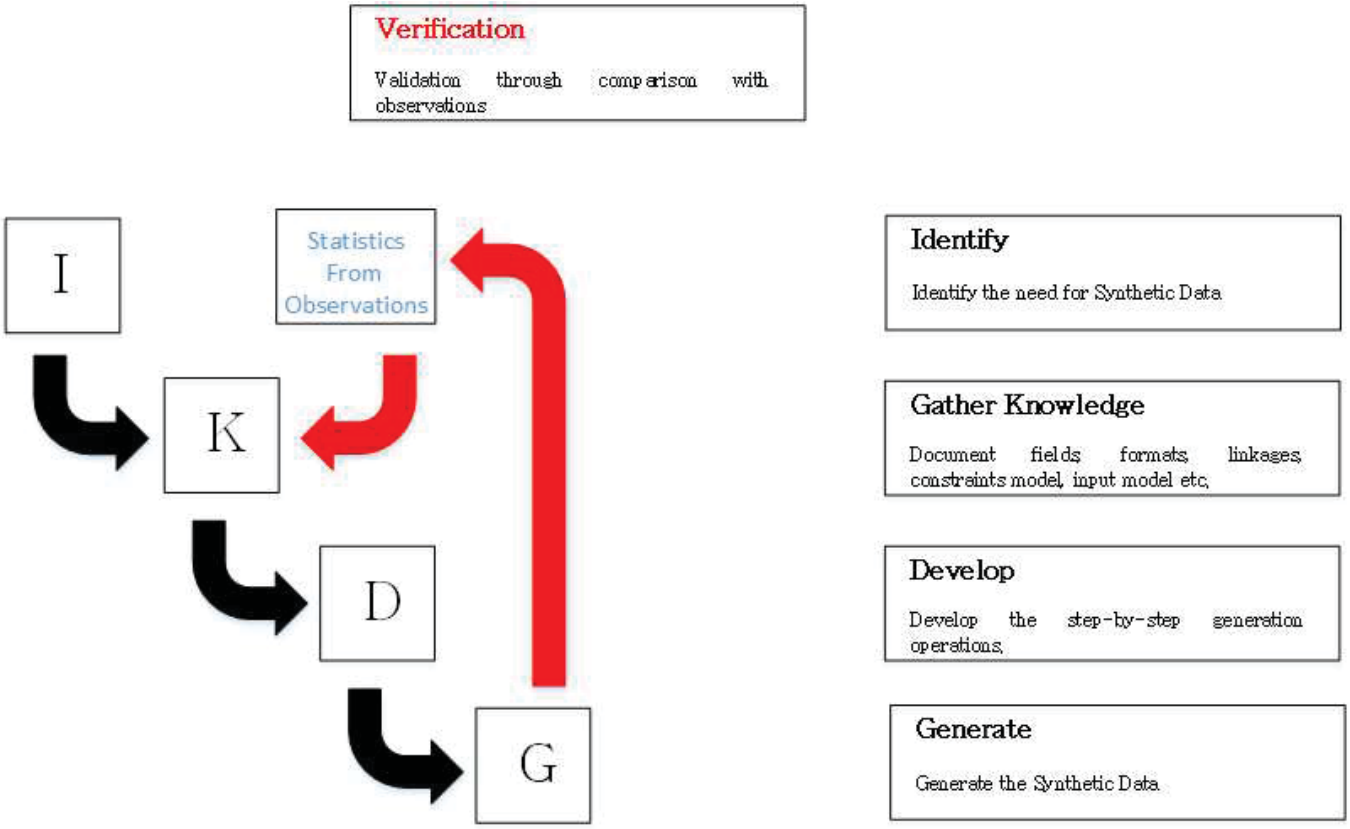


Figure 13: Verification Validation of the Generic Approach

6.4.4 Harmonising

The harmonising approach is the most complicated and its use was not observed in any of the literature included in this study. This approach requires two sets of real-world observational data which is used during the execution of a set of four steps (Carley, 1996). These steps are shown in Figure 14:

1. The first step requires a complete run of the SDG model (in Black). A calibration validation is performed to compare the synthetic data against the first set of real-world observations (in RED).
2. The second step involves recalibrating the model using an estimation benchmark generated by a linear model (in BLUE).
3. The third step involves cross-validation of the results after the first two steps. Cross-validation of these results involves validation against the second set of real-world data while using predictions taken from the first set (in GREEN).
4. The fourth and final step contrasts both the synthetic data and linear model's predictions against the second set of real-world data.

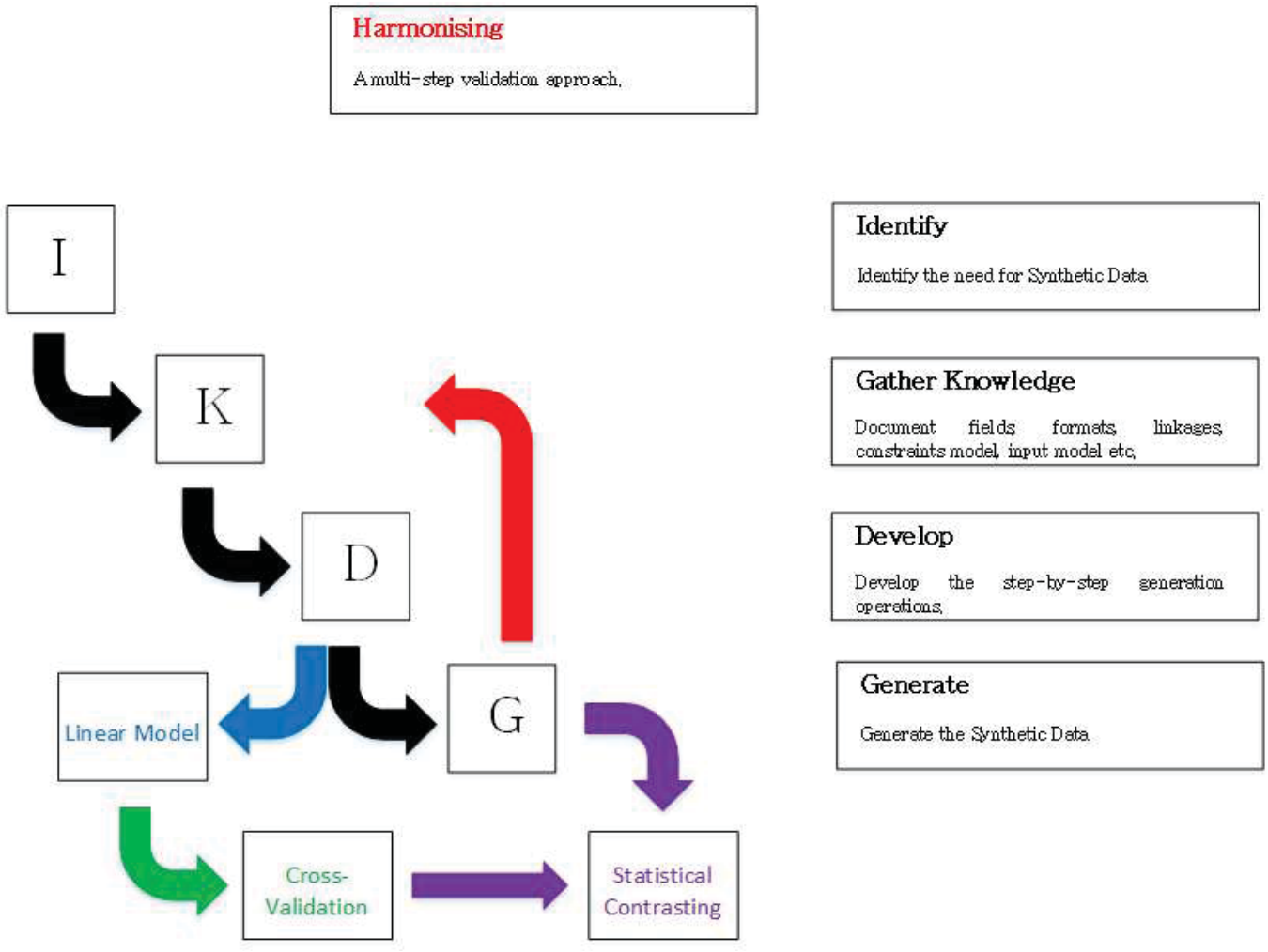


Figure 14: Harmonising Validation of the Generic Approach

6.5 Case Study

The following case study utilises the CoMSER SDG application that was introduced in section 4.5 to apply, discuss and contrast the CM validation approaches with the improved generic approach to validation presented in section 6.2 of this chapter.

6.5.1 Grounding Validation

A grounding validation approach might see the CoMSER generation model compared to other approaches used in the creation of synthetic electronic health records (S-EHR). An alternate grounding approach would compare the incidence and treatment statistics used to populate the CareMap with those publicly available from the Ministry of Health's website. The grounding approach's limitations come from; (a) the fact that it only checks whether prior work has used a similar generation method, and makes no assessment of whether that model represents either a best fit for the required solution; (b) that it provides only limited quantitative checks of the starting conditions used by the generation method against observed data, and; (c) that it makes no assessment of the appropriateness, applicability or correctness of the resulting synthetic data.

6.5.2 Calibration Validation

Use of the calibration validation model with CoMSER would see iterative generation batches run. After each batch run the synthetic data would be checked against the observed data it seeks to emulate, with the rules, constraints and other input variables receiving minor tweaks and changes until the synthetic data appears to match the original observed dataset. The calibration model makes no assessment as to whether the generation method is the most appropriate to create the required solution, and results in a solution that is difficult to repeat or apply to other datasets, even when they are for the same type of problem or within the same domain. The manipulations are uniquely specific to only that observed dataset and that set of generation conditions.

6.5.3 Verification Validation

Verification validation of the CoMSER method would extract a selection of statistical attributes from the input or observation data, seeking to match with these with same drawn from the synthetic data. Verification validation might be used by CoMSER to verify that patient ethnicities, treatment modalities and other quantifiable outcomes have been generated consistently. Verification validation does not seek to identify or isolate issues in the constraints, rules or even the generation algorithm. In that way it would let the researcher know that there was an issue in the synthetic data, but could not identify the root cause.

6.5.4 Harmonising Validation

The Harmonising validation approach would see CoMSER being used to generate two sets of synthetic health records, an initial set followed by a second after a single calibration validation pass. Each set of S-EHRs would need to be generated using observations, constraints and rules collected from one set of real-world EHRs. The generation model is then recalibrated with an estimated linear model before cross-validation of each set of synthetic health records against predictions drawn from a second set of

real EHRs. In this way the Harmonising validation model is both complicated, and requiring of two sets of real EHRs. This factor alone may render the harmonising model untenable for an example like CoMSER, as the CoMSER model's primary goal was the generation of S-EHR without the need for real EHRs.

6.5.5 The Improved Generic Approach

The Improved Generic Approach to validation presented by this work recommends that researchers should validate across the three main stages, validating the knowledge to be used as the input for the generation method, validating the generation method itself and finally, validating the synthetic data that has been returned. If an issue or inconsistency is identified, the Improved Generic Method encourages returning to the prior stage to address the issue. In this way it is the only validation approach identified that addresses all three primary stages of a generic SDG approach, and the only one that does so both during execution and after completion. Most validation models only perform their process after the entire execution has completed.

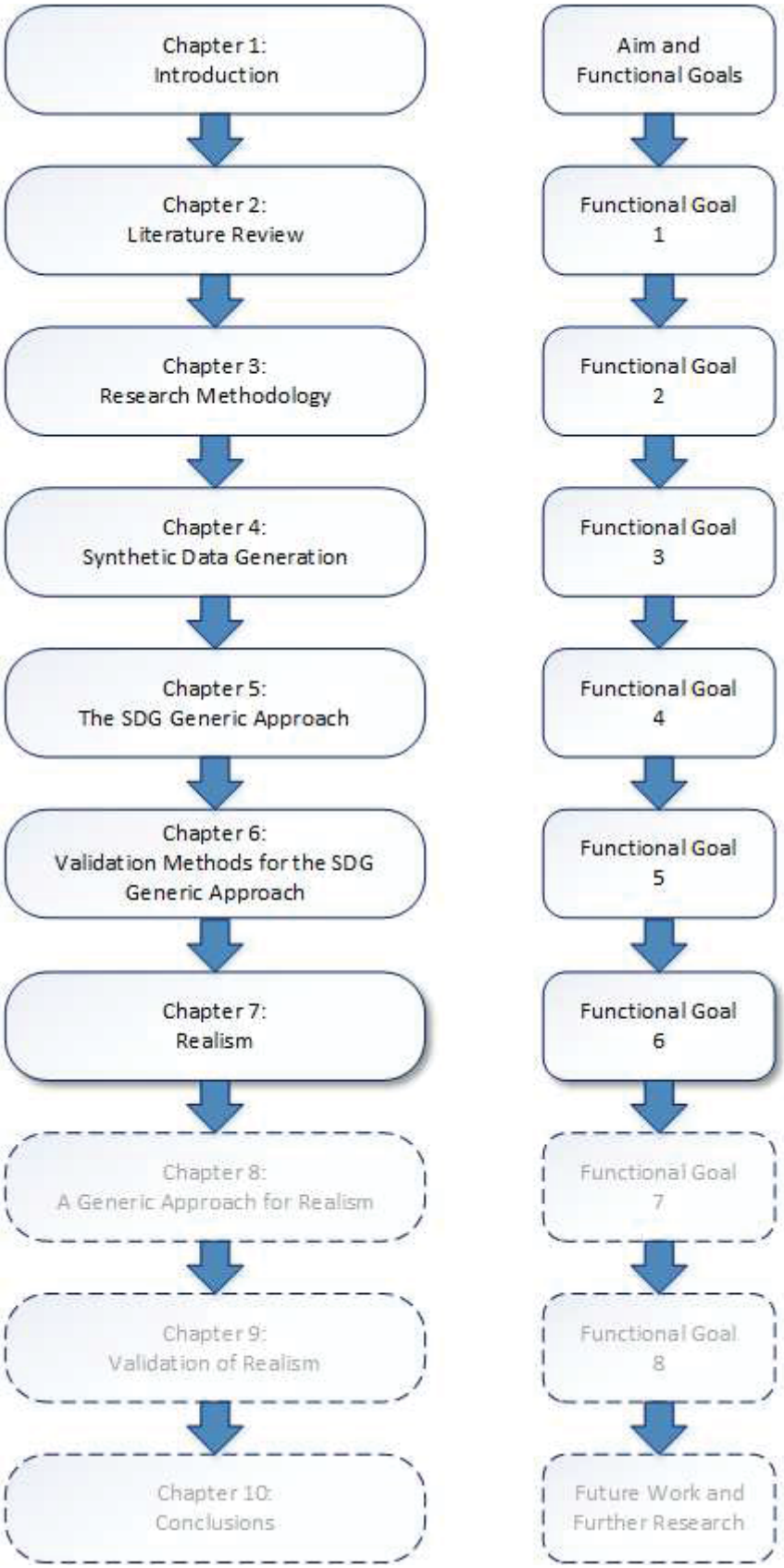
6.6 Conclusion

This chapter has presented a validation approach that incorporates validation at three key points in the generic SDG model; verifying the input, method and output components. We have seen that validation models exist that could have been drawn upon by synthetic data researchers, and while each on its own may not perform validation to the same standard as the approach presented in section 6.2, the use of any one model would have been significantly more effective than publication in validation's absence. To relate the proposed approach to established CM validation methods discussed in Section 2.5; method validation incorporates the overarching goals of the *internal, cross-model* and *conceptual validity* approaches, while input and output validation meets the *external* and *data validity* requirements.

The use of validation allows the SDG researcher to meet the scientific model documentation requirements, providing proof of claimed success and a platform for application assessment and repeatability by others. No experiment should be considered complete without rigorous testing and validation. Synthetic data modelling and generation are no different.

“A few years ago the city council of Monza, Italy, barred pet owners from keeping goldfish in curved bowls… saying that it is cruel to keep a fish in a bowl with curved sides because, gazing out, the fish would have a distorted view of reality. But how do we know we have the true, undistorted picture of reality?”

(Stephen Hawking)


Chapter 1:
Introduction
Chapter 2:
Literature Review
Chapter 3:
Research Methodology
Chapter 4:
Synthetic Data Generation
Chapter 5:
The SDG Generic Approach
Chapter 6:
Validation Methods for the SDG
Generic Approach
Chapter 7:
Realism
Chapter 8:
A Generic Approach for Realism
Chapter 9:
Validation of Realism
Chapter 10:
Conclusions
Aim and
Functional Goals
Functional Goal
1
Functional Goal
2
Functional Goal
3
Functional Goal
4
Functional Goal
5
Functional Goal
6
Functional Goal
7
Functional Goal
8
Future Work and
Further Research

7. Realism

This chapter begins the direct focus on realism in SDG, looking at how the term has been used, the types of validation applied and concluding by proposing a universal definition.

This chapter sets out to achieve functional goal 6:

Functional Goal 6. Develop a definition for realism in SDG.

This chapter is structured as follows:

7.1 Introduction to Realism

The property of *realism* in SDG is seen to bring a greater degree of accuracy, reliability, effectiveness, credibility and validity to the testing process (Bozkurt & Harman, 2011; Whiting et al, 2008; Williams et al, 2007). *Realism* is considered desirable and important; whether we are using real, partially real, de-identified or entirely synthetic datasets (Alessandri et al, 2015; Bolon-Canedo et al, 2013; Mouza et al, 2010; Sperotto et al, 2009; Tsvetovat et al, 2005; Whiting et al, 2008; Zanero, 2007) however the plethora of current literature demonstrates significant variation in the way that *realism* is discussed and defined, and in the approaches and methods used to produce the *realistic* dataset. In the absence of a clear definition and framework for understanding *realistic* data, the process to assess and validate the level of *realism* in the resulting dataset is severely challenged.

7.2 The Realism Component of Current SDG Literature

Many of the SDG projects in the appendices discuss some intention towards realism. This often consists of a brief statement about the need for a realistic synthetic dataset that can adequately mitigate the need for real data which may not be readily available or comes with issues that prevent release (Bozkurt & Harman, 2011; Domingo-Ferrer et al, 2012; Gianotti et al, 2004). While only 296 of the 7,746 SDG articles located also used the terms *realism* or *realistic*, for completeness a small sample of those that

did not were also reviewed. The results of this additional analysis identified that around one third of all SDG articles use alternate but functionally similar language for realism, such as: "*authentic"* (Barse et al, 2003), "*accurate to real structures"* (Ascoli et al, 2001), or provide that the "*replacement of real data"* is the goal of their SDG efforts (Domingo-Ferrer et al, 2012).

Very few authors (n = 2) define or provide a foundation for understanding realism within the context of their work. In both cases the definition was limited and vague; implying only that the aim of realism in SDG was that the resulting synthetic dataset needed to be a representative replacement for real data (Sperotto et al, 2009) and comparably correct in size and distribution (Killhourhy et al, 2007). In both cases validation of realism in the resulting synthetic dataset was not discussed.

7.3 Defining Realism from the Literature

If one were seeking to identify a definition for the concept of realism from SDG literature alone, one would not readily be found. Appendix B includes columns identifying the language and location where authors identified realism as a necessary component of their SDG output. It was intended to include a column within the annexed table to demonstrate the definition and characteristics of that realism, except that in all cases none was found. It can be seen from the table that validation methodology was also lacking or only superficially present, meaning that the vast majority of these projects fail to report their processes and outcomes in a manner consistent with the scientific method.

7.3.1 Understanding Realism

One characterisation for the term *realism* is the view that something *aims to be true or very nearly true*; that the object in question is equivalent to something observed (Levin, 1984). Authors discussing the issue of scientific realism propose a number of questions for consideration; Why should we believe a particular scientific theory to be true? Why should we believe that all of the entities suggested by our best theories are real? Why shouldn't we just consider these theories as mere instruments for the identification, systemisation and prediction of observable phenomena without attributing any reality to the invisible entities they theorise about? (McMullin, 1984; Psillos, 2005). McMullin (1984) asks that we also question how any belief in realism was formulated and how reliable that belief may be because if nothing else, realism must be clarified so that a belief in its existence can be defended. A theory or claim should only be considered confirmed when the realism of the input data, judged according to a scientific standard, leads to increasingly more accurate predictions (Laymon, 1984). In limited circumstances authors may find some acceptance for idealism or inference but this should not be relied upon in situations where overall realism is actually something we can ascertain and compare (Laymon, 1984). Actual knowledge cannot exist without truth and any case for scientific success must be based on the strength and appeal of that truth (Levin, 1984). Accordingly, those researchers developing SDG methods should not claim either overall success or more specifically, realism in the synthetic data they

delivered, without also demonstrating the realistic elements of whatever input data or knowledge was used in its creation, to an acceptable scientific standard.

In the case of synthetic data, we would seek to identify realism as the sum of two levels of knowledge. ***The first level*** is those things which are extrinsic or overtly obvious; the structure. The structure is made up of the data fields and generalised values easily realised in observational data or any other inputs being used in the generation process. ***The second level*** consists of those elements of knowledge that are intrinsic to the real data and not easily noticed by the naked eye; the relationships, concepts, rules and representations that can be drawn out from within the real data and which must be present in synthetic data in order for it to be truly representative. While each element must be separately and independently assessed, it is only as a validated, verified accumulation of knowledge that they signify and support that realism can and does exist.

7.4 Realism and the Scientific Method

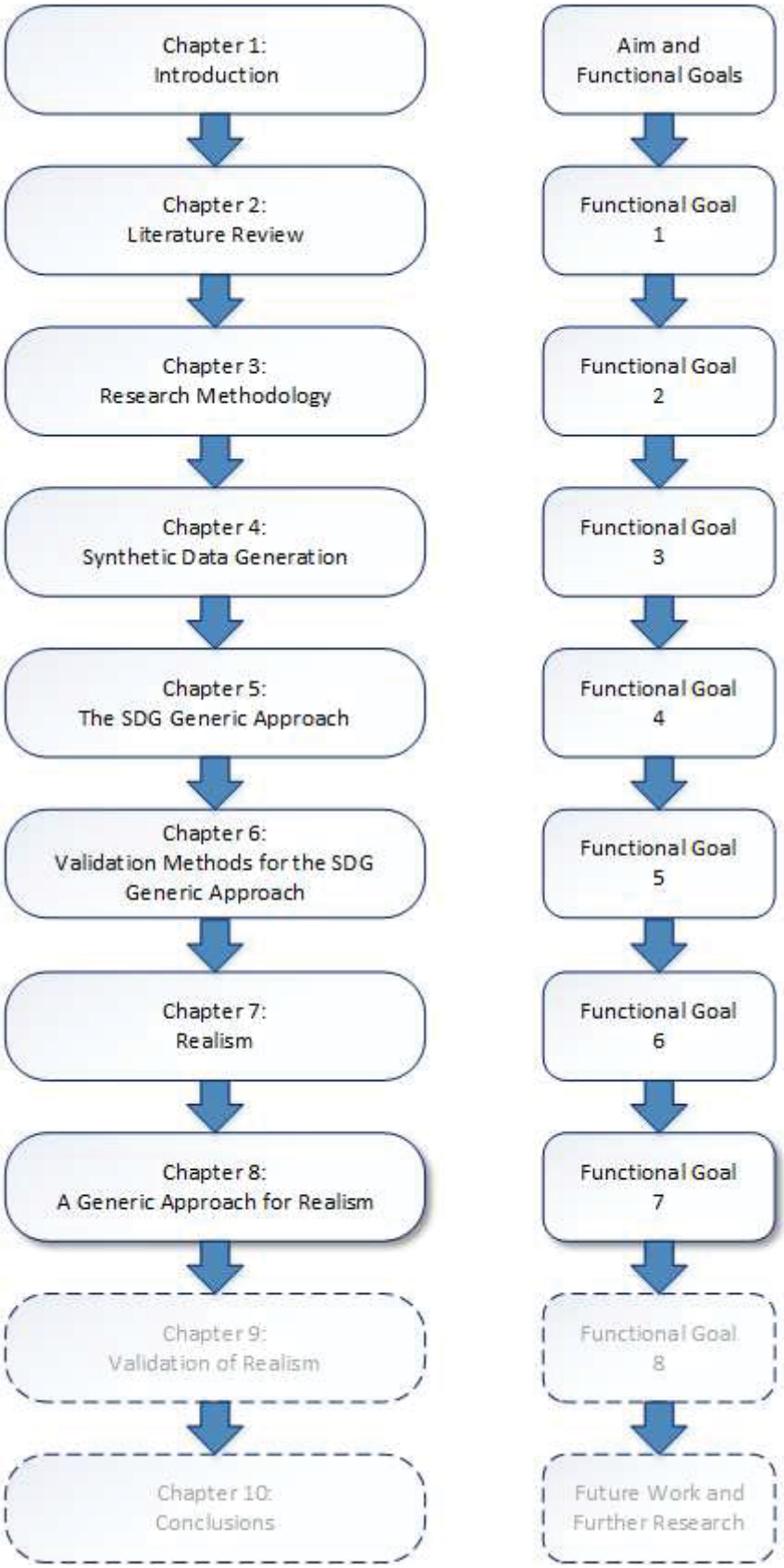
Readers may wonder whether discussion of validation methods, and indeed validation of realism in synthetic data is truly important. If proper definition, validation and documentation are all key to following and respecting the scientific method then this thesis would assert that the inclusion of a definition and validation approach for realism is both important and necessary to any published SDG method. Commitment to a complete and robust scientific method is essential for providing assurance that the expressed belief in a result is valid, repeatable and therefore of value to the scientific community at large.

It is easy to claim successful results in any research endeavour, especially if the authors are evaluating performance against imperfect or incomplete objectives (Voss et al, 2011). It has been suggested that computer science should not be considered as a true science, but more correctly an engineering discipline (Brooks, 1996); a field set apart from that of the natural sciences where phenomena is studied, because in computer science we manufacture computers and programs as creations and therefore the concepts of experimentation, testing and scientific method are misplaced (Tichy, 1997). This author agrees with Tichy (1997) when he exemplifies opinions such as that of Brooks' (1996) as an extremely narrow view of the domain of computer science. Computer science encompasses a far broader range of topics including the information sciences and related fields that go well beyond the narrow topic of computers (Tichy, 1997). The controlled method of inquiry more commonly known as the traditional scientific method should still be rigorously applied when research is conducted within the field of computer science (Nunamaker & Chen, 1990; Tichy, 1997; Zelkowitz et al, 1998). In order for future generation models to meet the basic tenets of the scientific method the identification, classification and presentation of adequately documented and verifiable characteristics is necessary. Necessary to the success of these SDG projects, especially when a project seeks to claim realism as a component of the synthetic data that is produced.

7.5 Conclusion

This research has seen that researchers frequently identify realism as a necessary quality for synthetic data to possess without defining realism or describing the elements that should be considered as realistic. This chapter began by discussing how the term realism has been used in the literature, and characterising how the term can be understood. If concluded by presenting a two-level framework that researchers can use to define realism as it applies directly to their own SDG project. The two levels consist of a) the obvious or extrinsic elements such as formats, structures and statistics, and; b) the inherent, embedded or intrinsic knowledge that can be extracted by understanding the relationships and rules from within. While many authors regularly describe the extrinsic knowledge without applying it to the realism context, too few describe go into the sufficient depth necessary to providing the reader an understanding of the intrinsic. A complete description of both the extrinsic and intrinsic knowledge can improve the generation approach, provides the basis for synthetic data validation and provides utility to the SDG approach by allowing others to repeat and validate the claimed results.

This page intentionally left blank


Chapter 1: Introduction
Chapter 2: Literature Review
Chapter 3: Research Methodology
Chapter 4: Synthetic Data Generation
Chapter 5: The SDG Generic Approach
Chapter 6: Validation Methods for the SDG Generic Approach
Chapter 7: Realism
Chapter 8: A Generic Approach for Realism
Chapter 9: Validation of Realism
Chapter 10: Conclusions
Aim and Functional Goals
Functional Goal 1
Functional Goal 2
Functional Goal 3
Functional Goal 4
Functional Goal 5
Functional Goal 6
Functional Goal 7
Functional Goal 8
Future Work and Further Research

8. RA: A Generic Approach for Realism

This chapter presents RA, a new systematic approach that is used to discover realistic elements, characteristic knowledge and rules to be used in synthetic data generation and facilitate claims of *realism*.

This chapter sets out to achieve the requirements of functional goal 7:

Functional Goal 7. Develop a generic approach to identifying and characterising realism for SDG

This chapter is structured as follows:

8.1 Introduction

This chapter presents a new structured *knowledge discovery in databases* approach built from an amalgam of the works of Fayyad et al (1996), Fernandez-Arteaga et al (2016), Holzinger et al (2014) and Mitra et al (2002); with new further enhancements that define and document the realistic properties that may be inherent to a given store of observed data. This combination relies on integrating a selection of established methods and algorithms to mine, describe and classify data. Each method is selected based on a solid foundation of tested and reviewed literature demonstrating its efficacious use in realising the unseen properties of datasets. Together they present a powerful and practical toolset for identifying structure and realistic properties.

8.2 Identifying Realistic Elements from the Real Data

One simple truth exists in information science: the amount of accessible data we possess is fast outstripping the knowledge we have gleaned from that data (Han et al, 1993; Mitra et al, 2002). Rarely is raw data of any direct benefit, especially to the SDG process (Mitra et al, 2002). We can acquire some knowledge from the sum of the entire dataset, but more still by extracting and illuminating relationships that exist between independent data fields within the dataset (Granger et al, 1993; Han et al, 1993). Together, the overall data along with its relational and other metadata can be used to develop knowledge rules that can either be *quantitative*, or where the rule does not relate to or associate quantitative aspects, *qualitative* (Han et al, 1993; Sobh & Perry, 2006).

A variety of theories and methods exist for drawing out meaningful knowledge from a collection or collections of raw data. Some authors discuss this knowledge extraction process as one of learning, in which concept hierarchies and realised new attributes are obtained from the data and assessed for relevance, to be used in the creation of a set of quantitative rules made up of a combination of characteristic rules and classification rules (Han et al, 1993). Alternative approaches focus on *data mining* (Castellani et al, 2008) while others consider data mining to be just one step in the much larger process of *Knowledge Discovery in Databases*, or KDD (Holzinger et al, 2014; Mitra et al, 2002; Prather et al, 1997). KDD uses such things as machine learning and neural networks to gather insight; to search for relationships and patterns that are essential to, but hidden in, the observed dataset (Holzinger et al, 2014; Mitra et al, 2002; Prather et al, 1997). In each case the key goal is finding meaningful patterns in otherwise large and complex datasets (Castellani et al, 2008; Prather et al, 1997).

8.3 Differentiating the Substance of Data

The synthetic data we generate can itself be either quantitative or qualitative. Quantitative when we seek statistical truths from the synthetic data to test theories and predictive models (Sobh & Perry, 2006) such as the forecasting SDG seen in Wan et al (2008). Qualitative when we seek truth from the words, meanings and categories to be seen in the synthetic data (Sobh & Perry, 2006) such as the data scrambling method seen in Mouza et al (2010) or the signal and noise models of Barse et al (2003) and Killhourhy et al (2007). SDG models can also generate data utilising both the quantitative and qualitative approaches at the same time, or cumulatively, in the sum of their output. This is especially true when the model seeks both statistical and contextual applications. Such approaches are exemplified in the model of McLachlan et al (2016) where the authors sought to generate synthetic patient records using publically available incidence and treatment statistics to model treatment pathways with the application of contextually appropriate care notes as each nodal point in the care pathway was illuminated.

Identifying whether their SDG model is qualitative, quantitative or requiring some aspects of both is a necessary step for the synthetic data developer; it can play a significant role in deciding the factors and characteristics that are necessary to create realism in the model. Collectively the method paradigm incorporated with those factors and characteristics that have been identified provides the basis for

developing a structure for validating the success or failure of the SDG model. More importantly for those that have identified realism as a requirement, they provide the starting point to verifying realism in the resulting dataset.

8.3.1 Quantitative Characteristics

The real or observed data may in itself be statistical, and therefore quantitative, such as one might see in collected meteorological data. Even if it is not it may still be possible to draw significant statistical information, for example; consider an instance where the researcher has been presented with a spreadsheet of people who voted at a selection of polling booths and is asked to generate a synthetic version for release, as the real data cannot be made public for privacy and confidentiality reasons. On the surface this may appear to be qualitative data however it would be possible to draw a number of statistical representations from it such as: how many people of each genealogical nationality voted in each hour and what percentage of them were male, what percentage of the total population of the area as reported in the census voted in each booth and so on. Such statistical observations, once identified and documented, represent the quantitative characteristics of the real data and should be sought in the data output by any synthetic model claiming to be realistic.

8.3.2 Qualitative Characteristics

The qualitative characteristics of the real or observational dataset should be identified and documented for any SDG project, but especially for those projects that seek to create realistic synthetic data. The first step would be to identify and qualify the database schema. The database schema, or structure, describes how the data is organised (Nijssen & Halpin, 1989). In the relational database example this includes expression of the tables, the fields within those tables, constraints such as those identifying the primary key or limiting field values along with any referential integrity constraints, or foreign keys (Nijssen & Halpin, 1989).

8.4 Knowledge Discovery in Databases (KDD)

While traditional methods of data mining often involved a manual process of scouring through databases in search of previously unknown and potentially useful information, these processes can be slow and an inefficient use of time (Fayyad et al, 1996; Holzinger et al, 2014; Mitra et al, 2002). Modern approaches, where the human is accentuated by machine learning and neural network algorithms, are considered a more expedient way of realising insights from today's extremely large datasets (Fayyad et al, 1996; Holzinger et al, 2014; Mitra et al, 2002).

The KDD process follows a logical progression of steps, with a number of the cited authors (Fayyad et al, 1996; Fernandez-Arteaga et al, 2016; Holzinger et al, 2014; Mitra et al, 2002) agreeing on the basic structured approach shown in Figure 15:

Step 1: Develop and document

Develop and document: a) an understanding of the application domain, b) relevant prior knowledge and, c) the goal of the KDD process.

Step 2:

Composed of two sub-steps: First, collect together the raw data by selecting relevant datasets, or focusing specifically on a subset of data from which discovery is to be performed. Second, extract or select the target data to be used in later steps of the KDD process from the pool of RAW data.

Step 3:

Cleansing and pre-processing functions are performed on the data to eliminate noise by removing rows containing such issues as incomplete or inconsistent data from the target data pool.

Step 4:

Transforming the data can also consist of multiple sub-steps including *integration*, *projection* and *reduction*. The i*ntegration* of data from multiple sources into a single consistent dataset, *projection* of the data by finding or identifying useful features used to represent data and *reduction* of the number of variables in the dataset down to those that are necessary for consideration.

Step 5:

This step requires matching the KDD process to a particular data mining method. These can include summarisation, classification, regression, clustering, web mining and so on (as described in Fayyad et al, 1996). This enables the selection and execution of a data mining algorithm that will be used to search for patterns within the target data.

Step 6:

The mined patterns are then interpreted and evaluated through the use of analysis using visualisation or other methods. This can be done automatically within the KDD process or with the direct intervention of the human to identify the truly interesting or useful patterns.

Step 7:

The discovered knowledge is then available for use.

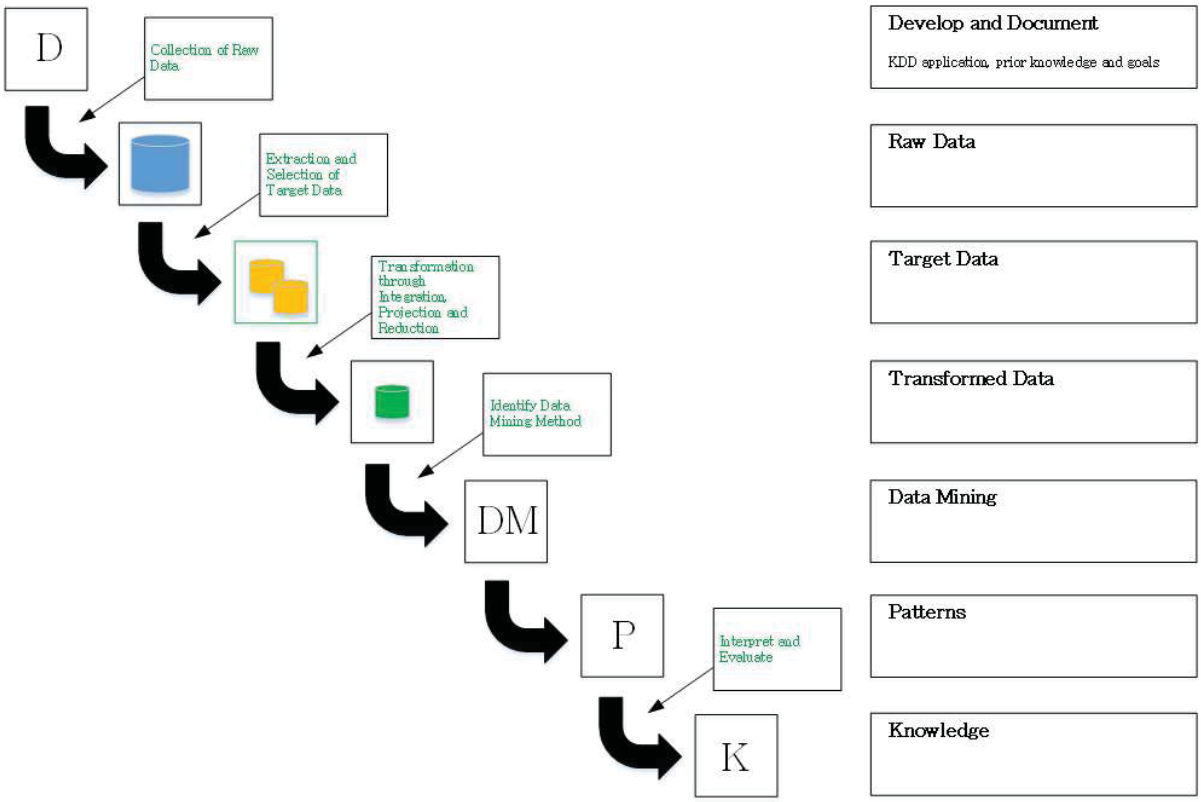


Figure 15: The KDD Process

8.4.1 HCI-KDD

As KDD has demonstrated its application, utility and computational prowess to improve on the generally basic manual techniques of data mining, a number of authors have built upon and extended Fayyad et al's (1996) original definition. A number of these extensions would be relevant to and support efforts to define, develop and assess the realism component of many of the SDG methods tabled in this research.

Holzinger et al (2014) focused their KDD work on problem-solving multiple issues with data mining in the bioinformatics realm. They identified that the standard computer-borne KDD process added knowledge value more from the very physical side of data, sometimes missing the more cognitive and necessary human functions of interaction, communication and sense-making (Holzinger et al, 2014). They propose an extended approach described as Human-Computer Interaction for Knowledge Discovery in Databases, or HCI-KDD (Holzinger et al, 2014). The HCI-KDD approach suggests that involvement of one or more domain experts may result in the identification and comprehension of additional qualitative patterns and information that the standard KDD machine learning approaches wouldn't find (Holzinger et al, 2014).

8.5 RA: The Enhanced KDD Approach

To ancient Egyptians the sun, Ra, represented light, warmth and growth. As a deity, it was believed that Ra was the king of gods and personal patron to the pharaohs. The RA approach prescribes use of the established Knowledge Discovery in Database (KDD) processes described in Section 8.4, updated and extended through the application of Human Computer Interactive KDD, or HCI-KDD principles. Step 5 of the KDD process shown in Figure 15 is enhanced through the application of additional knowledge discovery methods; the realisation of Concept Hierarchies (CH), Formal Concept Analysis (FCA) and the identification of Characteristic and Classification rules that can be used to describe the data. Each element of the enhanced KDD method is explained in the following sections.

8.5.1 Concept Hierarchies

Han et al (1993) recommend the deduction of attribute-oriented quantitative rules from large and very large databases through gaining an understanding of the concept hierarchies, data relevance and expected rule forms contained within the dataset. The concept hierarchy allows a synthetic data generation researcher to infer general rules from a dataset by analysing extracted and hierarchically arranged trees of relevant terms and phrases (Han et al, 1993; Sanderson et al, 1999). Developing a concept hierarchy involves organising levels of concepts identified within the data into a taxonomy, reducing down candidate rules to formulas with a particular vocabulary (Han et al, 1993).

Concept Hierarchy information is organised according to general-to-specific ordering, with the most general being the null ANY and the most specific corresponding to the specific values of the attribute in the database. The hierarchical relationship can be represented schematically, rising from the most specific (at the bottom) to the most general (at the top), such as "*city* **in** *state* **in** *country*".

8.5.1.1 Extending Concept Hierarchies with Statistics

One clear example this research proposes is the extension of concept hierarchy trees to include occurrence statistics. Once the concept hierarchy tree is identified, another pass across the source data should occur to summate the occurrence of each of the specific terms.

The extended concept hierarchy provides for a set of statistical tables that could be used to constrain the generation model, and a method for accurately assessing and validating the resulting synthetic data. The concept hierarchy could just as easily be about the relationships between data as it is about the value of attributes in the dataset itself. It can even be used to represent attributes and terms observed in documents and raw text (Sanderson et al, 1999) and offers a very powerful way of seeing into and representing the data while ensuring we do not actually need to rely on, or risk exposing, any of the actual source data in the SDG process.

8.5.2 Formal Concept Analysis

Formal Concept Analysis (FCA) is another well-developed approach utilised in data mining and KDD (Stumme et al, 2002; Willie, 1992). FCA has been used successfully as a knowledge discovery and representation method in a wide range of fields not limited to medicine, psychology, social and informational sciences and civil engineering (Stumme et al, 2002).

FCA starts with a formal context delivered via a triple, where an *object* {G} and *attribute* {M} are shown with their *incidence* or *relationship* {I} (Stumme et al, 2002; Willie, 1992). A context table (shown later in Table 12) is created that identifies with an *X* instances where a relationship exists between an object and its corresponding attribute/s.

The creation of concepts, which are interrelated collections of contexts, comes from review of the context table. For example; one might seek to identify the smallest or largest concept containing one of the objects contextualised with or without a particular attribute. These concepts can then be represented as rules. Examples of these might be:

> The smallest concept containing Object *x* may be with {Object *b*}
> The largest concept containing Object *y* may be with {Objects *l, b, d*} and attribute {Attribute *f*}

The second step in the FCA approach involves the creation of a concept lattice (shown later in figure 19). The simplest way to consider a concept lattice is as a mapping of the formal context or intersections of objects and attributes. The concept lattice visually delivers the powerset structure, mathematically described as $\wp${G,M,I} (Willie, 1992). It allows the researcher to easily identify not only sets of objects with common attributes, but also the order of specialisation of objects with respect to their attributes (Rodriguez-Jimenez, Cordero, Enciso & Rudolph, 2016).

Whereas such a lattice demonstrates the intersections of *incidence* between objects and attributes, more recent research has demonstrated that the lattice can also be used successfully to demonstrate the incidence of objects *not* possessing of an attribute (Rodriguez-Jimenez et al, 2016). The absence of something, or negative information, can sometimes be just as significant as confirmation of its presence (Rodriguez-Jimenez et al, 2016). The resulting lattice is described as the *negative information concept lattice* (Rodriguez-Jimenez et al, 2016).

8.5.3 Characteristic and Classification Rules

Han et al (1993) propose a set of strategies that can be used to learn characteristic and classification rules from within a dataset. These rules can act as constraints during generation, and tools to compare against the resulting synthetic data to validate its accuracy and realism.

8.5.3.1 Characteristic Rules

Firstly; data relevant to the learning process is collected. All non-primitive data should be mapped to the primitive data using the concept hierarchy trees (e.g. *Vancouver* would be mapped to *City*, *Professor* would map to *Position*, *Applied Sciences* would map to *Faculty Department* and so on).

Second; Generalisation should be performed on the smallest decomposable components to minimise the number of concepts and attributes to only those necessary for the rule we endeavour to create. Han et al (1993) give an example seeking to generalise the attributes of a professor. In this example they identify that the *Name* attribute is too general and not characteristic to a generalised set of data from which we could make rules about the concept *Professor*; that is, the name of each individual professor does nothing to further enhance, categorise or describe the general population of professors. They further go on to identify that in some cases it is sufficient to substitute a lower level concept (say the city or suburb where the professor was born) with its corresponding higher level value, the state or country. In another example they identify that where an attribute value is too specific, resulting in a large number of distinct values, replacing this value with a higher level general term should be performed. As such, a specific salary could be substituted with the identifier for the pay scale range that the salary amount was drawn from, or could be ranked into a range spread corresponding to the terms; *low, medium* and *high*.

The third step in the rule creation process transforms the final generalisation into a logical formula that recognises rules within the data. In this way we identify a characteristic rule that can operate as a constraint during the generation phase and later as a validation tool to verify both the generation model and realism in the synthetic dataset.

8.5.3.2 Classification Rules

Han et al (1993) provide a classification knowledge discovery process that discriminates the concepts of a target class from those of a contrasting class. This method provides weightings for the occurrence of a given set of attributes for the target class in the source dataset, and accounts for occurrences of attributes that overlap, or apply to both the target and contrasting class. To develop a classification rule, first the classes to be contrasted, their attributes and relevant data must be identified. Attributes that overlap form part of the generalisation portion of the target class only. Collections of attributes that are specific only to the target class become the basis for the classification rule.

8.5.3.3 Summary of Characterisation and Classification Rules

In summary; characterisation rules describe reduced collections of generalised attributes for a class occurring together in the dataset; where for any query of the dataset specifying *n-1* attributes from the rule, the remaining attribute is the only one that can be true. Classification rules describe those specific collections of attributes that differentiate one class from one or more remaining classes; where the target class is the only response for a query against the dataset specifying all of the attributes defined in the rule.

8.6 Case Study: Validation of the RA Approach

The following applies the theoretical approach of each KDD extension to the practical domain of midwifery. This case study utilises an anonymised patient dataset that was made available publicly to researchers and other interested parties by the Australian Department of Health during 2016 as well as recorded aspects from the Clinical Practice Guidelines (CPGs) and clinician input utilised in the realisation of the CoMSER SDG method (McLachlan et al, 2016).

8.6.1 Quantitative Aspects

In the CoMSER example the authors describe the quantitative aspects as coming from the New Zealand Ministry of Health (MoH) published dataset of birth statistics (McLachlan et al, 2016). The source data consists of multiple excel spreadsheet tables describing pregnancies, births, medical interventions and outcomes that were recorded in each of New Zealand's District Health Board (DHB) areas (Ministry of Health, 2014). An example of the quantitative aspects taken from CoMSER are shown in Tables 9 and 10.

Table 9: Ethnicity Statistics for births at CMDHB in 2012 (expressed as percentages)

Ethnicity (%)	
European	22.24
Maori	25.13
Pacific Islander	34.30
Asian	16.14
Other	2.11
Not Stated	0.08

Table 10: Age Statistics for births at CMDHB in 2012 (expressed as percentages)

Age	
Under 20	8.26
20-24	22.93
25-29	26.74
30-34	23.96
35-39	14.58
40 and Over	3.53

8.6.2 Qualitative Aspects

Defining the qualitative aspects realises the tables, fields, constraints and relationships that either already exist in a real dataset that is to be emulated by synthetic data, or which is necessary to the requirements of the synthetic data being created. A schema example taken from the anonymised patient data released by the Australian Department of Health (DoH) in 2016 is shown in Figure 16 and Table 11.

8.6.3 Applying KDD

The first step of the KDD approach requires the development of an understanding and knowledge of the application domain and the goal of the KDD process. In this case we will use a domain familiar to this author, that of electronic health records (EHRs).

The second step necessitates the collection and reduction of the dataset to be investigated and eventually, recreated synthetically. The EHRs of those patients who had given birth were extracted from the larger dataset of anonymised EHRs from the Australian DoH. This pregnancy dataset was further reduced in width to only those attributes that related to the labour and birth event.

patient	(patientID#, title, lastName, firstName, dateOfBirth, gender, ethnicity, primaryLanguage)
inPatientUnit	(inPatientUnitID#, unitName, building, floor, maxBedsAlloc, activeBedsAlloc)
admissionEvent	(admissionID#, patientID, inPatientUnitID, bedNumber, admissionDate, releaseDate)

Figure 16: Midwifery Patient Database Relational Schema extract

Table 11: Midwifery Patient Database Patient Relational Table Schema extract

patient		
PK	patientID	INT
	title	TEXT(10)
	lastName	TEXT(30)
	firstName	TEXT(30)
	dateOfBirth	DATETIME
	gender	CHAR(10)
	ethnicity	CHAR(20)
	primaryLanguage	VARCHAR(100)

To cleanse the dataset any patient whose labour and birth records were incomplete or who had given birth en route to hospital were removed. For the sake of example; the number of variables, or attributes, was reduced to the delivery type, interventions and whether or not the patient had been administered two particular medications, steroids and intravenous antibiotics. This completed the third and fourth steps of the KDD process.

Step five, which is built around the enhanced and extended data mining components was then conducted on the resulting labour and birth dataset. The results of each data mining process follows.

8.6.4 Concept Hierarchy

An example concept hierarchy for Child Birth is shown in Figure 17. The general term *Child Birth* breaks down into the two modes by which birth occurs, *Caesarean* and *Vaginal*. As an example; Caesarean births break down even further into the two specific types that occur, the *elective* or requested/planned caesarean and the *emergency* caesarean. In this way we are moving from the most general concept at the top to the most specific at the bottom. This is extended in Figure 18 with the addition of quantitative statistics (in brackets) identified from the MoH source data.

8.6.5 Formal Concept Analysis

While the number of possible attributes in a midwifery patient records dataset is much larger, for this example the attribute set has been limited to those of labour and birth and two medications administered during those processes. For the purpose of example ten patients were selected at random and their intersection with these attributes has been populated into the FCA binary reference shown at Table 12.

The Formal Context data from Table 12 was entered into the into the GaLicia 3 open source application (Valtchev et al, 2003) resulting in the concept lattice in Figure 19. Patients are identified numerically on the left of each context and the attributes are shown alphabetically on the right. The upper-most node in red represents all of the patients, the lowest node represents all of the attributes (birth and labour types). The lattice accordingly displays contexts built of the lowest number of attributes with the largest number of patients at the top, through to those with the largest number of attributes and lowest number of patients at the bottom.

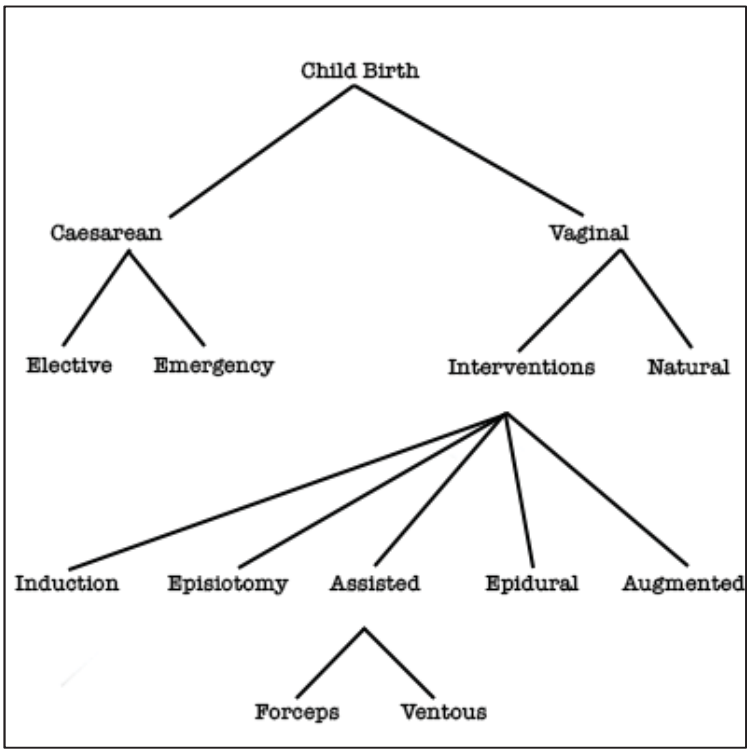


Figure 17: Concept Hierarchy for Child Birth

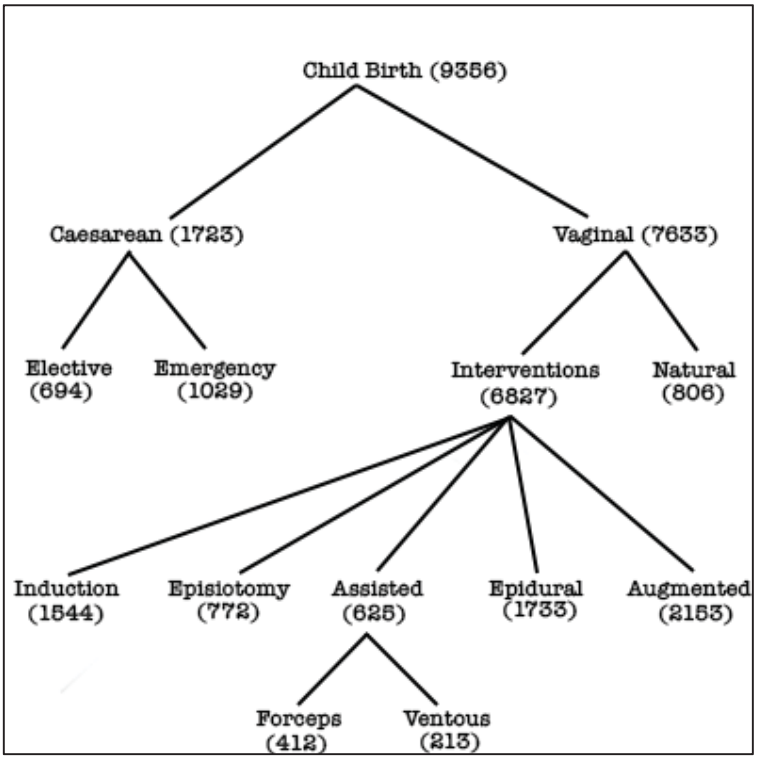


Figure 18: Concept Hierarchy for Child Birth with Statistics

Table 12: Formal Concept Analysis for 10 Random Labour and Birth Patients

			Birth Type						Labour Type				
	(a) Induction	(b) Epidural	(c) Natural	(d) Elective Caesarean	(e) Emergency Caesarean	(f) Forceps	(g) Ventous	(h) Episiotomy	(i) Premature Labour	(j) Term Labour	(k) Post Dates Labour	(l) Steroids	(m) IV Antibiotics
Patient 0	X	X				X		X			X		X
Patient 1		X	X					X		X			
Patient 2			X								X		
Patient 3	X	X			X						X		X
Patient 4			X					X	X			X	X
Patient 5		X		X									X
Patient 6	X	X			X						X		
Patient 7	X						X	X			X		X
Patient 8				X					X			X	X
Patient 9		X	X							X			

8.6.6 Characteristic Rule

A characteristic rule exemplar identified through analysis of the DoH labour and birth data provides the following four attributes for the target class “Patient”:

n1: **Pregnancy Status:** Low Risk

n2: **Gender:** Female

n3: **Pregnant:** Yes

n4: **Fetal Heart Monitoring:** Intermittent in Labour

The potential *n-1* scenarios are:

(a) A *pregnant*, *female* where the fetal heart monitoring is conducted on an *intermittent* basis;
   a. Can ONLY have the pregnancy status: *low risk*.

(b) A *low risk, pregnant* patient receiving *intermittent* fetal heart monitoring;
   a. Can ONLY be gender: *female*

(c) A *female* with a pregnancy status of *low risk* receiving *intermittent* fetal heart monitoring;
   a. Can ONLY have a pregnancy status: *pregnant*

(d) A *female* patient who is *pregnant* and *low risk*;
   a. Requires ONLY *intermittent* fetal heart monitoring

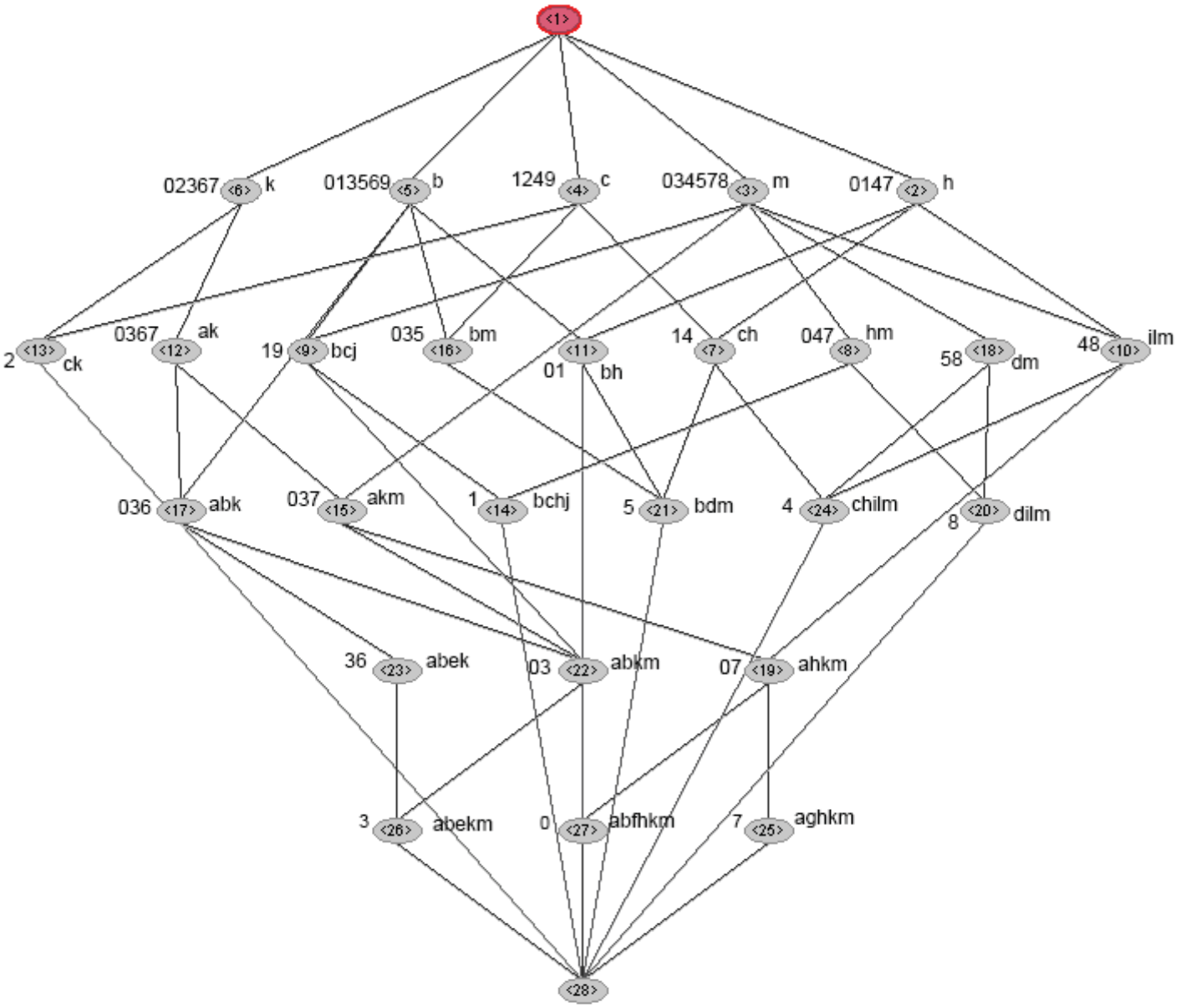


Figure 19: Concept Lattice example

Expressed as a conditional formula, this rule could be used in the post-SDG validation process to verify that all pregnant patients are female, and those receiving intermittent fetal monitoring are all of the low risk pregnancy status. The logical rule condition formula is shown in Figure 20.

$$\forall\chi\ \mathbf{(}\text{midwiferyPatient}(x) \rightarrow ((\mathit{Sex}(x) = \text{female})\ \wedge\ (\mathit{Pregnant}(x)\ \varepsilon\ \text{Yes})\ \wedge\ (\mathit{pregnancyStatus}(x)\ \varepsilon\ \text{Low Risk})\ \wedge\ (\mathit{fetalHeartMonitoring}(x)\ \varepsilon\ \text{Intermittent}))\mathbf{)}$$

Figure 20: Characteristic Rule from the domain of Midwifery

8.6.7 Classification Rule

A classification rule exemplar established through analysis of the DoH labour and birth patient dataset that demonstrates operation of a healthcare facility's CPG (Eastern Health, 2014) for medically planned or recommended mode of delivery based on a patient's previous mode/s of delivery is shown in Table 13.

Table 13: Generalised Relation Table

Mode of Delivery	Multip[1]	Primip[2]	Previous Delivery = CSect <2	Previous Delivery = CSect >2
Caesarean	No	yes	Yes	No
	Yes	No	Yes	No
	Yes	No	No	Yes
Vaginal Delivery	No	yes	Yes	No
	Yes	No	Yes	No

[1]Multiparous – a patient who has given birth multiple times
[2]Primiparous – a patient who is pregnant for the first time

The qualitative classification rule for Caesareans as a medically planned mode of delivery based on the mother's previous mode/s of delivery is extracted from the relational data in table 13 and shown below in Table 14:

Table 14: The qualitative classification rule for Caesarean based on previous mode/s of delivery

Multip	Primip	Previous Delivery = CSect <2	Previous Delivery = CSect >2
Yes	No	No	Yes

This disjunctive is represented as a logical rule. The rule identifies and weighs the attribute that creates the disjunct (*d*) as shown in Figure 21. The CPG and patient advice factsheet based on that CPG state that it was clinically indicated for all pregnant patients who have had two previous caesareans are to have a caesarean delivery for subsequent pregnancies due to significant risk of post-partum bleeding (Eastern Health, 2014). Analysis of the DoH patient records showed that patients whose medical history included two or more prior caesareans delivered subsequent babies by caesarean on medical advice.

The rule is effected through operation of all of the cumulative triggers; (1) that the previous pregnancies trigger is activated, as the first-time or *primiparous* mother is not the target of the CPG. (2) that the previous caesareans trigger is activated, in this case by two or more previous caesareans. If both triggers are established, then we can be assured from both the CPG and analysis of anonymised patient records for that medical district that her next pregnancy will in 100% of cases should be another caesarean.

When used as a constraint in the generation algorithm this rule would ensure that where a multiparous mother has had two previous caesareans, the probability of her next pregnancy being a caesarean is automatically updated to 100%.

$$\forall\chi\ (\text{modeOfDelivery(x)} \rightarrow ((Multip(\text{x}) = \text{Yes}) \wedge (Primip(\text{x}) \in \text{No}) \wedge (previousDelivery{=}CSect{<}2(\text{x}) \in \text{No}) \wedge (previousDelivery{=}CSect{>}{=}2(\text{x}) \in \text{Yes}[d{:}100\%])))$$

Figure 21: Classification Rule from the domain of Midwifery

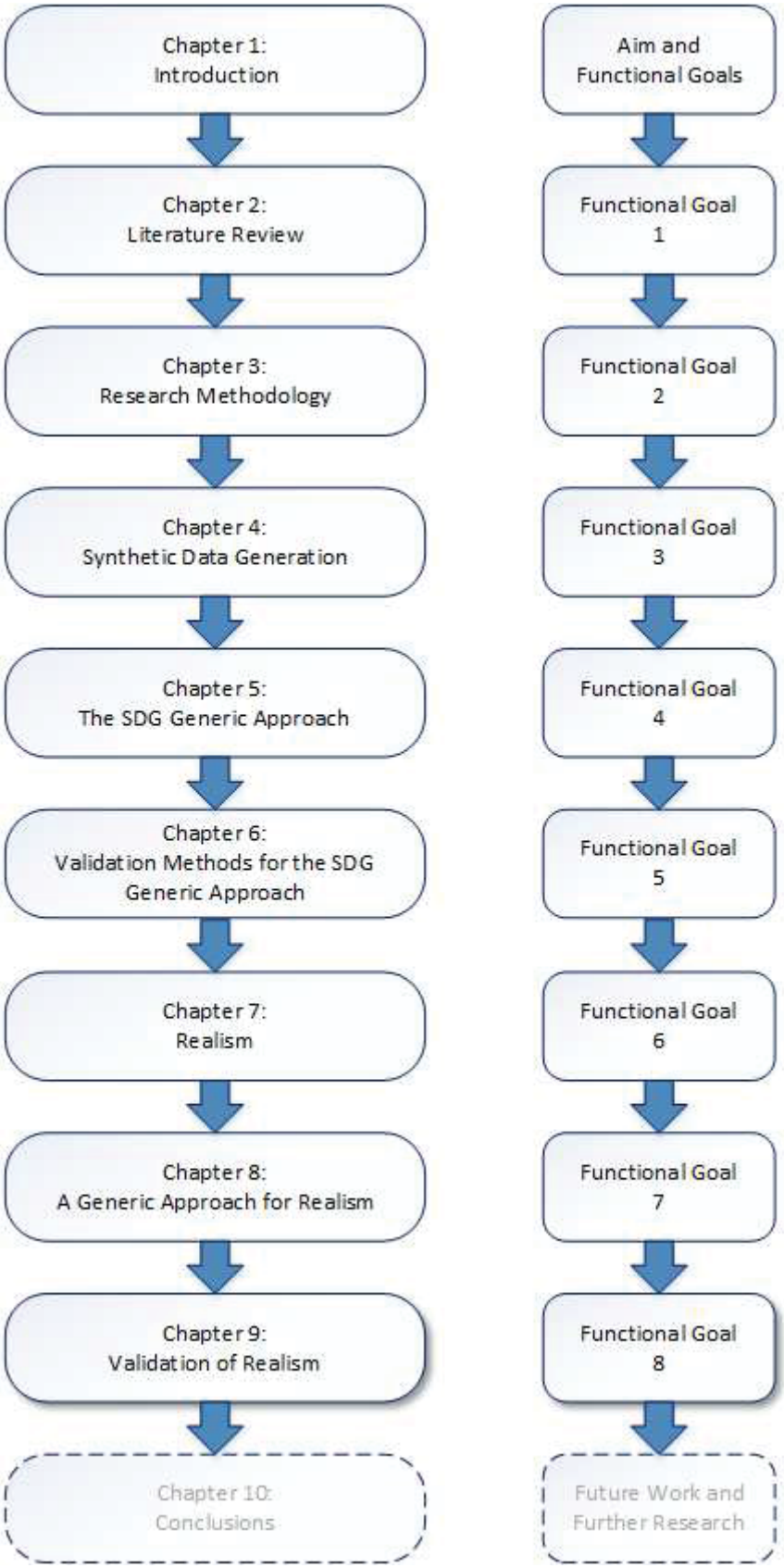
8.6.8 Case Study: Discussion

This case study has applied the extended HCI-KDD methodology to a published application of SDG that generates electronic healthcare records in the midwifery domain for the labour and birth event. Through description of each of the described elements, the case study investigates a comparatively simple SDG model that only utilises one statistical dataset, some published guidelines and a procedural flowchart known as a CareMap in the healthcare domain. It demonstrates that while obvious data exists on the surface, a wealth of knowledge is hidden below the surface that the methodology described in this chapter is able to expose. Having more knowledge about the inner aspects and characteristics of the data we seek to synthesise can only serve to improve the likeness, accuracy and realistic nature of the synthetic dataset.

8.7 Conclusion

It was discussed earlier that a large number authors claim some requirement for realism in their SDG methods (see Chapters 2 and 7). All of the SDG methods reviewed in this research claimed success, giving rise to the impression that unsuccessful methods are not published. We also saw that the vast majority of SDG methods provide little or no evidence of a structured approach to identifying and recording the elements of the dataset they seek to synthesise, beyond ensuring superficial comparability to the obvious structural and statistical elements (see chapters 4 and 7). One final note is that it should be reinforced here is that commencing any validation of an SDG method with anything less than the most complete set of highest quality knowledge available to the researcher will be hampered and unable to provide justification for the resulting synthetic data. Therefore, the use of a structured knowledge discovery methodology is possibly the only sound way of bringing together the seed knowledge required for any such validation process.

This chapter presented an enhanced and extended KDD method that can resolve both issues. The proposed method utilises a range of qualitative and quantitative observations followed by and incorporating HCI-KDD principles that are extended through the identification of concept hierarchies, formal concept analysis and systematic analysis that delivers characteristic and classification rules. The knowledge recorded from these efforts resolves the issue of ensuring we have abundant high quality information about the observed data that we seek to synthesise. In identifying and recording that framework of knowledge we have also provided a systematic approach to validating both the generation method used and the resulting synthetic data.


Chapter 1: Introduction
Chapter 2: Literature Review
Chapter 3: Research Methodology
Chapter 4: Synthetic Data Generation
Chapter 5: The SDG Generic Approach
Chapter 6: Validation Methods for the SDG Generic Approach
Chapter 7: Realism
Chapter 8: A Generic Approach for Realism
Chapter 9: Validation of Realism
Chapter 10: Conclusions
Aim and Functional Goals
Functional Goal 1
Functional Goal 2
Functional Goal 3
Functional Goal 4
Functional Goal 5
Functional Goal 6
Functional Goal 7
Functional Goal 8
Future Work and Further Research

9. The HORUS Approach to Validation of Realism

This chapter presents the HORUS approach to validating and justifying the existence of realism in synthetic data.

This chapter sets out to achieve the requirements for functional goal 8:

Functional Goal 8. Develop and integrate the additional processes to validate realism in SDG.

This chapter is structured as follows:

9.1 Introduction

The presence of realism should only be asserted if it is verified (Penduff et al, 2006; Putnam, 1997). The domain of science should always be concerned primarily with testing; the validation and justification of any claim (Gallagher, Ritter, Satava, 2003; Haig, 1995). Validation is necessary to ensure the synthetic data we create is not skewed. This is why the ability to validate should be built in from the beginning as many models that use synthetic data may become unreliable in the company of skewed data (Gao et al, 2007). It has been argued that self-validation of your own methodology is meaningless (Forer, 1949), moreover it is observed, and now confirmed by this research, that very few published models are validated (Barlas, 1996; Carley, 1996). In Appendix B we find many SDG models that claim success in the absence of a rigorous method of scientific validation. Some form of validation would be absolutely necessary to support claims for realism in synthetic data (Penduff et al, 2006).

If the reader finds no documented evidence demonstrating reliability of the SDG model then the validity of the approach must be questioned (Moss, 1994). More than one of the SDG models reviewed discussed the fact that their generator could be or was run through many permutations or tweaks with each rendering different synthetic data, however the authors only discuss and display the outcome of one such operation (see Efstriadis et al, 2014; Gafurov et al, 2015; Ngoko et al, 2014). This gives rise to the question of whether the operation presented represents the only generation pass to deliver synthetic data remotely close to the real observations being modelled.

Each of the four CM validation approaches discussed in chapter 2 at section 2.6 is used to confirm a CM model's relationship to observed data, and therefore, it could be argued each detects some degree of realism. At the very least, these CM validation approaches represent a legitimate starting point for the discussion of validating realism in SDG models. The CM validation methods demonstrate a largely quantitative approach, or at best only possess some minor qualitative properties. The common tendency observed in the literature has been to measure some number of statistical properties in the synthetic data and draw comparisons to the same in real datasets. When validating realism in numerical or forecasting models quantitative methods are highly regarded. However, while statistical approaches may be of benefit to those models they should not be relied upon solely or to the complete exclusion of qualitative assessments of the interactive metrics, structure and characteristics that should have been identified from the real dataset (Penduff et al, 2006).

9.2 Application of the Horus approach

One of ancient Egypt's oldest and first national gods, Horus was revered as the god of the sky; that which contains both the sun and the moon (Ludwig, 2016; Porter, 2011). In the same way, the Horus approach to realism validation draws on both the THOTH enhanced generic SDG and Ra enhanced KDD approaches, effectively containing the sun and moon in order to validate for the presence of realism in synthetic data.

The validation approach occupies five steps; (1) Input validation; (2) Realism validation of the Input data; (3) Method validation; (4) Output validation, and finally; (5) Realism validation of the output data. Each analyses separate elements of the SDG process. These steps and how they fit within the SDG structure are identified in green in Figure 22. Used collectively, the five steps provide the information necessary for grounded confirmation of whether synthetic data is consistent with and compares realistically with the information seen in the observed data that the SDG model is seeking to emulate.

9.2.1 Input Validation

The process of input validation is a precautionary step. Similar to the five "rights" of medication administration: right patient, drug, dose, route and time (Koppel et al, 2008), input validation seeks to ensure the right data with right rows, is aggregated in the right manner and presented in the right format to the right constraint model. The input validation process should alert the user to any situation where input information has been prepared or presented incorrectly; be it an incomplete data row, incorrectly entered, transformed or applied statistic, or an inapplicable constraint (Lertpalangsunti et al, 1999).

While each stage of the HCI-KDD approach in chapter 8 creates knowledge, the input validation step concerns itself only with that knowledge presented in the form of data tables and statistics that are being used to seed or guide the generation process. The input validation process verifies each item, confirming that correct and sound input data is being presented to the generator, thus ensuring smooth

operation of the automated synthesis process (Bex et al, 2006). Input validation addresses many robustness issues before they can corrupt the SDG process (Laranjeiro, Vieira & Madeira, 2009).

9.2.2 Realism Validation 1

The first realism validation process seeks to verify all of the concepts and rules derived from the information used in the HCI-KDD process, as well as the statistical knowledge that may have been applied to these. It reviews and tests the premise and accuracy of each rule to ensure consistency with the intention of any guidelines used in their creation (such as Clinical Practice Guidelines in healthcare) or where available, tests them in real circumstances to ensure they are not rendered nugatory through interaction with observational data. Where any item of knowledge is found to be at issue, the researcher is returned to the knowledge gathering phase as shown in Figure 22.

9.2.3 Method Validation

Method validation requires the researcher to review the efforts of others both inside and outside of the given research domain. Attention would be paid to identify what methodological approaches are common for the researcher's domain, as well as what methods other domains have used for similar types of SDG. Assessment of this collected research would be performed to ensure that the chosen method is appropriate for the proposed solution, and capable of delivering the required synthetic data. Validation is not a search for absolute truth, more correctly and in this instance it is a search to establish legitimacy (Oreskes, Shrader-Frechette & Belitz, 1994). Table 15 provides the six key questions that should be asked of the SDG methodology or algorithm to be used.

Table 15: Realism Validation Questions

Validation Type	Validation Focus
Conceptual	Does the theoretical model adequately represent the real world?
Internal	Is the algorithm and computer code that employs it free from error?
External	Does the algorithm and computer code adequately and accurately represent the real world?
Alignment	How does this model's output compare to that of other models?
Data	How does the synthetic data compare to real observed data?
Security	Have there been any undocumented changes or manipulations to the model or code that may contribute to or alter the results?

The algorithm or mathematical components within a computer program are verifiable (Oreskes et al, 1994). Method validation also requires verifying that the algorithm to be used has been faithfully and correctly constructed and is complete in its execution.

9.2.4 Output Validation

Output validation seeks to both validate the structures in the output data as well as verify its basic statistical content. This step demonstrates the difference between the terms *validation* and *verification*. Oreskes et al (1994) describe the act of validation as ensuring the model is free from known or detectable flaw and internally consistent, while it is verification that seeks to establish that the output or predictions of the SDG model are consistent with observational data. Output validation ensures that the synthetically generated data conforms to the qualitative and quantitative aspects derived and defined for the generation project during the knowledge discovery phase.

9.2.5 Realism Validation 2

The second realism validation process performs all of the same tests as the first except that in this case the tests are performed against the newly created synthetic dataset to ensure it is consistent with the knowledge (rules, constraints and concepts) that were used in its creation. While some may feel that the statistical similarity verified during the output validation process is sufficient to proceed to using the synthetic data, the second realism validation step is arguably the most important for establishing and justifying any claim that the synthetic data is a proper substitute for the real data it was created to replace.

9.2.6 Validation: Discussion

The five-step validation approach presented in this chapter draws upon best practice validation knowledge identified from computational modelling, extending it into a complete approach that verifies and validates the input, method and output, and more importantly, establishes whether realism has been achieved. Validation of the input phase delivers a baseline, establishing requirements for the synthetic data and the elements available for its creation. This baseline has been used both to feed the generation process, and now verify that the synthetic data delivered is not just superficially coherent to real data, but demonstrates consistency through every concept and relationship. To a level that if verification is successful, can only mean realism has been achieved.

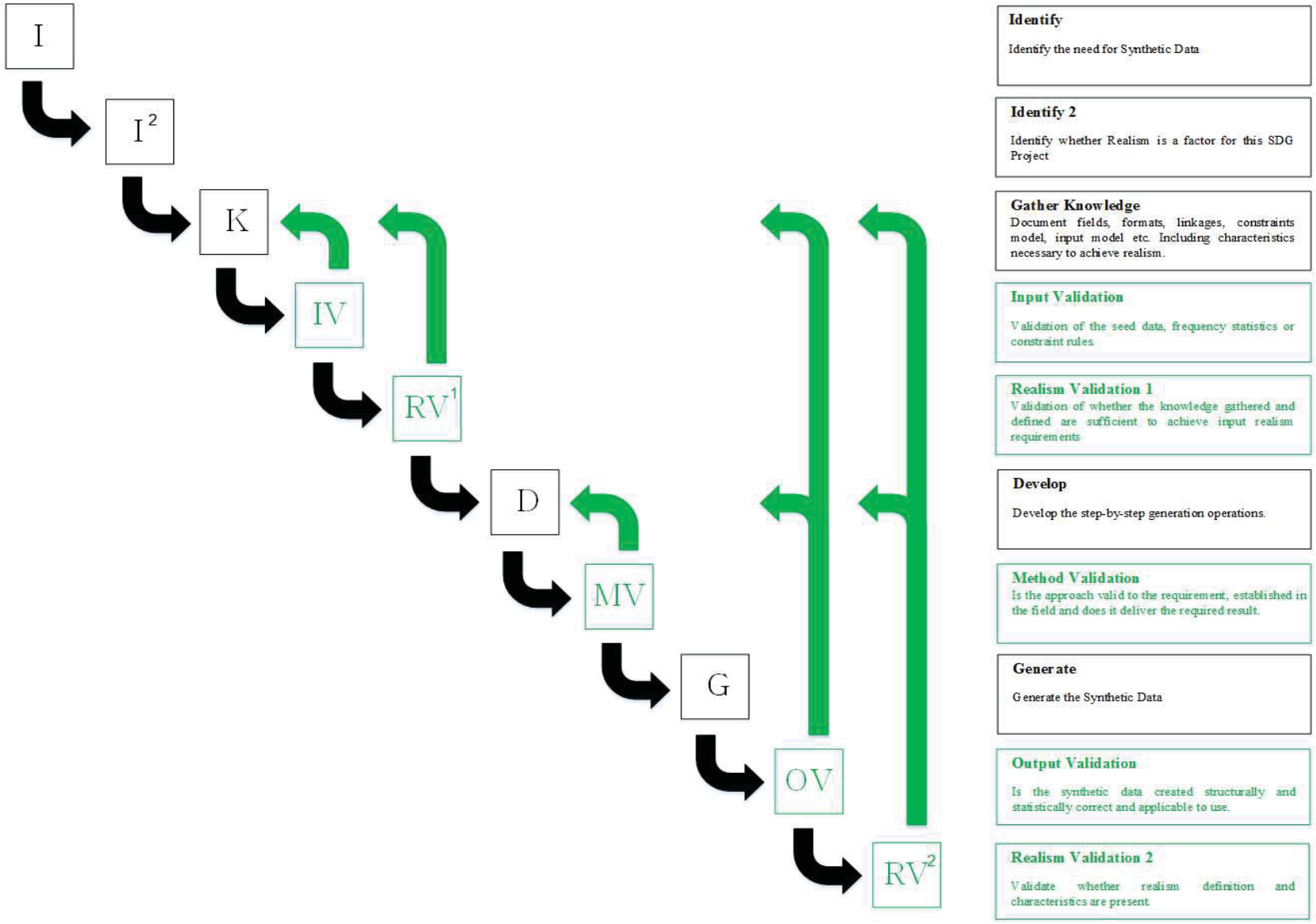


Figure 22: The HORUS approach embedded into THOTH

9.3 Case Study: Application of the Validation Approach

The following case study presents a detailed application of the validation steps against the knowledge already gathered from the midwifery patient data and input data used for CoMSER. The input validation step ensures that all data being provided to the generation process is correct in type, format and structure. The first realism validation step analyses on the knowledge that has been induced and deduced, be it concepts, concept hierarchies, rules or the statistical relations that may have been realised for any of these. The method validation step ensures we use the most appropriate method for generation, and verifies that we have faithfully prepared the algorithm or application to render the desired synthetic data. The final two validation steps are conducted on the synthetic data that has been output by the generation method. The first of these is output validation, which works through the output data to ensure that the fields, format and structure are consistent with either the observation data being synthesised, or with the requirements as defined for the synthetic data prior to creation. The final step, the second of the realism validation processes, ensures correctness and consistency by testing and validating each component of the concept and rules knowledge against the output data. Once each of the five validation steps has been passed, the validation process is complete.

Successful validation through the use of this process provides something that is difficult to observe within many of the published methods reviewed and tabled in the appendices, a basis for demonstrating and claiming successful delivery of the realism element sought and often claimed as present in the synthesised data. If the written documentation is maintained and presented to the degree prescribed by the scientific method, it can provide a basis for repeatability of each and every step of the process used.

9.3.1 Input Validation

Given the requirements and claims of the CoMSER SDG model (McLachlan et al, 2016), input validation at a minimum would need to verify each of the items described in Table 16.

Validating that the statistical representations being presented to the generation model are accurate may also take the form of manual verification that those entered are correct when compared to the tables that have been generated through the operation of the knowledge gathering step discussed at subsection 8.8.1. If any issues are identified the researcher should return to the knowledge gathering phase to correct them, as shown in Figure 22.

Table 16: CoMSER Input Validation Case Study

Validation Item	Validation Element	Description
Health Incidence Statistics (HIS)	Publicly available HIS	The authors state the model is required to use publicly available HIS in order to ensure protection of individual patients privacy. Accordingly, it is necessary to their model to verify that all statistics being used or aggregated have been derived from publicly available sources
	Format and breakdown	Any synthesis project should ensure that the statistics presented to the SDG process are able to deliver synthetic data with the required structure. Given the authors' approach, consistently formatted statistics would be required for each node of their state transition machine (STM). The HIS would also need to provide linked data for the demographics statistics to accurately model each field of their demographics generation approach and have the synthetic patients be accurate to the CareMap STM model.
	Completeness	Incomplete data may result in errors or exceptions in the operation of the generation algorithm. Each row and column of the dataset should be verified complete and without gaps and the data occupying each field must be the expected data in the correct format.
CareMap	CPGs	The authors would need to validate that the Clinical Practice Guidelines from which each CareMap is derived are correct for the geographical area or health board being modelled. Some variation on the care processes and procedures will exist between different health departments, boards or hospitals. These differences would impact the CareMap model and resulting STM.
	Clinician Input	Whenever a CareMap is created or there is any alteration made to the CareMap or STM, the resulting model should be verified with one or more experienced and currently practicing clinicians in that healthcare domain and geographical area.
Qualitative Characteristics	Data Fields	Verification and validation of the output data specification to be used in the generation operation. This step ensures that the definition of field types, sizes and the format specification for each will deliver data that is either structurally comparable to the dataset being recreated synthetically, or can fulfil the requirements prescribed for the output dataset.

9.3.2 Realism Validation 1

Realism validation of the CoMSER model requires a combination of validation and verification. ***Firstly;*** validation of the elements and structures of the concept hierarchy in Figure 18 while ***secondly*** verifying the statistics applied (in brackets) from the publicly available labour and birth statistics. ***Thirdly;*** confirming all binary relationships from the FCA shown in Table 12 to ensure validity of the concept lattice constructed from the FCA table and shown in Figure 19. ***Lastly;*** validating the form of each characteristic and classification rule while ensuring application of all rules results in the expected response.

9.3.3 Method Validation

Method validation of the CoMSER model requires an understanding of the research that went into identifying ways of describing clinical knowledge and how this knowledge could be presented in a structured model that readily demonstrates how each step in a health treatment process relates to those that come before or after. The CareMap was identified as an established method that clinicians were already using to develop and display the steps in diagnosing, treating and managing a clinical health process. The CareMap was most often visually represented by clinical researchers in the form of a flow

diagram. When imbued with rules and knowledge that establish how the patient moves from one point in the flow diagram to the next, the CareMap can more correctly be described as a state transmission machine (STM) which is something that comes with established verifiable methods for representation in a computer model.

Consideration was then drawn to the method prescribed by CoMSER for resolving the path taken by each synthetic patient through the STM. Given that the authors sought to generate patients whose incidence and treatment records would be statistically similar to those of a real patient population, a weighted probability distribution random generation approach was developed. This approach relied on the Walker Alias Method which describes one Monte Carlo type approach to probability distribution modelling. Research into Monte Carlo methods resolves that the Alias method is considered more advantageous as it is easier to understand and brings greater speed and accuracy to traditional Monte Carlo approaches (Edwards, Rathkopf & Smidt, 1991; Salvat, Fernandez-Varea & Sempau, 2006).

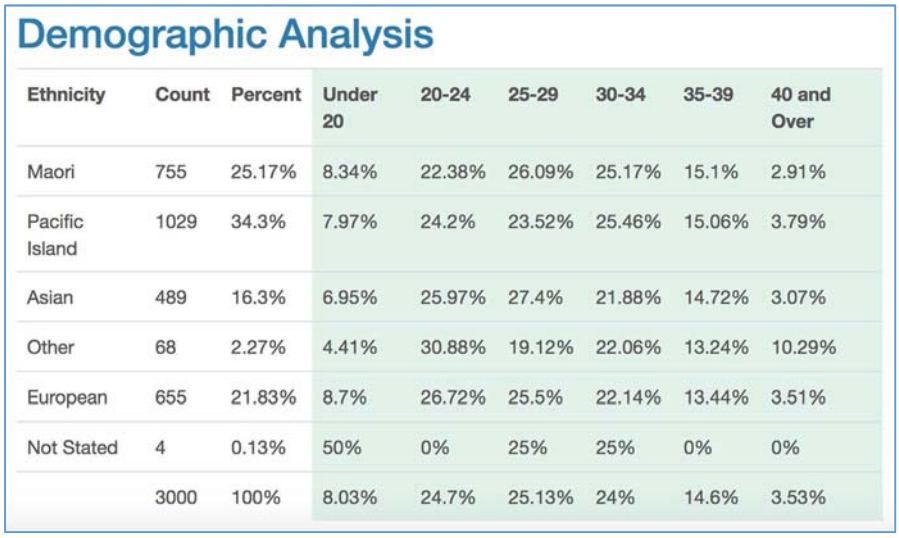
The method validation approach concludes that the use of CareMaps extended with descriptive rules and statistics, presented as State Transition Machines, and the Walker's Alias Method for probability distribution Monte Carlo modelling are appropriate given the available input knowledge, purpose and output data required of the CoMSER model.

9.3.4 Output Validation

The output phase for CoMSER begins with direct comparison of the generated data against that which was defined for the SDG project in order to validate the structure. In the case of CoMSER there was no use of real patient records and therefore no input dataset that was being emulated. CoMSER was required to generate patient records containing of fields for the patient's demographics and path through the Labour and Birth CareMap along with applicable treatment notes for each node and edge the patient encounters. These structures were present and valid.

In order to validate the statistical correctness within generated data an SDG operation of the CoMSER CoMENGINE application was conducted, generating patient records using input parameters built from the 2012 statistics for patients giving birth at Middlemore Hospital in Auckland, New Zealand (extracted from the larger public release of MoH birth statistics). The demographic statistics from a generation operation of the CoMSER CoMENGINE are shown in Table 17. Simple percentage comparison using the ethnicity and age input statistics from subsection 8.6.1 in Tables 9 and 10 is shown here in Tables 18 and 19. This demonstrates that CoMSER's SDG approach using the Walker Alias Method to produce a small (n=3000) collection of records is producing records that are within 0.4 percent of absolute accuracy to the publicly available health incidence statistics used in the input phase for ethnicity, and within 1.3 percent for age. These are considerably small error margins that testing has shown get closer to absolute 0 when much larger generation runs (n>10,000) are performed.

Table 17: Demographic Analysis Table from CoMSER CoMENGINE



Demographic Analysis

Ethnicity	Count	Percent	Under 20	20-24	25-29	30-34	35-39	40 and Over
Maori	755	25.17%	8.34%	22.38%	26.09%	25.17%	15.1%	2.91%
Pacific Island	1029	34.3%	7.97%	24.2%	23.52%	25.46%	15.06%	3.79%
Asian	489	16.3%	6.95%	25.97%	27.4%	21.88%	14.72%	3.07%
Other	68	2.27%	4.41%	30.88%	19.12%	22.06%	13.24%	10.29%
European	655	21.83%	8.7%	26.72%	25.5%	22.14%	13.44%	3.51%
Not Stated	4	0.13%	50%	0%	25%	25%	0%	0%
	3000	100%	8.03%	24.7%	25.13%	24%	14.6%	3.53%

Table 18: Ethnicity Statistics Comparison

Ethnicity	Statistical %	CoMSER SDG
European	22.24	21.83
Maori	25.13	25.17
Pacific Islander	34.30	34.30
Asian	16.14	16.30
Other	02.11	2.27
Not Stated	00.08	0.13

Table 19: Age Statistics Comparison

Age Range	Statistical %	CoMSER SDG
Under 20	8.26	8.03
20-24	22.93	24.70
25-29	26.74	25.13
30-34	23.96	24.00
35-39	14.58	14.60
40 and Over	3.53	3.53

9.3.5 Realism Validation 2

While the previous Output phase has validated the general data structure and verified the statistical content, the second realism phase focuses on establishing that each of the knowledge elements defined for the model is consistently represented in the synthetic data. Each concept hierarchy is comparatively verified against the generated data, and where consistency has been maintained, the related concept lattice is generated. The concept lattice is validated through observation, ensuring that relationship intersections of the same patients and phenomena occur in a similar way in the synthetic data.

The more important step is that of rule validation. Where rules like shown in Figure 20 have established that a particular confluence of variables do not occur in an observed dataset, the same must be true of the generated data. Where a rule like that shown in Figure 21 reports that a particular knowledge element always occurs, that must also be true of the generated data. Other rules may identify the intersection of particular elements on a scale between the characterisation and classification examples shown in subsection 8.6, that is, falling somewhere between never occurring (0%) and always occurring (100%). In these cases, the particular element should occur with a similar percentage to that defined by the rule, with the authors defining a reasonable standard deviation for statistical consistency that is applicable for their generation method.

For example; if a patient was generated who had performed a vaginal birth in hospital with no interventions or extended patient care and she also had a history of three previous and recent caesarean surgeries, this patient would be breaching the rule in Figure 21. This rule is based on CPGs advising that any patient who has had two previous caesareans is at significant risk of uterine abruption and in such cases a further caesarean should be performed. The rule was verified against the anonymous midwifery dataset and found to be true in clinical observation. It was important to the validity and realism of the synthetic patient data to ensure successful validation of the rule against the 3000 CoMSER synthetic patients, which in this case was achieved. No synthetic patient who had endured two or more previous caesareans was generated as having performed an in-hospital vaginal delivery.

9.4 Discussion and Summary

This case study has discussed and demonstrated the operation of a synthetic data validation approach that has been enhanced to also validate for the element of realism as defined in Chapter 7 and identified through operation the HCI-KDD process described in Chapter 8. Rather than just focus on possibly superficial checks of the rendered synthetic data, this approach operates during the entire SDG project. It firstly validates and verifies the knowledge being fed to the generation mechanism, the method or algorithm that mechanism relies upon and therefore the researcher could delay performing the SDG operation until those checks have completed and delivered results that are to his or her satisfaction.

The input and first realism validation processes allowed the user to assess the more obvious elements of knowledge from the data structures and statistics, as well as dive deep into that embedded detail that

existed but was far less obvious; the concepts and rules. The obvious and less obvious elements of knowledge are collectively described as the *input knowledge*. At this point in the process the authors of the CoMSER model would have had a greater degree of confidence that their approach, algorithm and overall model can deliver the synthetic data required.

The output and realism validation phases that round out the approach presented in this chapter assessed the synthetic data for its degree of compliance to the input knowledge used to guide and constrain its creation. Claiming that the structural and statistical form of the synthetic data is similar or consistent with observation proves that the data is realistic and completely valid, as some authors in the appendices have done, is a declaration made on the most superficial of grounds. Assessment of the CoMSER model has shown that there are a lot of factors under the surface that should be presented as part of the overall picture to support these claims. Consider for example how invalid a claim of realism would be if a genetically male patient was found to have been generated in the labour and birth dataset. This might seem a preposterous example, but if either of the demonstrated rules in Figures 20 and 21 were found to have been violated the synthetic data and generation method would be called into question.

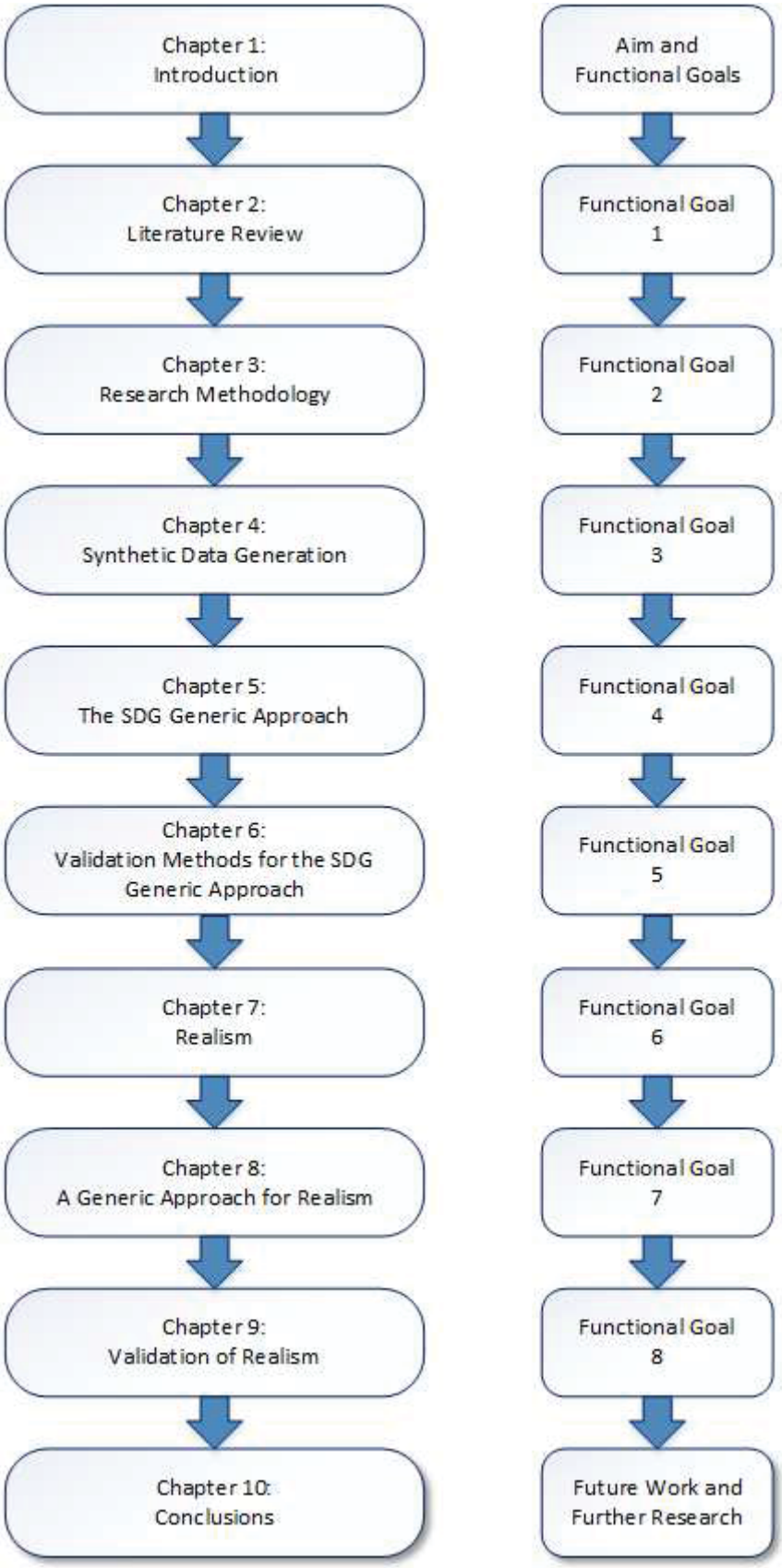
Realism validation in CoMSER consisted of a survey of clinical practitioners who read up to five different synthetic patient records and rated their correctness and visible realistic properties (McLachlan et al, 2016). This survey is reported in the CoMSER article and demonstrates that clinicians felt the health records were highly realistic (McLachlan et al, 2016). In their article the authors did not contemplate rigorous validation and the use of KDD of the type suggested in this thesis (McLachlan et al, 2016). The statistical (quantitative) and rules-based (qualitative) analyses of this case study have validated that a significant degree of realism does exist within CoMSER-generated synthetic patient record, verifying the claims made by the authors that at the time they were only able to support with statistical data from their study and the small number of survey responses (McLachlan et al, 2016). The validation approach presented here incorporates many of the elements of CM validation discussed in chapter 6. It meets the requirements of ***grounding*** (subsection 6.3.1) in that it encourages the researcher to ensure that the generation method is one which is used consistently in the same domain and for the same type of SDG. It performs ***verification*** (subsection 6.3.3) through comparison of quantitative and qualitative aspects of the generated data with the knowledge gathered from input or seed data. Given multiple observational input datasets it is possible to perform ***harmonising*** validation (subsection 6.3.4). While repeat runs may be less necessary, conducting ***calibrating*** validation (subsection 6.3.2) becomes a much simpler matter as the ATEN framework provides a many data points that can facilitate identifying where an issue may have arisen, and therefore where calibration could correctly be applied.

The experience of applying the validation approach described in this chapter exposed that the bulk of effort is actually expended during the knowledge gathering processes described in subsections 8.3 – 8.6. These processes improve the chance that the input data, constraints and method used to generate the synthetic data can deliver the required result, accordingly the effort expended should not be

considered as a burden to those developing a new SDG approach. Rather, given the potential for benefit it should be welcomed. These benefits may come from improved accuracy and a reduction in the number of iterations and lead time required to achieve realistic and useable synthetic data. These benefits are important, especially when we appreciate from reviewing SDG articles that most SDG project's output is simply a means to an end, that is, the SDG project's role is simply to produce synthetic datasets that are required for other research or researchers.

Additional case studies are included in Appendices C and D. Each of these demonstrate some application of the Horus validation approach up to and so far as is practical for the SDG method being reviewed. Each also demonstrates the limitations inherent to the Horus validation approach when it is applied without the full benefit of each of the prior cumulative processes.

This page intentionally left blank


Chapter 1: Introduction
Chapter 2: Literature Review
Chapter 3: Research Methodology
Chapter 4: Synthetic Data Generation
Chapter 5: The SDG Generic Approach
Chapter 6: Validation Methods for the SDG Generic Approach
Chapter 7: Realism
Chapter 8: A Generic Approach for Realism
Chapter 9: Validation of Realism
Chapter 10: Conclusions
Aim and Functional Goals
Functional Goal 1
Functional Goal 2
Functional Goal 3
Functional Goal 4
Functional Goal 5
Functional Goal 6
Functional Goal 7
Functional Goal 8
Future Work and Further Research

10. Conclusion

True adherence to the scientific method should be necessary for any research endeavour. It proved impossible to locate an SDG article seeking realism that provided complete documentation of the input knowledge, generation methodology and experimental approach and the corrections made to arrive at the output data. Some provided solid understanding of the problem and source materials they had used, but then gave only limited snippets of the generation method used. Others focused on the generation method and code, and only briefly touched on the source materials. It was very rare to be provided access to a dataset or subset of synthetic data so that a third party could establish the application and usefulness of the data for themselves. The incomplete nature of the majority of SDG publications constrains reusability, inhibits repeatability and prevents any external validation of the claims made by the authors.

The importance of supporting claims by demonstrating that the method and results have been verified and validated cannot be understated. Validation was absent from the majority of published SDG methods seeking realism, or where it was contemplated, it concerned extrinsic factors like the generation speed or CPU and memory usage or was only presented for a small portion of the output and therefore did not validation the input materials or generation method. In rare instances where validation consisted of overlaying observed data graphs with plots of the synthetic data, we were not told whether the synthetic data used constituted a single generation operation, or was selected as the most accurate or best fit from a number of operations. Use of an established validation method can deliver scientifically acceptable justification for claims of success and constitutes the single most effective proof that synthetic data possesses the necessary element of realism and is therefore a suitable replacement for real or observed data.

This thesis contributes the following:

10.1 The concept of Realism

Realism is a collection of two levels of knowledge. It is defined as the sum total of the extrinsic, or obvious, and the intrinsic, or hidden. The ***extrinsic*** includes structural and readily observable statistical information easily identified from the input materials; How the data is laid out in its columns and rows, the representation and consistency of the fields and values and other things like how many records of a given type and the scale of values. These are also discussed in this research as the quantitative and qualitative elements of the data. ***Intrinsic*** knowledge requires more effort to extract, but presents a far

more detailed picture of the concepts contained within the input materials, the relationships those concepts have with each other and with each independent row of data.

10.2 The THOTH enhanced generic SDG approach

This research presented three previously undocumented elements of the approach to creating synthetic data. The ability to ***classify*** the level of *syntheticness* and ***characterise*** the generation method goes some way to mitigate the inherently linear waterfall nature of SDG efforts. Identification of these two elements provides the researcher with a degree of foresight and direction, which can only improve the resulting synthetic data and bring greater efficiency to the overall SDG project. The third and final element was realisation of the existence of a generic SDG approach itself; an approach that many researchers were using but which had remained unpublished until now. THOTH systematically brings the three together, providing an easy to follow structure that begins with selection of the classification level, identification of the most appropriate method characterisation to be used in the generation process, and finally using both selections to enhance the generic SDG approach.

10.3 The RA approach to systematically identifying realistic elements for SDG

The elements and qualities of realism can be identified through the engagement of the Ra extended knowledge discovery in databases (KDD) approach. This approach first establishes the quantitative and qualitative aspects of the data (the *extrinsic*), followed by an in-depth and structured investigation of the concepts, relationships and rules that exist within the data (the *intrinsic*). The results are documented, integrating statistics collected during the quantitative analysis with the elements of knowledge realised during the KDD operation. An example in the domain of midwifery of each step in the collective process was provided in the case study of Chapter 8.

10.4 The HORUS approach to Validating Realism in SDG

The process of validating realism is one which requires the researcher to verify that each of the identified realistic elements or qualities can be found in its correct form within the synthetic data. The Horus validation process follows that of the Ra knowledge discovery, in that the extrinsic quantitative and qualitative aspects are first assessed and only when they are accurately established does the researcher goes on to verify each of the intrinsic components of knowledge. In this way the validation process is greatly simplified through the benefits gained by having already identified the realistic elements of knowledge prior to generation. The outcome at any step of the validation process may be to continue, or return to one of the previous steps in the enhanced SDG approach shown in Figure 22 in order to address the issue identified. If completed successfully, Horus validation supports the making of claims as to the correctness of the SDG model, quality of the synthetic data and the existence of realism.

The validation process shown in Figure 22 consists of five validation points. If used, each on its own would go some way to support claims seen in the SDG methods explored during this research. Use of

the complete five-step method results in a wealth of both broad and specific analytical knowledge realised from within the observed data. This knowledge guides the SDG researcher in all aspects of their project; as they design the generation method, select or develop tools, constraints and algorithms, and even as they structure the synthetic data that they will output. It is an approach that if applied allows SDG projects to faithfully adhere to established scientific methods of the current literature and empowering and providing confidence to both creators and users of synthetic data.

10.5 Benefits and Limitations

A benefit of the THOTH approach is the ability for the researcher to better plan for the generation method and prepare the knowledge elements and techniques that will be used in the creation of synthetic data. This benefit arises from classification of the level or degree of syntheticness required, and characterisation of the generation method. This characterisation establishes the algorithm, constraints and inputs that may be required for successful generation. To the best of our knowledge a generic approach to SDG was not encountered or previously contemplated in the literature. The THOTH approach can be easily implemented and comes with little overhead. A limitation of THOTH is the unidirectional linear nature of its waterfall-type model, however classification and characterisation may greatly mitigate the effects of this limitation.

When following THOTH, RA benefits the researcher through assurance as to the quality of the synthetic data they will create. This benefit is achieved through establishment of the elements and characteristics that define realism for the generation project; the extrinsic quantitative, intrinsic qualitative aspects of and conditions that inhabit the input data. A number of SDG authors (Bozkurt & Harmon, 2011; Domingo-Ferrer et al, 2012; Gianotti et al, 2004) considered realism as a relevant goal or outcome they sought from their generation method. However, while two authors (Sperotto et al, 2009; Killhourhy et al, 2007) provided vague definitions that *the synthetic data needed to be representative of real data*, no prior work considered what form this realism should take (Section 7.2), the approach needed to accomplish it (Section 7.3), nor even how to verify that it had been successfully achieved (Section 7.3.1). Another benefit of RA is that as additional items of input or seed data are introduced, the statistics, knowledge, constraints and rules are accordingly further refined, increasing the potential accuracy and realism of the synthetic data. A limitation that arises from operation of the RA approach is that it is presently manual, requiring the researcher to possess an eye for detail along with sound logic, analytical and problem solving skills.

The HORUS approach benefits through being an inherently straight-forward model to validate and verify synthetic data. HORUS benefits the researcher by identifying rules, constraints or datasets that may be causing issues that reduce accuracy, realism and therefore the utility of generated synthetic data. Another benefit is that it is not inconceivable that the number of SDG iterations may be reduced, significantly reducing the time taken to produce accurate and realistic synthetic data that can be used in other experiments. No comparable works were located during this research; however, the closest

relatable work encountered was that of Carley (1996), whose work is discussed in chapter 6. Carley's work presented four separate approaches to validation of synthetic data produced in the domain of computational modelling. Each of these approaches appears, even in Carley's summation, to not be representative of a single validation solution. The strength of HORUS is that it represents a single operational validation solution and one which exceeds any of the four separate solutions Carey had proposed. HORUS has a significant limitation in that it is wholly dependent on the researcher having already engaged the RA approach to identify the statistics, knowledge and rules that will be key to providing assurance that the synthetic data is a competent and accurate representation of real data. Another limiting issue is that the case studies conducted in this research identified that where the extrinsic quantitative aspects of the synthetic data are found wanting, continued engagement in the HORUS validation approach looking at the intrinsic knowledge, rules and constraints may be of little additional benefit until the extrinsic issues are resolved.

Each approach contributed by this work is enhanced through interaction and engagement of the others. THOTH provides framework and approach knowledge that improves RA, RA provides the extrinsic and intrinsic knowledge to seed HORUS, and the results of engaging HORUS either identify where an issue may exist in the first two and therefore target where additional work is required, or confirms their successful operation and therefore justify the claim of realism in the synthetic data. The ATEN framework therefore presents as one of the strongest structures seen in engineering and nature, the triangle, each component represented as a corner communicating with its adjacent neighbours.

10.6 Future Work

There are a number of avenues open for future work, including operation of the ATEN framework during the entire lifecycle of a significant SDG project. This would necessitate the considered operation of a new SDG project where every element was documented rigorously, and where two streams or processes are conducted concurrently. In the first or normal stream the SDG project would operate in the manner that the majority do now, following the SDG generic approach described in Chapter 5. No input or other validation steps would be taken and realism would be given no more consideration than it is in the majority of SDG cases reviewed. In the second stream another researcher would collect the input materials and documentation from the first and use it to follow the complete and validated SDG approach described in Figure 22. The second researcher would ameliorate his input materials and generation method as prompted by the enhanced HCI-KDD approach and deliver a second synthetic dataset enhanced with the knowledge discovered. The synthetic datasets should then be validated to assess which is more successful through its proximity to real or observed data and its accuracy to being realistic. As previously discussed a new SDG experiment such as this is required primarily due to the incompleteness of every SDG project reviewed during this research. Another avenue for future work would be development of AI or machine learning models that can automate some or all of the knowledge discovery and validation components.

10.7 Summary

Validation of realism in a modern synthetically generated dataset represents a complicated challenge that until now had not been conquered or even contemplated in the literature. Many claim to have created realistic synthetic data yet few even approach simple validation of their realism proposition. The approach proposed in this thesis draws on, expands and enhances established methods to result in a complete end-to-end validation solution. If SDG authors reported the use of any one validation method their claims of having created realistic synthetic data may have at least had some merit. Use of the entire approach contained herein firstly ensures a complete analysis of the source data. More information can only improve the generation approach, and a better generation approach delivers better synthetic data. Secondly, the knowledge realised also provides a solid base with which to validate the synthetic data that has been created, ensuring its ability to actually replace real data. The approach presented is simple and not overly burdensome, with many of the component steps being activities that data synthesisers may already be undertaking in an albeit unstructured or unconsidered way.

References

Abowd, J., & Woodcock, S. (2001). Disclosure limitation in longitudinal linked data. *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, 215277.

Agarwal, R., Aggarwal, C., & Prasad, V. (2000). A tree projection algorithm for generation of frequent itemsets. *J. of Parallel and Distributed Computing*, 61.

Agrawal, D., Butt, A., Doshi, K., Larriba-Pey, J., Li, M., Reiss, F., Raab, F., Scheifer, B., Sozumura, T., & Xia, Y. (2015). SparkBench – A spark performance testing suite. *Performance Evaluation and Benchmarking: Traditional to Big Data to Internet of Things*, 9508.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. (1996). Fast discovery of association rules. In *Advances in knowledge discovery and data mining*, Am. Assoc. for Artificial Intelligence, Menlo Park: California.

Ahrens, J., & Dieter, U. (1989). *An alias method for sampling from the normal distribution.* Computing 42.

Alessandri, M., Craene, M., Bernard, O., Giffard-Roisin, S., Allain, P., Waechter-Stehle, I., Weese, J., Saloux, E., Delingette, H., Sermesant, M., & D'hooge, J. (2015). A pipeline for the generation of realistic 3D synthetic echocardiographic sequences: Methodology and open-access database. *IEEE Transactions on Medical Imaging*, 34(7).

Alfons, A., Kraft, S., Temple, M., & Filzmoser, P. (2010). *Simulation of synthetic population data for household surveys with application to EU-SILC.* Research Report CS-2010-1. Dept. of Statistics and Probability Theory, Vienna University of Tech.

Alter, H. (1974). Creation of a synthetic data set by linking records of the Canadian survey of Consumer Finances with the Family Expendature Survey 1970. *Annals of Economic and Social Measurement*, 3(2).

Ascoli, G., Krichmar, J., Nasuto, S., & Senft, S. (2001). Generation, description and storage of dendritic morphology data. *Phil. Trans. Of the Royal Society London*, 356.

Assmann, J. (2001). *The search for god in ancient Egypt*. pp. 80-81.

Axtell, R., Axelrod, J., Epstein, J., & Cohen, M. (1996). Aligning simulation models: A case study and results. *Computational and Mathematical Organisational Theory*, 1(2).

Barlas, Y. (1996). Formal aspects of model validity and validation in system dynamics. *System Dynamics Review*, 12(3).

Barnard, K., Martin, L., Coath, A., & Funt, B. (2002). A comparison of computational colour constancy Algorithms – Part II: Experiments with Image Data. *IEEE Transactions on Image Processing*, 11(9).

Barse, E.L., Kvarnström, H., & Jonsson, E. (2003). Synthesizing test data for fraud detection systems. *19th Annual Computer Security Applications Conference, Proceedings of*. pp 384-395

Bate, A., Lindquist, M., Edwards, I., Olsson, S., Orre, R., Lansner, A., & De Freitas, R. (1998). A bayesian neural network method for adverse drug reaction signal generation. *Pharmacoepidemiology and Prescription*, 54(315).

Begue, N., Cramer, J., Bargen, C., Myers, K., Johnson, K., & Morris, R. (2011). Automated method for determining hydrocarbon distributions in mobility fuels. *Energy Fuels*, 25(4).

Bex, G., Neven, F., Schwentick, T., & Tuyls, K. (2006). Inference of concise DTDs from XML data. *Proceedings of the 32nd Int. Conference on Very Large Databases*, pp. 115-126.

Bierkens, M., & Geer, F. (2008) GEO4-4420 Stochastic Hydrology. *Utrecht University.*

Birkin, M., & Clarke, M. (1988). SYNTHESIS – A synthetic spatial information system for urban and regional analysis: Methods and examples. *Environment and Planning A*, 20(1).

Birkin, M., Turner, A., & Wu, B. (2006). A synthetic demographic model of the UK population: Methods, progress and problems. *Regional Science Association, British and Irish Section, 36th Annual Conference.* The Royal Hotel, St Helier, Jersey, Channel Islands.

Bolon-Canedo, V., Sanchez-Marono, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge Information Systems*, 34(1).

Bozkurt, M., & Harman, M. (2011). Automatically generating realistic test input from web services. *Service Oriented System Engineering (SOSE), IEEE 6th International Symposium on.*

Bregt, A. (1997). GIS support for precision agriculture: Problems and possibilities. *Precision Agriculture: Spacial and temporal variability of environmental quality*, Wiley: Chichester. pp173-181.

Brinkhoff, T. (2003). Generating traffic data. *IEEE Data Eng. Bull.*, *26*(2), 19-25.

Brissette, F. P., Khalili, M., & Leconte, R. (2007). Efficient stochastic generation of multi-site synthetic precipitation data. *Journal of Hydrology*, *345*(3), 121-133.

Brooks, F. (1996) Toolsmith II. *Communications of the ACM*, 39(3).

Burton, R., & Obel, B. (1995). The validity of computational models in organisational science: From model realism to purpose of the model. *Computational and Mathematical Organisation Theory*, 1(1).

Cable, G. (1994). Integrating case study and survey research methods: an example in information systems. *European J. of Information Systems*, 3(2).

Carley, K. (1996). Validating computational models. *Carnegie Mellon University.*

Cassa, C., Olson, K., & Mandl, K. (2004). System to generate semisynthetic data sets of outbreak clusters for evaluation of Outbreak-Detection performance. *Morbidity and Mortality Weekly Report (MMWR) September 24, 2004*.

Castellani, B., & Castellani, J. (2008). Data mining: Qualitative analysis with Health Informatics data. *Qualitative Health Research*, 13(7).

Chawla; N., Japkowics, N., & Kotcz, A. (2004). Editorial: Special issue on Learning from Imbalanced Data Sets. *SIGKDD Explorations*, 6(1).

Chen, H., & Han, J. (2010). Stochastic computational models for accurate reliability evaluation of logic circuits. *GLSVLSI '10, ACM*, Providence, Rhode Island.

Cockburn, A. (2003). Research Methods in Information Systems Research: Matching method to researcher. *Humans and Technology*. Sourced from: http://alistair.cockburn.us

Cohen, K., & Cyert, R. (1965). Simulation of Organisational Behaviour. March, J. (ed.) *Handbook of organisations*. Chicago, Il: Rand McNally.

Collins, H. (1992). *Changing order: Replication and induction in scientific practice.* University of Chicago Press.

Crawford, S., & Stucki, L. (1990). Peer review and the changing research record. *J. Am. Society of Information Science*, 41.

Creswell, J. (2003). *Research design: Qualitative, quantitative and mixed methods approaches* (2nd Ed). Sage Publications.

Davis, C. (1993). *The computer generation of multinominal variants*. Computational Statistics and Data Analysis, 16.

Delleur, J., & Kavvas, M. (1978). Stochastic models for monthly rainfall forecasting and synthetic generation. *J. of Applied Meteorology*, 6282.

Deming, W., & Stephan, F. (1940). "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known". *Annals of Mathematical Statistics,* 11(4).

Dey, A. (2001). Understanding and using context. *Personal and Ubiquitous Computing*, 5(1).

Domingo-Ferrer, J. (2012). Marginality: A numerical mapping for enhanced exploitation of taxonomic attributes. In *International Conference on Modelling Decisions for Artificial Intelligence*, Springer: Berlin, Heidelberg.

Drecshler, J. (2014). Synthetic Data: Where do we come from? Where do we want to go? *Synthetic Data Workshop of the Institute for Employment Research.*

Drechsler, J. (2011). *Chapter 2: Background on multiply-imputed synthetic datasets* in *Synthetic datasets for statistical disclosure control: Theory and Implementation*, Springer Science and Business Media, pp 7-11

Drechsler, J. & Reiter, J. (2011). An empirical evaluation of easily implemented, non-parametric methods for generating synthetic datasets. *Computational Statistics and Data Analysis*, 55(12).

Eastern Health (2014). Birth after previous caesarean section. *Eastern Health Maternity Service*. Sourced from: https://www.easternhealth.org.au/images/services/Maternity_-_Birth_after_Previous_Caesarean_Section_APPROVED_0814_189.1.pdf

Edwards, A., Rathkopf, J., & Smidt, R. (1991). Extending the Alias Monte Carlo sampling method to general distributions. *American Nuclear Society International Topic Meeting*.

Efstratiadis, A., Dialynas, Y. G., Kozanis, S., & Koutsoyiannis, D. (2014). A multivariate stochastic model for the generation of synthetic time series at multiple time scales reproducing long-term persistence. *Environmental Modelling & Software*, *62*, 139-152.

Eisenhardt, K. (1989). Building theories from case study research. *Academy of Management Review*, 14(4).

Ekstrom, A. (2015). *The moral bias behind your search results*. TED Oslo.

Elliot, L., Ingham, D., Kyne, A., Mera, N., Pourkashanian, M., & Wilson, C. (2002). The optimisation of reaction rate parameters for chemical kinetic modelling using genetic algorithms. *ASME '02 Proceedings*.

Fabrega, L., Vila, P., Careglio, D., & Papadimitriou, D. (2013). Measurement Methodology and Tools: First European Workshop. Springer: Denmark.

Fabrizio, E., & Monetti, V. (2015). Methodologies and advancements in the calibration of building energy models. *Energies*, 8.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Feldman, D. C., & Arnold, H. J. (1983). *Managing individual and group behavior in organizations*. McGraw-Hill College.

Fernandez-Arteaga, V., Tovilla-Zarate, C., Fresan, A., Gonzalez-Castro, T., Juarez-Rojop, I., Lopez, Narvaez, L., & Hernandez-Diaz, Y. (2016). Association between completed suicide and environmental temperature in a Mexican population, using the KDD approach. *Computer Methods and Programs in Biomedicine*, 135.

Fidel, R., Pejtersen, A., Cleal, B., & Bruce, H. (2004). A multidimensional approach to the study of human-information interaction: A case study of collaborative information retrieval. *J. Am. Soc. For Inf. Sc. And Tech.* 55(11).

Forer, B. (1949). The fallacy of personal validation: A classroom demonstration of gullibility. *The J. of Abnormal and Social Psychology*, 44(1).

Freed, R. E., D'Auria, S. H., & Markowitz, Y. J. (1999). *Pharaohs of the Sun*. Bulfinch Press.

Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyse expression data. *J. of Computational Biology*, 7(3).

Gafurov, T., Usaola, J., & Prodanovic, M. (2015). Incorporating spatial correlation into stochastic generation of solar radiation data. *Solar Energy*, *115*, 74-84.

Galan, J., Izquierdo, L., Izquierdo, S., Santos, J., Olmo, R., Lopez-Paredes, A., & Edmonds, B. (2009). Errors and Artifacts in agent-based modelling, *J. of Artificial Societies and Social Simulation*, 12(11).

Gallagher, A., Ritter, E., & Satava, R. (2003). Fundamental principles of validation and reliability: Rigorous science for the assessment of surgical education and training. *Surgical Endoscopy*, 17(10).

Gao, J., Fan, W., Han, J., & Yu, P. (2007). A General framework for mining concept-drifting data streams with skewed distributions. *International Conference on Data Mining, Proceedings of the 2007*.

Gargiulo, F., Ternes, S., Huet, S., & Deffuant, G. (2010). An iterative approach for generating statistically realistic populations of households. *PloS ONE*, 5(1).

Geweke, J., & Porter-Hudak, S. (1983). The estimation and application of long memory series models. *J. of Time Series Analysis*, 4(4).

Giannotti, F., Mazzoni, A., Puntoni, S., & Renso, C. (2005, November). Synthetic generation of cellular network positioning data. In *Proceedings of the 13th annual ACM international workshop on Geographic information systems* (pp. 12-20). ACM.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond Heuristics and Biases. *European Review of Social Psychology*, 2.

Gore, R. (2001). Pharaohs of the Sun. *National Geographic*, *199*(4), 35-57.

Graham, V. A., Hollands, K. G. T., & Unny, T. E. (1988). A time series model for Kt with application to global synthetic weather generation. *Solar Energy*, *40*(2), 83-92.

Granger, C., & Terasvirta, T. (1993). Modelling non-linear economic relationships. *OUP Catalogue.*

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational evaluation and policy analysis*, *11*(3), 255-274.

Green, P. E., & Rao, V. R. (1971). Conjoint measurement for quantifying judgmental data. *Journal of Marketing research*, 355-363.

Green, T. (1992). *The city of the moon god: Religious traditions of Harran.* Brill.

Gunn, B. (1923). Notes on the Aten and his names. *The Journal of Egyptian Archaeology*, *9*(3/4), 168-176.

Haggstrom, M. (2009). Blood values sorted by mas and molar concentration. Sourced from: http://goo.gl/E5wnAv

Haig, B. (1995). Grounded theory as scientific method. *Philosophy of Education*, 28(1).

Han, J., Cai, Y., & Cercone, N. (1993). Data-driven discovery of quantitative rules in relational databases. *IEEE Transactions on Knowledge and Data Engineering*, 5(1).

Harvey, N. (2007). Use of heuristics: Insights from forecasting research. *Thinking & Reasoning*, 13(1).

Hoffman, P. (1998). *The man who loved only numbers: The story of Paul Erdos and the search for mathematical truth.* New York: Hyperion.

Holzinger, A., Dehmer, M., & Jurisica, I. (2014). Knowledge discovery and interactive data mining in Bopinformatics – State-of-the-art, future challenges and research directions. *BMC Bioinformatics*, 15(6).

Houkjaer, K., Torp, K., & Wind, R. (2006). Simple and realistic data generation. *VLDB '06*.

Ishigami, M., Cumings, J., Zetti, A., & Chen, S. (2000). A simple method for the continuous production of carbon nanotubes. *Chemical Physics Letters*, 319(5).

Jaderberg, M., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Synthetic data and artificial neural networks for natural scene text recognition. *arXiv:1406.2227*.

Jain, A., & Zongker, D. (1997). Feature Selection: Evaluation, application and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2).

Kanungo, T., & Resnik, P. (1999). The Bible, truth and multilingual OCR evaluation. *Proceedings SPIE 3651.*

Killourhy, K., & Maxion, R. (2007). Toward realistic and artefact-free insider-threat data. *23rd Annual Computer Security Applications Conference, Proceedings of (CSAC).*

Kleijnen, J. (1995a). Verification and validation of simulation models. *European J of Operational Research*. 82(1).

Kleijnen, J. (1995b). Statistical validation of simulation models. *European J. of Operational Research.* 87(1).

Klein, M., & Sinha, B. (2015). Likelihood-based finite sample inference for synthetic data based on exponential model. *Thailand Statistician,* 13(1).

Kohnen, C. N. (2005). *Using Multiply-Imputed, Synthetic Data to Facilitate Data Sharing* (Doctoral dissertation, PhD Dissertation, Institute of Statistics and Decision Sciences, Duke University).

Koppel, R., Wetterneck, T., Telles, J., & Karsh, B. (2008). Workarounds to barcode medication administration systems: Their occurences, causes and threats to patient safety. *J. of the Am. Medical Informatics Assoc.* 15(4).

Kosseim, P. & Brady, M. (2008). Policy by procrastination: Secondary use of electronic health records for health purposes. *McGill JL & Health*, 2(5).

Kneppell, P., & Aragno, D. (1993). *Simulation validation: A confidence assessment methodology*. Los Alamitos, CA. IEEE Computer Society Press.

Kuiper, J., Heuvel, E., & Swertz, M. (2015). b. *Biopreservation and Biobanking*. 13(3).

Kvale, S. (1994). Ten standard objections to qualitative research interviews. *J. Phenomenological Psychology*, 25(2).

Laranjeiro, N., Vieira, M., & Madeira, H. (2009). Improving web services robustness. *Proceedings of the IEEE Int. Conference on Web Services ICWS'09*, pp. 397-404.

Laymon, R. (1984). The path from data to theory. In Leplin, J. (Ed.) *Scientific Realism*. University of California Press.

Lee, A. (1989). A scientific method for MIS case studies. *MIS Quarterly*, March 1989, pp. 33-50.

Lee, J., & Carley, K. (2004). OrgAhead: A computational model of organisational learning and decision making. *CASOS Technical Report, Carnegie Melon University*.

Leggett, D. & McBryde, W. (1975). General computer program for the computation of stability constants from absorbance data. *Analytical Chemistry*, 47(7) pp. 1065.

Lertpalangsunti, N., Chan, C. W., Mason, R., & Tontiwachwuthikul, P. (1999). A toolset for construction of hybrid intelligent forecasting systems: application for water demand prediction. *Artificial Intelligence in Engineering*, *13*(1), 21-42.

Levin, M. (1984). What kind of explanation is truth? In Leplin, J. (Ed.) *Scientific Realism*. University of California Press.

Little, R. J. (1993). Statistical analysis of masked data. *Journal of Official statistics*, *9*(2), 407.

Ludwig, J. (2016). Horus Falcon. *Analecta*, 29. pp 74-75.

Lujano-Rojas, J., Dufo-Lopez, R., & Bernal-Agustin, J. (2013). Probabilistic modelling and analysis of stand-alone hybrid power systems. *Energy*, 63.

Lundin, E., Kvarnstrom, H., & Jonsson, E. (2002). A synthetic fraud data generation methodology. *Information and Communications Security.* 2513.

Lydiard, T. (1992). Overview of the current practice and research initiatives for the verification and validation of KBS. *The Knowledge Engineering Review*, 7(2).

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., & Vilhuber, L. (2008). "Privacy: Theory meets Practice on the Map". *2008 IEEE 24th International Conference on Data Engineering*. Pp 277–286.

Mahmoud, E., (1984). Accuracy in forecasting: A survey. *J. of Forecasting*, 3(2).

Manno, I. (1999). *Introduction to the Monte Carlo method.* Budapest, Hungary: Akademiai Kaido.

Mateo-Sanz, J., Martinez-Balleste, A., & Domingo-Ferrer. J. (2004). Fast generation of accurate synthetic microdata. In *International Workshop on Privacy in Statistical Databases* pp. 298-306. Springer, Berlin Heidelberg.

Markov, A. (1971). Extension of the limit theorems of probability theory to a sum of variables connected in a chain. Reprinted in Appendix B of R. Howard, *Dynamic Probabilistic Systems, Vol 1: Markov Chains*. New York: John Wiley and Sons.

McDonough, J. & McDonough, S. (1997). *Research methods for English Language Teachers*. London: Arnold.

McHugh, J. (2000). Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA Intrusion detection system evaluations as performed by Lincoln Laboratory. *ACM Transactions on Information and Systems Security*, 3(4).

McLachlan, S., Dube, K., & Gallagher, T. (2016). Using CareMaps and health statistics for generating the realistic synthetic EHR. *Manuscript accepted for publication, ICHI'16.*

McMullin, E. (1984). A case for scientific realism. In Leplin, J. (Ed.) *Scientific Realism* (pp. 248-281). University of California Press.

Meng, D., Wang, J., Gu, Z., Zhang, Q., & Wang, W. (2013). Military information security based on modelling and simulation. *SPBEI'13.*

MerriamWebster, (nd). *Network*. Sourced from: https://www.merriam-webster.com/dictionary/network

Ministry of Health, (2014). *Maternity and newborn data and stats.* Sourced from: http://goo.gl/httD3b

Mitra, S., Pal, S., & Mitra, P. (2002). Data mining in soft computing framework: A survey. *IEEE Transactions on Neural Networks*, 13(1).

Mora, M., Fernandez, M., Gomez, F., Cantero, D., Lafuente, J., Gamisans, X., & Gabriel, D. (2015). Kinetic and stoichiometric characterisation of anoxic sulphide oxidation by SO-NR mixed cultures from anoxic biotrickling filters. *Environ. Biochem.* 99(1).

Moss, P. (1994). Can there be validity without reliability? *Educational Research*, 23(2).

Mouza, C., Metais, E., Lammari, N., Akoka, J., Aubonnet, T., Comyn-Wattiau, I., Fadili, H., & Cherfi, S. (2010). Towards an automatic detection of sensitive information in a database. *Advances in database knowledge and database applications, 2nd International conference on.*

Mwogi, T., Biondich, P., & Grannis, S. (2014). An evaluation of two methods for generating synthetic HL7 segments reflecting real-world health information exchange transactions. *AMIA Annu Symp Proc.*

Nicoletti, I., Migliorati, G., Pagliacci, M., Grignani, F., & Riccardi, C. (1991). A rapid and simple method for measuring thymocyte apoptosis by propidium iodide staining and flow cytometry. *J. of Immunological Methods.* 139(2).

Ngoko, B. O., Sugihara, H., & Funaki, T. (2014). Synthetic generation of high temporal resolution solar radiation data using Markov models. *Solar Energy*, *103*, 160-170.

Nguyen, Q., & Leung, P. (2009). Do fishermen have different attitudes toward risk? An application of prospect theory to the study of Vietnamese fishermen. *J. of Agricultural and Resource Econom.* 34(3).

Nijssen, G. M., & Halpin, T. A. (1989). *Conceptual Schema and Relational Database Design: a fact oriented approach*. Prentice-Hall, Inc.

Nunamaker, J., & Chen, M. (1990). Systems development in information systems research. *System Sciences*, 23rd Annual *International conference on.* Hawaii, USA.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation and confirmation of Numerical models in the earth sciences. *Science*, 263(5147).

Oxford, (2016). *Definition of validation in English.* Sourced from: https://en.oxforddictionaries.com/definition/us/validation

Parker, S. (2003). *McGraw-Hill Dictionary of Scientific and Technical Terms*. Sixth Edition. McGraw-Hill Education.

Parnas, D., & Clements, P. (1986). A Rational design process: How and why to fake it. *Software Engineering, IEEE Transactions*: 251-257.

Peak, D., Guynes, C., & Kroon, V. (2005). Information Technology Alignment Planning – A case study. *Information and Management*, 42.

Penduff, T., Barnier, B., Molines, J., & Madec, G. (2006). On the use of current meter data to assess the realism of ocean model simulations. *Ocean Modelling.* 11(3).

Penzotti, J., Lamb, M., Evensen, E., & Grootenhuis, P. (2002). A computational ensemble pharmacophore model for identifying substrates of P-Glycoprotein. *J. Medicinal Chemistry*, 45(9).

Peters, F., Drummer, O., & Musshoff, F. (2007). Validation of new methods. *Forensic Science International*, 165.

Pickard, A. J. (2013). *Research methods in information*. Facet publishing.

Ponzini, R., Biancolini, M., Rizzo, G., & Morbiducci, U. (2012). Radial bias functions for the interpolation of hemodynamics flow pattern: A quantitative analysis. Giamberatdino, D et al (Eds.) *Computational Modelling of Objects Represented in Images.* London: Taylor and Francis.

Porter, R. (2011). Insights into Egyptian Horus Falcon Imagery by Way of Real Falcons and Horus Falcon Influence in the Aegean in the Middle Bronze Age: Part I. *Journal of Ancient Egyptian Interconnections*, *3*(3), 27-38.

Prather, J., Lobach, D., Goodwin, L., Hales, J., Hage, M., & Hammond, W. (1997). Medical Data Mining: Knowledge discovery in a clinical data warehouse. *AMIA Annual Fall Symposium, Proceedings of.*

Pressman, R. (1998). Can internet-based applications be engineered? *IEEE Software*, *Sept/Oct 1998.*

Psillos, S. (2005). *Scientific Realism: How Science tracks truth.* Routledge.

Pudjijono, A. & Christen, P., (2009). Accurate synthetic generation of realistic personal information. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 507-514). Springer Berlin Heidelberg.

Pukelsheim, F., & Simeone, B (2009). On the Iterative Proportional Fitting Procedure: Structure of Accumulation Points and L1-Error Analysis. *Augsburg University: Institute for Mathematics.*

Putnam, H. (1977). Realism and reason. *Proceedings and Addresses of the American Philosophical Assoc.* 50(6) pp483-498.

Radhakrishnan, R., Kharrazi, M., & Memon, N. (2005). Data Masking: A new approach for stenography? *J. of VLSI Signal Processing*, 41.

Raghunathan, T., Reiter, J., & Rubin, D. (2003). Multiple imputation for statistical disclosure limitation. *J. of Official Statistics*, 19(1).

Reiter, J. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, 30(235).

Reiter, J. (2004a). New Approaches to data dissemination: A glimpse into the future. *Chance,* 17(3), pp 11-15.

Reiter, J., & Kinney, S. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *J. of Official Statistics*, 28(4).

Robey, D. & Sahay, S. (1996). Transforming work through information technology: A case study of geographical information systems in county government. *Information Systems Research*, 7(1).

Rodriguez-Jiminez, J., Cordero, P., Enciso, M., & Rudolph, S. (2016). Concept lattices with negative information: A characterisation theorem. *Information Sciences*, 369(51).

Rosevear, A. (1984). Immobilised biocatalysts – A critical review. *J. of Chemical technology and Biotechnology*. 34(3).

Rubin, D. (1976). Inference and missing data. *Biometrika*, 63(3).

Rubin, D. (1987). *Multiple imputation for non-response in surveys*. New York: John Wiley and Sons.

Rubin, D. (1993). Discussion: Statistical disclosure limitation. *J. of Official Statistics.* 9(2). pp 461-468.

Rubin, D. (1996). Multiple Imputation after 18+ years. *J. of the Am. Statistical Assoc.* 91.

Sakshaug, J. (2011). Synthetic data for small area estimation. *PhD Thesis*, University of Michigan.

Salvat, F., Fernandez-Varea, J., & Sempau, J. (2006). PENELOPE-2006: A code system for Monte Carlo simulation of electron and photon transport. In *Workshop Proceedings*, (Vol. 7).

Sanderson, M., & Croft, B. (1999). Deriving concept hierarchies from text. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 206-213). ACM.

Sarkar, S., & Boyer, K. (1994). A computational structure for pre-attentive perceptual organisation: Graphical enumeration and voting methods. *IEEE Transactions on Systems, Man and Cybernetics*, 24(2).

Simonovic, S. (2012). Floods in a changing climate. *Cambridge University Press*.

Skalka, A. (2009). Understanding human dynamics: Report of the Defence Science Board Task Force. DIANE Publishing: Washington.

Smith, D., Clarke, G., & Harland, K. (2009). Improving the synthetic data generation process in spatial microsimulation models. *Environment and Planning*, 41(1).

Smithson, S. (1994). Information retrieval evaluation in practice: A case study approach. *Information Processing & Management*, 30(2).

Sobh, R., & Perry, C. (2006). Research design and data analysis in realism research. *European J. of Marketing*, 40(11).

Sperotto, A., Sadre, R., Van Vliet, F., & Pras, A. (2009). A labelled data set for flow-based intrusion detection. *Proceedings of the 9th IEEE International Workshop on IP Operations and Management (IPOM '09)*, pp 39-50.

Srikanthan, R., & McMahon, T. (2001). Stochastic generation of annual, monthly and daily climate data: A review. *Hydrology and earth system sciences.* 5(4). pp 653-670.

Stratigopoulos, H., Mir, S., & Makris, Y. (2009). Enrichment of limited training sets in machine-learning-based analog/RF test. *DATE '09*.

Stedinger, J., & Taylor, M. (1982). Synthetic streamflow generation: Model verification and validation. *Water Resources Research*, 18(4).

Stodden, V. (2010). The scientific method in practice: Reproducibility in the computational sciences. *SSRN Paper 1550193*, MIT Sloan School of Management Working Paper 4773-10.

Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., & Lakhal, L. (2002). Computing iceberg concept lattices with TITANIC. *Data & knowledge engineering*, *42*(2), 189-222.

Tellis, W. (1997). Application of a case study methodology. *The Qualitative Report*, 3(3).

Thiele, J., Kurth, W., & Grimm, V. (2014). Facilitating parameter estimation and sensitivity analysis of agent-based models: A cookbook using Netlogo and R. *J. Artificial Societies and Social Simulation*, 17(3).

Tichy, W. (1997). Should computer scientists experiment more? *IEEE Computer*, 31(5).

Tsuzuki, S., & Tanabe, K. (1991). Refinement of molecular mechanics parameters for ethers based on conformation energies of Me-O-X obtained by ab initio molecular orbit calculations. *J. Chem. Soc.* 2.

Tsvetovat, M., & Carley, K. (2005). Generation of realistic social network datasets for testing of analysis and simulation tools. *DTIC Document, Technical report 9*.

Valtchev, P., Grosser, D., Rourne, C., & Hacene, M. (2003). Galicia: An open platform for lattices. In *De Moor, A. & Ganter, B.* (Eds) *Using Conceptual Structures: Contributions to 11th Intl. Conference on Conceptual Structures*. pp. 241-254.

Van den Bulcke, T., Van Leemput, K., Naudts, B., van Remortel, P., Ma, H., Verschoren, A., & Marchal, K. (2006). SynTReN: A generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, 7(1).

Van Holde, K., & Weischet, W. (1978). Boundary analysis of sedimentation-velocity experiments with monodisperse and paucidisperse solutes. *Biopolymers*, 17(6) pp 1387-1403.

Velasquez, J. (1997). Modelling emotions and other motivations in Synthetic Agents. *Proceedings of Am. Assoc. for Artificial Intelligence*, AAAI'97.

Venti, S. (1984). The effects of income maintenance on work, schooling and non-market activities of youth. *The Review of Economics and Statistics.* 66(1).

Voss, D., Clements, J., Cole, K., Ford, M., Handy, C., & Stoval, A., (2011). Real science, real education: The university nanosat program. *Small Satellites, the 25th Annual Conference on.*

Walker, A. (1974). *New fast method for generating discrete random numbers with arbitrary frequency distributions.* Electronic Letters, 10(8).

Walker, A. (1977). *An efficient method for generating discrete random variables with general distributions*. ACM Transactions on Mathematical Software, 3(3).

Wan, L., Zhu, J., Bertino, L., & Wang, H. (2008). Initial Ensemble generation and validation for ocean data assimilation using HYCOM in the pacific. *Ocean Dynamics*. 58. pp 81-99.

Weinberg, D., Abowd, J., Steel, P., Zayatz, L., & Rowland, S. (2007). Access methods for United States microdata. *US Census Bureau Centre for Economic Studies* Paper No. CES-WP-07-25.

Weston, J., Bordes, A., Chopra, s., Rush, A., Merrienboer, B., Joulin, A., & Mikolov, T. (2015). Towards AI-complete question answering: A set of prerequisite toy tasks. *Under review as a conference paper at ICLR*.

Whiting, M., Haack, J., & Varley, C. (2008). Creating realistic, scenario-based synthetic data for test and evaluation of information analytics software. *Proceedings of the 2008 Workshop on beyond time and errors: Novel evaluation methods for Information Visualisation (BELIV '08)*.

Wikipedia, (2009a). Synthetic Data. (Historic version). Sourced from: https://en.wikipedia.org/w/index.php?title=Synthetic_data&oldid=329971695

Wikipedia, (2009b). Synthetic Data. (Historic version). Sourced from: https://en.wikipedia.org/w/index.php?title=Synthetic_data&oldid=330434049

Wilkinson, R. (2003). *The complete gods and goddesses of ancient Egypt*. p. 217.

Wilkinson, R. (2008). Anthropomorphic Deities. *UCLA Encyclopedia of Egyptology*, *1*(1).

Williams, K., Ford, R., Bishop, I., Loiterton, D., & Hickey, J. (2007). Realism and selectivity in data-driven visualisations: A process for developing viewer-oriented landscape surrogates. *Landscape and Urban Planning*, 81.

Willie, R. (1992). Concept lattices and conceptual knowledge systems. *Computers and Mathematical Applications*, 23(6).

Winkler, W. (2004). Masking and re-identification methods for public-use microdata: Overview and research problems. In *Workshop on privacy in Statistical Databases*, Springer: Berlin, Heidelberg.

Wolff, E. (1980). *Estimates of the 1969 size distribution of household wealth in the US from a synthetic data base*, in: Smith, J (Ed.) Modelling the distribution and intergenerational transmission of wealth. University of Chicago Press.

Wu, X., Wang, Y., & Zheng, Y. (2003). Privacy preserving database application testing. *WPES '03*.

Yin, R. (1984). *Case study research: Design and methods*. Beverly Hills, California: Sage Publishing

Yin, R. (2011). *Applications of Case Study Research*. Sage.

Yu, M. (2008). Disclosure risk assessments and control. *Ph.D. Dissertation* University of Michigan.

Yuan, Y. (2005). Multiple imputation for missing data: concepts and new developments. Rockville. *MD, SAS Institute*.

Yuan, C. (2010). Multiple imputation for missing data: Concepts and new development (Version 9.0). *SAS Institute, Rockville* (MD 49).

Zainal, Z. (2007). Case study as a research method. *Jurnal Kemanusiaan*, 9.

Zanero, S. (2007). Flaws and frauds in the evaluation of IDS/IPS technologies. *Forum of Incident Response and Security Teams (FIRST '07).*

Zelkowitz, M., & Wallace, D. (1998). Experimental models for validating technology. *IEEE Computer*, 31(5).

Appendix A: Synthetic Data Generation Literature

Table 20: Synthetic Data Generation Literature

Lead Author	Title	Year	Academic Field	Synthesis Approach	Synthesis Scale	Realism
Ascoli, G	Generation, description and storage of dendritic morphology data	2001	Neuroscience (bioinformatics sim)	Network Generation (rules-based algorithms)	Synthetic	Intended / implied
Barse, E	Synthesising test data for fraud detection systems	2003	Computer Science (fraud detection)	Signal and Noise (probability distribution)	Semi-synthetic	Intended / inherited
Bozkurt, M.	Automatically generating realistic test input from web services	2011	Computer Science (software dev)	Format Masking (structural validity)	Synthetic	Intended / implied
			Computer Science (software dev)	Format Masking (semantic validity)	Semi-synthetic	Intended / inherited
Brinkhoff, T	Generating traffic data.	2003	Data Engineering (Geoinformatics)	Network Generation (Spatio-temporal Algorithms)	Synthetic	Intended / implied
			Data Engineering (Geoinformatics)	City Simulation (Spatio-temporal Algorithms)	Synthetic	Intended / implied
Brissette, F.	Efficient stochastic generation of multi-site synthetic precipitation data.	2007	Environmental Sc. (Weather model)	Multi-site precipitation model (improved Wilkes model)	Synthetic	Not intended / implied
Cao, J.	Stochastic models for generating synthetic HTTP source traffic	2004	Computer Science (tcp/http modelling)	FSD Time Series Model (user connection sim)	Synthetic	Intended/ implied
Carapellucci, R.	A new approach for synthetically generating wind speeds: A comparison with the Markov chains method.	2013	Energy Science (Wind Modelling)	Stochastic Generation (Markov chains)	Semi-Synthetic	Intended / Inherited

Cooper, C.	Realistic synthetic data for testing association rule mining algorithms for market-based databases	2007	Computer Science (Database Analysis)	Transactional Modelling (referential attachment)	Synthetic	Intended / implied
Domingo-Ferrer, J	Marginality: A numerical mapping for enhanced treatment of nominal and hierarchical attributes.	2012	Computer Science (SDC)	Numerical Hierarchical Mapping (Data Masking)	Semi-Synthetic	Intended / Inherited
Efstratiadis, A	A multivariate stochastic model for the generation of synthetic time series at multiple time scales reproducing long-term persistence.	2014	Environmental Sc. (Rainfall Modelling)	Time Series Generation (Stochastic Model)	Synthetic	Intended / Inherited
Gafurov, T	Incorporating spatial correlation into stochastic generation of solar radiation data.	2015	Energy Science (Solar Modelling)	Temporal Statistic Gen. (Stochastic Model)	Synthetic	Intended / Inherited
Gargiulo, F.	An iterative approach for generating statistically realistic populations of households	2010	Computer Science (Pop^n modelling)	Weighted Random Generation (probabilistic)	Synthetic	Intended / inherited
Giannotti, F.	Synthetic generation of cellular network positioning data	2005	Computer Science (telecoms)	Spatio-temporal Data Generation	Synthetic	Not intended / implied
Houkjaer, K.	Simple and realistic data generation	2006	Computer Science (database dev)	Weighted selection (monte carlo type)	Synthetic	Intended / implied
Jaderberg, M.	Synthetic data and artificial networks for natural scene text recognition	2014	Computer Science (scene text recog)	Text Rendering (image generation)	Synthetic	Intended/ inherited
Jeske, D.	Generation of synthetic data sets for evaluating the accuracy of knowledge discovery systems.	2005	Mathematical Sc. (Data mining)	Semantic Graphs (Constraint based)	Synthetic	Intended / implied
Killourhy, K.	Toward realistic and artefact-free insider-threat data	2007	Computer Science (intrusion detection)	Signal and Noise (sanitised user data)	Non-synthetic	Intended / inherited
Mateo-Sanz, J	Fast generation of accurate synthetic microdata	2004	Computer Science (mathematics)	Weighted random generation (non-iterative method)	Synthetic	Intended / inherited
McKenna, E	Four-state domestic building occupancy model for energy demand simulations	2015	Electrical Eng. (occupancy model)	Stochastic Markov Chain (Time-Use Survey)	Synthetic	Intended / implied

Mouza, C.	Towards an automatic detection of sensitive information in a database	2010	Computer Science (software dev)	Data Scrambling (detection-replacement-propagation)	Semi-synthetic	Intended / inherited
Mwogi, T.	An evaluation of two methods for generating synthetic HL7 segments reflecting real-world information exchange transactions.	2014	Health Informatics (data generation)	Markov Chain Model (weighted probability)	Synthetic	Intended / implied
			Health Informatics (data generation)	Music Box Model (random selection seed)	Semi-synthetic	Intended / inherited
Ngoko, B.	Synthetic generation of high temporal resolution solar radiation data using Markov models	2014	Energy Science (Solar Modelling)	Temporal Statistic Gen. (Stochastic & Markov chains)	Synthetic	Intended / Implied
Pei, Y	A synthetic data generator for clustering and outlier analysis	2006	Computer Science (data mining)	Test Dataset Generation (distribution and transformation)	Synthetic	Not intended
Pudjijono, A.	Accurate synthetic generation of realistic personal information	2009	Computer Science (SDG)	Frequency-constrained and dependency-aware random generation	Synthetic	Intended / implied
Richardson, I.	A high resolution domestic building occupancy model for energy demand simulation.	2008	Electrical Eng. (Energy use)	Markov Chain Model (demand modelling)	Synthetic	Intended / Inherited
Sperotto, A.	A labelled dataset for flow-based intrusion detection	2009	Computer Science (intrusion detection)	Signal and Noise (flow analysis)	Non-synthetic	Intended / inherited
Tsevotat, M.	Generation of realistic social network datasets for testing of analysis and simulation tools	2005	Computer Science (Human interaction)	Network Generation (constrained random)	Synthetic	Intended / implied
Van den Bulcke, T.	SynTReN: A generator of synthetic gene expression data for design and analysis of structural learning algorithms.	2006	Computer Science (health informatics)	Network Generation (random graph gen)	Synthetic	Intended / implied
Wan, L.	Initial ensemble generation and validation for ocean data assimilation using HYCOM in the Pacific	2008	Meteorology (forecasting)	Data Assimilation (weighted statistical estimation)	Synthetic	Intended / inherited

Whiting, M.	Creating realistic, scenario-based synthetic data for test and evaluation of information analytics software	2008	Computer Science (threat detection)	Signal and Noise	Semi-synthetic	Intended / implied
Winkler, WE.	Masking and re-identification methods for public-use microdata: Overview and research problems.	2004	Statistical Sc. (Census Data)	SDC (Data Masking)	Semi-synthetic	Intended / inherited
Yu, Y.	Synthetic data generation to support irregular sampling in sensor networks.	2003	Environmental Sc. (sensor networks)	Spatial Interpolation (Spatio-temporal modelling)	Synthetic	Intended / implied

Intended: The authors specifically mention their want or intention that the generated data be realistic.
Inherited: If realism is found in the resulting dataset, it is inherited from the seed data used in the generation process
Implied: If realism is found in the resulting dataset, it is implied by virtue of the constraints used to model the data

Appendix B: Realism in SDG Approaches

Table 21: Realism in SDG Approaches

Lead Author	Year	Realism Introduction	Barriers to Realism	Benefits from Realism	Realism Approach	Realism Validation
Ascoli	2001	The nearest thing to realism suggested by the authors is a comment in the abstract regarding their generating 'accurate' neuronal structures.	The issue for this project is the acquisition of real neuronal cell maps – which is costly and time consuming (and complicated in species such as humans) (Introduction part a).	Benefit gained from the development of a simple and cheap statistical method for generating 'virtual neurons'	The authors developed a method for digitally representing the structure observed in neurons.	Validation discussed as two forms of verification: one to ensure that the output matches the parameters set prior to generation, the second to verify against experimental observations.
Barse	2003	The method ensures that important statistical properties of authentic data are preserved by using authentic normal data and fraud to seed synthetic data (abstract).	Issues might include that the service being tested is in development, and may therefore produce irregular data, there may be limited control over the volume of data being captured by a real system and uncertainty as to whether real data contains actual fraud traces.	Synthetic data can be created to demonstrate or test particular issues or properties. Can be generated to cover large periods of time.	The authors collect real data over a small period to be used to generate synthetic data based on synthetic user profiles and scenarios.	Validation consists of analysing synthetic data for user fraud instances.

Bozkurt	2011	Generating realistic test data is a major problem for software testers (introduction). What is needed is a technique that is both realistic and automated (introduction). The article gives in depth discussion of realism.	Automated testing approaches can generate test data that is unrealistic. Creating test data manually is laborious, but can produce some realistic test data.	Automated, Tailored, Applicability, with minimized dependence on existing data sources.	Approach categorises input data and uses the services that generate real data as the solution to generating synthetic data. Approach is limited to semantic systems as these systems allow for the discovery of required data, and the relations between different data structures. Their examples use simplistic systems that are easily understood, such as a library's ISBN data type. They use basic field descriptions in some random generation runs, and semantic rules in others.	The authors test various incarnations of their system in generating ISBN and ZIP code numbers. They compare these to real ISBN and ZIP codes.
Brinkhoff	2003	"…demands for generators that produce realistic data sets." (abstract)	Location based services research is limited by a lack of realistic simulation of moving traffic within given infrastructures. Is a pioneering approach but requires significant testing to prove itself.	Authors demonstrate two pioneering approaches to modelling moving traffic in time and space.	User modelling and constraints applied to generate completely synthetic data structures.	None discussed.
Brissette	2007	The author's talk of the fact that in realistic cases multiple sites should be generated (not a single site) and that the number should be limited in most cases to less than 20.	Research limited by a lack of weather generation approaches for climate change studies.	Benefit comes from the demonstration of an approach that can maintain some consistency even when 'noise' is deliberately introduced into the generation method.	Modelling of precipitation across multiple sites using a basin area model for demonstration.	None discussed
Cooper	2007	"…generates structures that are closer to real-life market basket data." (abstract)	Existing testing models most often use data generated by the QUEST engine which many consider does not adequately represent real data.	The provision of a simpler model and method that can produce data the authors believe to be closer to real-life	The authors performed analysis on some (sanitised) publicly available real data sources and compared these to QUEST generated data. They used the sum of these analyses to produce a simplified QUEST model with different production properties.	Validation consists of comparisons between Author's model and Quest and of produced datasets against real datasets. This is not thoroughly explained or discussed, however.

Domingo-Ferrer	2012	"…by generating synthetic data preserving some statistical features of the original, or hybrid data obtained by combining both…"	…publication of statistical data in such a way that the individual responses of specific users cannot be inferred… Privacy through SDC	Preserves analytical validity	Data masking through numerical mapping for selected nominal hierarchical data.	No validation of realism or a sample application, however some complicated mathematical proofs of the algorithm in the appendices.
Efstratiadis, A	2014	"…properly represent not only the spatial and temporal variability of rainfall, but also the statistical properties…" "…generated realistic patterns of daily discharge."	Management of water resource systems is difficult to model due to intrinsically uncertain nature of hydrometeorological phenomena.	Claims to generate realistic predictions of rainfall patterns	Performs statistical analysis of large number of real observations in order to build generation model that largely self-corrects to mimic	Visual overlay of predictions to actual.
Gafurov, T.	2015	"…so that the resultant synthetic values… have realistic spatial correlations."	Apparently there are no known methods for multisite generation of solar radiation data from reduced (easily available) inputs.	Claims to be first model to incorporate spatial correlations for solar radiation to improve data quality.	Uses seed data observations to then generate hourly predictions	Visual overlay of prediction to actual (wide variance).
Giannotti	2004	Authors declare their interest in '*generating logs with realistic semantics*' (top of page two).	Real cell phone tracking data from network providers is personally sensitive information and therefore not publicly available for research.	Behaviour patterns from mobile trajectories may allow inducing traffic flow information, helping people to get around more efficiently and effectively. It may also help mobile operators to optimise bandwidth and power allocations.	The primary realism approach in the method described is the imputation of given real-world parameters, such as the location of cell phone towers and obstacles such as high-rise buildings. The actual movement data is synthetic and generally only constrained by the location and type of obstacles.	None described
Houkjaer	2006	That test data for software and database developers 'must be realistic' which they define as being 'correct in size and distributions'. (Introduction) They also discuss enhancing the realism of the database by using a non-empty schema and the	The authors claim that the specificity inherent to most models has a hampering effect.	Benefits claimed by their model is generalism. Speed. None of these is a benefit of realism, however. The authors don't define the benefits that their model may derive from realism besides claiming it has highly accurate characteristics.	Majority of paper is about performance of model. The model essentially populates fields of database based on architecture/graph model and a number of basic dependency rules.	Only tests were on performance, no validation of realism.

		addition of user-definable data.				
Jaderberg	2014	The authors claim to have trained their neural network models with synthetic data that is 'highly realistic and sufficient to replace real data' (abstract).	After discussing the current state of OCR systems and a number of issues and conditions that these systems cannot deal with, the authors conclude in the introduction that text recognition represents a challenging problem.	There is no discernible discussion as to any real world benefits or future uses that might come from this research.	They used a font rendering method to produce single-word text images that are all of the same fixed height for their text recognition engine to decode. This results in a system that unrealistically 'knows what to expect' and just like the OCR systems they discuss in the introduction, has difficulty working outside of the set height parameter.	None described
Jeske	2005	"…attractive alternative for testing an IDAS when realistic datasets are not readily available" (introduction)	Generating a realistic dataset for all types of IDAS testing is a formidable challenge. Authors promote concept of generating data of *sufficient quality*.	Generates datasets for training IDAS systems that can be used to trawl large datasets and alerting to anomalies	Generates synthetic data to populate database built on semantic graph with known field constraints.	None discussed
Killourhy	2007	Research in the chosen area is considered 'limited by a lack of realistic, publicly available, real-world data.' (abstract)	Privacy and confidentiality prevent people from exposing sensitive data to researchers. Sanitisation reduces accuracy by introducing artefacts.	Validation of a method of sanitisation that offers higher levels of accuracy (fewer artefacts).	The authors use an established validation method to assess the efficacy of each sanitisation method. The chosen method uses word-tokens that can be used to ensure that sensitive data can be redacted in a way that reduces artefact in the usage models before release to researchers.	Discussion of a goal (estimation of real-world performance) and the effectiveness of their synthetic data. Evaluation of the effects of sanitisation methods.
Mouza	2010	Many methods aim to replace real data with false but realistic data.	Privacy and confidentiality. Requires the creation of semantics based rules sets in order to identify the data that needs to be replaced.	No other model is automated to seek out the sensitive data that needs to be replaced.	The authors model aims to replace real data using a system that detects sensitive data and replaces it using a rules based approach.	No evaluation or validation of the generation method or realism is discussed.
Ngoko, B	2014	"…as long as the model used to synthetically generate the data produces data of similar statistical characteristics to typical measured data."	Historical data at high sampling rates is unavailable for many locations.	Allows for studies of planning and system performance	Uses seed data regarding weather conditions converted to states and markov model generation.	Graph-based statistical comparisons to observed data.
Pudjijono	2009	The published method has been designed to deliver	Privacy and confidentiality. Not easy to create data	The benefits come from being able to repeat experiments	Creates first-state synthetic data using tables of real data, frequencies	There is suggestion that

		'flexible creation of synthetic data with realistic characteristics' (abstract).	which has characteristics similar to real-world data.	with the synthetic generator and share datasets between research teams and approaches without exposing sensitive information.	and dependencies (gleaned from sources like telephone books) as look-up tables. Other attributes like age, sex and SSN are generated by constraint rules. In the second step these are modified further before being stored as duplicate records. The realism comes from the frequencies and dependencies used to model the first state data, and the constraints used to link family data together in the second stage.	validation could consist of verifying the pre and post frequency and distribution of values (both of field values and linkages - i.e. family structures)
Richardson	2008	"This paper presents a thorough and detailed method for generating realistic occupancy data" (abstract)	A lack of synthetic data with occupancy and time use activity data.	Benefit comes from understandings gained from having high resolution energy use data that closely models real Time Use Survey data at greater scale with occupant activity modelling.	The authors use Time Use Survey data from a small number of anonymous real people to model thousands of synthetic energy users.	Verification of generated data points against the statistics from the TUS data used to constrain/seed the generation process.
Sperotto	2009	The data set aims to be realistic (abstract) which the authors define as being 'representative of real traffic and complete from a labelling perspective'. The most 'realistic' traces are those collected in the wild (introduction).	The problem of balance between privacy and realism (Introduction). Network traces from ISPs are rarely published due to the sensitive nature of the material.	Benefit gained from the capture and creation of labelled datasets.	The method consists of capturing network logs of normal and IDS traffic using a honeypot computer, sorting, categorising and labelling.	No validation discussed.
Tsvetovat	2005	The authors state that the 'main concern in generation of artificial data is its realism' (introduction).	Unclear and difficult to define parameters are available to achieve high levels of fidelity in simulation. The use of assumptions from other work.	The ability to more accurately, robustly and repeatability test network analysis algorithms without the use of real data.	The authors recognise that purely random data is not a good approximation of a real phenomenon. They formulate terrorist (social)networks using a formalisation of the way a cellular telephone network operates, where each cell only touches the edge of the next (so where each person knowns how to make contact with one proximal cell).	The few sentences eluding to validation in the introduction are so complicated and poorly written as to be nonsensical.

Van den Bulcke	2006	"… generated networks more closely resemble the characteristics of real transcription networks" (abstract)	Experimental data of the appropriate size and design is not usually available.	Algorithms to infer the structure of gene regulatory networks based on expression data is an important subject. Results are considered by the authors to be extremely scalable and biologically plausible.	Network generator that produces simulated gene expression data based on established topologies from known biological networks rather than relying on simple randomness.	Demonstration by simulation but no validation method shown.
Wan	2008	The resulting six sets of initial ensembles are then analysed in terms of sustainability of the ensemble spread and realism of the correlation patterns (abstract).	Initial ensemble data suffers from being based on random perturbations, or model outputs taken at different times with unwarranted temporal variances.	Their initialisation method allows their model to accept a range of variation (it is not a fixed algorithm method)	Authors populate HYCOM equation with high frequency observation data, using this to make predictions of future measurements. Validation is anticipated to result from using observed data for 2000/01 to generate data for 2002 and then comparing to actual observations for that year.	The authors work does not appear to show any comparison between the experiments and real data for the period tested.
Whiting	2008	Synthetic data has suffered from the criticism that it is unrealistic due to the common approach of using random data that looks like real data (motivation).	Most tools focus on large database testing and not information analytics, the data created is not realistic or believable.	Having known ground truth in the data set provides a baseline for evaluating how an analytic tool supports an analyst in their job.	Primary focus was to come up with a method for creating and evaluating datasets that has realism and is therefore believable. Focus was on network threat generation and detection. Very narrow area of work with most of the work seemingly inapplicable to other SDG areas.	The authors designed a contest to evaluate methods and datasets.
Yu	2003	"Our goal is to more realistically evaluate sensor network system designs" (abstract)	Sensor network research was in its infancy. Authors state that they could not locate prior work on modelling data input in a sensor network context.	Pioneering approach in a new area of synthetic generation research.	Observations were collected from one location and used as the statistical basis for generation of data for unmonitored locations.	Evaluation metrics were defined and plots made comparing observation with synthetic data

Appendix C: A Review of the Kartoun SDG Method

Kartoun, U. (2016). A methodology to generate virtual patient repositories. *arXiv Preprint.* arXiv: 1608.00570

Kartoun's August 2016 paper entitled "A methodology to generate virtual patient repositories" begins by presenting similar justifications for engaging in the generation of synthetic electronic health records (EHRs) as Dube & Gallagher (2014), McLachlan et al (2016) and others; real records come with the risk of patient exposure as they contain private and sensitive information that even data scrambling, redaction or anonymisation techniques fail to mitigate. Kartoun's primary motivation comes through as the need of an EHR dataset that an academic can provide to and use with students in a practical way when teaching bioinformatics/health informatics in the computer sciences (Kartoun, 2016).

In keeping with the incomplete documentation seen in most other SDG articles, very little of Kartoun's actual method or algorithm is presented to the reader. A table of weightings is provided in Supplemental Table 1 that Kartoun describes as the source for population-level preconfiguration during virtual patient generation. A second supplemental table is provided that describes the maximum and minimum values for a selection of medical laboratory blood tests. The source for these values is not given and most appear to be at or within 10% +/- of the normal published reference ranges (Haggstrom, 2009). Kartoun (2016) states that only chief complaints common for both men and women were allowed in the cohort, but that most of the other generated information, number of admissions, selection of the common chief complaint, the laboratory test results and date and time of the admission, release and laboratory test were randomly generated. In any event the two supplemental tables, statement regarding the use of gender-common complaints and the randomness of generation otherwise are the only detail the reader is given on the input and generation method.

Kartoun makes available via a URL three sample repositories that he appears to suggest are foundational to the creation of large databases that *highly resemble real patient records*. The reader must provide a name and email address and on doing so is directed to a download page with compressed zip files containing of 100, 10,000 and 100,000 EHR record sets respectively. The second option (n=10,000) was selected for the purposes of this review.

Given the author's claim that *only chief complaints common for both men and women were allowed in the cohort* it was surprising to find a high degree of gender-specific ailments in the chief complaints data table, including the following sample conditions shown in Table 22:

Table 22: Sample gender-specific conditions from the Kartoun (2016) EMR dataset.

Gender-Specific Condition	Count
Inflammatory disorders of scrotum	19
Benign neoplasm of scrotum	16
Renal failure following incomplete spontaneous abortion	49
Protozoal diseases complicating pregnancy, first trimester	14
Diseases of the digestive system complicating pregnancy, first trimester	16
Pre-existing hypertensive heart disease complicating pregnancy, second trimester	14
Diseases of the digestive system complicating pregnancy, second trimester	18
Diseases of Bartholin's gland	13

While the list in Table 22 constitutes only a small sample, conditions of a gender-specific nature in the dataset included the following key words; pregnancy (n=1524), childbirth (n=557), abortion (n=49), uterus (n=58), ovary (n=202) and scrotum (n=45). While invalidating Kartoun's claim of having only selected gender-common ailments, this would not present as an issue against his claims of success, or indeed realism, were it not for the fact that in many instances these complaints were linked to entirely inappropriate patients. Given that this was a very specific area that could be tested, this component of the input knowledge was used in the validation of his method and the synthetic data that is provided.

A random sample of ten patients were selected from Kartoun's patient dataset. These patients are identified in Table 23. Validation issues observed in the records of these ten patients included; pregnancy and/or childbirth that occurred at unusually advanced ages >49 [Patients 01 and 05]; male patients who were pregnant [Patients 02, 05 and 08]; a birth male who was admitted after the incomplete spontaneous abortion of his pregnancy [Patient 08]; a 75yo birth female who had previously given birth and was now being admitted for an inflammatory condition of her scrotum [Patient 04]; an 84yo birth male suffering a disease of the Bartholin's glands, which for reference lubricate the female vagina during arousal and even in gender reassignment surgery could not be created or transplanted into a previously male person [Patient 07]; a 53yo birth male patient with inflammation of the uterus [Patient 09] and finally; a 24yo birth male patient with cancer of the right ovary [Patient 10].

Table 23: Ten Random Patients from Kartoun (2016)

	Patient ID	Gender	Date of Birth
01	43556DC2-BCFC-45A8-84C3-1D3E4A11B02F	Female	1952-10-06
02	67A4744C-EB52-4AA3-8886-BB29BDED0DC1	Male	1953-03-13
03	6E48CC29-E0B6-4D5F-9D56-7BB45E203223	Female	1958-06-07
04	B6B35809-4552-422C-A62D-E425CA241AAC	Female	1934-04-25
05	4731F86A-9225-424F-9944-AEF4001D0C55	Male	1953-10-07
06	EDA9123E-E39F-420B-8D42-29ACCA6B9375	Female	1930-01-13
07	75862B23-C3EC-44E2-A5C8-6A40EDF0A573	Male	1922-01-11
08	DF598997-15DE-4658-A4D4-48A4ABA08179	Male	1941-03-16
09	6BBD8B7A-BC00-4F33-8139-536C1C90B568	Male	1954-02-19
10	12FCF5A6-B9F7-4AF9-ADD6-D63464AFAF13	Male	1975-04-19

The anatomical components described as part of the condition or complaint leading to admission for Patients 07,09 and 10 are organs that would not exist in someone who was birth male, even if that male had later undergone successful gender reassignment surgery to become female. In total, eight out of the ten randomly selected patients fell to even a common sense level of scrutiny. This left aside other interesting incongruities that gave pause, such as the American woman of Asian ethnicity whose primary language was Icelandic [Patient 01].

Conclusion

Kartoun's synthetically generated patient data is provided free firstly as part of his educating information science students in health informatics, but secondly to the general public. Limited detail is available in Kartoun's (2016) unpublished paper that could lead to an understanding of how these records are generated, except that he claims that his method is to be considered as a credible teaching aid and base for future realistic data generation. If 80% of a random selection from this generated EHR dataset fail basic validation using the limited input knowledge that we were provided by the author, then it can only be concluded that the entire dataset and the method used to create it cannot be relied upon even to be, as the author describes; *a foundation to further generate large, longitudinal artificial EMR databases that highly resemble real patient records*.

Appendix D: A Review of the Synthea SDG Method

Mitre. (2016). About SyntheticMass. Sourced from: https://syntheticmass.mitre.org/about.html

The Synthea application was created by a team from The Mitre Corporation, a US-based research and development organisation. Synthea forms the backbone of their SyntheticMass project that aims to create a realistic dataset comparable to that of the 7 million real health consumers of Massachusetts (Mitre, 2016).

For this validation review the Mitre team that developed Synthea provided a small group of documentation resources (n=3) and a table of clinical factors and observations that were used in the initial development of their type-2 diabetes generator module. They also provided a sample collection of synthetic patients generated using Synthea's type-2 diabetes module (n=1620). An older version of the Synthea source is publicly available on Github. The current updated code was not provided. A list of the documents provided is included in Table 24. Each of these resources concerns the disease modelling and decision making approach necessary to model diabetes. The Garber (2016) paper is a set of PowerPoint slides that delve into the various components that would be necessary to an algorithm that can make decisions about the treatment regimen for diabetic patients. Toussi et al's (2009) paper provides knowledge in the form of twenty-seven treatment rules developed from a review of almost 500 type-2 diabetic patient health records. Given that no documentation of the statistical basis for generation was provided, additional knowledge for logical and statistical validation of the patient cohort was developed by this author from publicly available health statistics. The statistics gathered were representative of both the state of Massachusetts, and the United States in general. These additional knowledge sources are listed in Table 25.

This review will look at key knowledge drawn from both source tables, using it within the validation framework described in this thesis to assess the correctness and appropriateness of the Synthea application. It begins with assessment of the qualitative and quantitative qualities of the synthetic data. Only if these are sound should validation proceed to knowledge discovery and assessment.

Table 24: Documents provided by the Synthea Team

Barhak, J., Isaman, D., Ye, W., & Lee, D. (2010). Chronic Disease Modelling and Simulation Software. *J. Biomed. Informatics*, 43(5).
Garber, A. (2016). AACE/ACE Comprehensive Type2 Diabetes Management Algorithm. *American College of Endocrinology.*
Toussi, M., Lamy, J., Le Toumelin, P., & Venot, A. (2009). Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: Methods and example for type2 diabetes. *BMC Medical Informatics and Decision Making*. 9(28).

Table 25: Additional Sources for Type2 Diabetes Validation Data

Source	Title	URL
A	CDC National Diabetes Report	https://www.cdc.gov/diabetes/data/statistics/2014statisticsreport.html
B	Mass. Health and Human Services	http://www.mass.gov/eohhs/gov/departments/dph/programs/community-health/diabetes/facts/diabetes-statistics.html
C	CDC Data Sources, Methods & References	https://www.cdc.gov/diabetes/data/statistics/2014statisticsreport.html
D	Type2 Diabetes in MA	Goldoftas, B. (2014). Type2 Diabetes in Massachusetts: A population perspective and its importance for public policy.
E	Fact Sheet: Massachusetts	http://www.amputee-coalition.org/resources/massachusetts/
F	Diabetic Foot Complications	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4422657/
G	Long-term survival end stage renal disease	Beladi-Mousavi, S. S., Alemzadeh-Ansari, M. J., Alemzadeh-Ansari, M. H., & Beladi-Mousavi, M. (2012). Long-term survival of patients with end-stage renal disease on maintenance hemodialysis: a multicenter study in Iran. *Iranian Journal of Kidney Diseases*, *6*(6), 452.
H	Changes in age at diagnosis of Type2	Koopman, R., Arch, G., Diaz, V., & Geesey, M. (2005). Changes in age at diagnosis of type 2 diabetes mellitus in the United States, 1988-2000. *The Annals of Family Medicine*, 3(1).
I	Facts about Diabetes in MA	http://www.mass.gov/eohhs/docs/dph/com-health/diabetes/facts-in-mass.pdf
J	Heart disease and stroke statistics	Go, A., Mozaffarian, D., et al (2013) Heart Disease and Stroke Statistics - 2014 Update: A report from the American Heart Association.
K	Global trends in the incidence of type-2	Farsani, S., Van der Aa, M., Van der Vorst, M., Knibbe, C., & De Boer, A. (2013). Global trends in the incidence and prevalence of type 2 diabetes in children and adolescents: A systematic review and evaluation of methodological approaches. *Diabetologia*, 56(1471).

Prevalence

Prevalence of type-2 diabetes in Massachusetts tends generally to be close to that of the entire United States. Massachusetts Government statistics show only a slight increase from 7.0% in 2008 to 7.2 in 2010 (Source B) during a time when the national CDC figures were stable at 8.3% (Source A). This is in contrast to Synthea where prevalence is 76.79% (1244 of 1620 patients).

While white non-Hispanics are least likely to succumb to type2 diabetes both in MA (Source B) and the wider US (Source A), in Synthea they are statistically the second-most prevalent racial type to receive the diagnosis. Another difference between Synthea and observed statistics comes when we look at the prevalence in Black non-Hispanics. In MA and the wider US, prevalence in this group is significantly higher than that observed in the white population, close to double in both cases. In Synthea this is reversed with Black non-Hispanics seeing much lower prevalence rates when compared to white non-Hispanics. Synthea does appear to represent the spirit of the relationship between Black non-Hispanics,

Asians and Hispanics similar to that observed in MA, as shown in Figure 24, however the overall incidence of diabetes in the Synthea population is around 70% higher than real world observations.

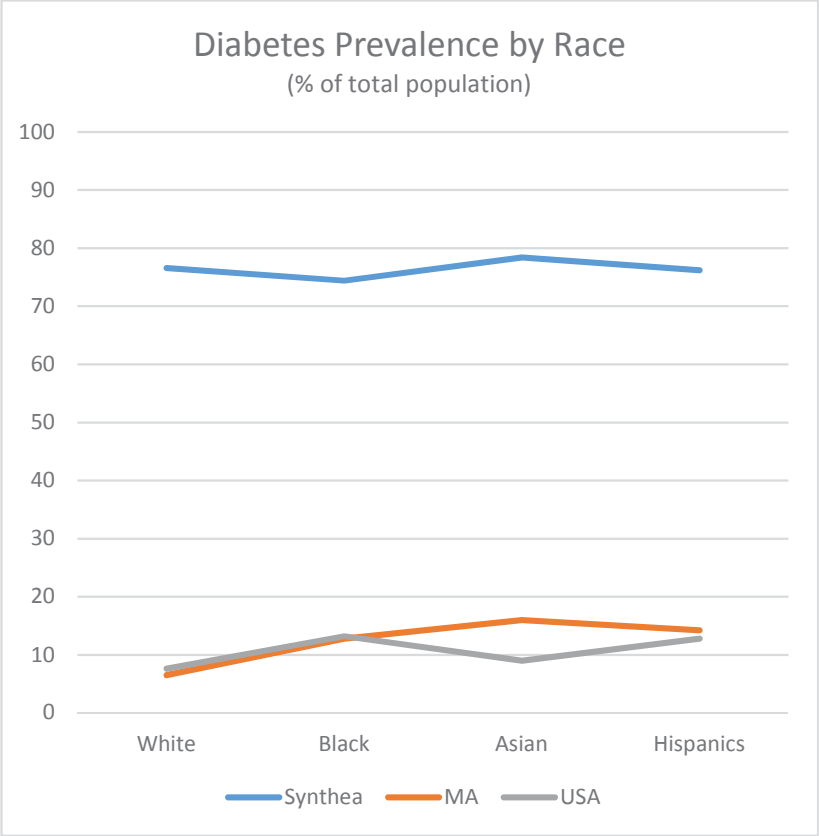


Figure 23: Synthea Validation Review: Diabetes Prevalence

Amputation in Diabetes

Amputations resulting from type-2 diabetes in the United States occur in 0.25% of the diabetic population (Source C). In Massachusetts, lower limb amputation occurs in 0.65% of the diabetic population, with upper limb amputations occurring in an even fewer 0.034% of cases (Source E). Diabetes related foot amputations occur nationally in 1.06% of diabetics annually, however this amount includes those for both type 1 and type 2 diabetes (Source F).

In Synthea, 100% of those patients diagnosed with type-2 diabetes also show at least one amputation event. Lower limb amputations occur in 83.4% of Synthea's diabetic patients with upper limb amputations occurring in 81.19%. A large percentage of the Synthea population (n= 64.54%) receive multiple amputations, including at least one of each type. Synthea codes foot amputations separate to general lower limb amputations, with 83.44% of diabetics showing the SNOMED code for foot amputation as part of their medical history.

Synthea's statistics for amputations are far in excess of both State and national averages and no correlation could be identified between the synthetic and real world statistical outcomes.

Kidney Failure

Diabetes accounts for 44% of all kidney failure diagnoses in the US, affecting 0.17% of the national diabetic population (Source A and C). 0.79% of diabetics are either on regular dialysis or are receiving a kidney transplant annually (Source A and C). Even when treated with regular dialysis, End Stage Renal Disease (ESRD) in diabetes comes with a prognosis of death within ten years (Source G).

In Synthea, 87.06% of diabetic patients are diagnosed with ESRD, a technical description for kidney failure. This is a condition that necessitates immediate and ongoing dialysis and kidney transplant else death rapidly occurs. None of the synthetic patients are seen to receive dialysis treatment yet those diagnosed with ESRD tend to live in excess of ten years, with patients that have lived between 15 and 21 years beyond the diagnosis regularly observed. Such longevity is contraindicated by published research (Source G).

Age at Diagnosis

Type-2 Diabetes Mellitus is generally diagnosed in adults ≥40 years of age (Source J). The diabetes generally found in very young children under 10 years of age is type-1 (Source J). A recent systematic review showed that across the US reported incidence rates for type-2 diabetes in the <10 age group were 1.1 in 100,000 or less, depending on ethnicity (Source K). The mean age at diagnosis of type-2 diabetes in the US is 46 (Source H). While the national average for type-2 diabetes in the 10-19 age group is 0.01% (Source C), MA reports a higher than average 0.06% (Source I). The MA Department of Public Health report that they do not presently have accurate statistics for the incidence of any diabetes in the <10 age group, but state that clinicians believe incidence is increasing (Source I).

The age at diagnosis of the Synthea type-2 diabetic patients has been plotted and compared with the state and national averages, and is shown in Figure 25. While Synthea manages to weigh the peak of adult diagnosis around 46-47 years, similar to the national average of 46, there are several significant issues with the Synthea patient cohort.

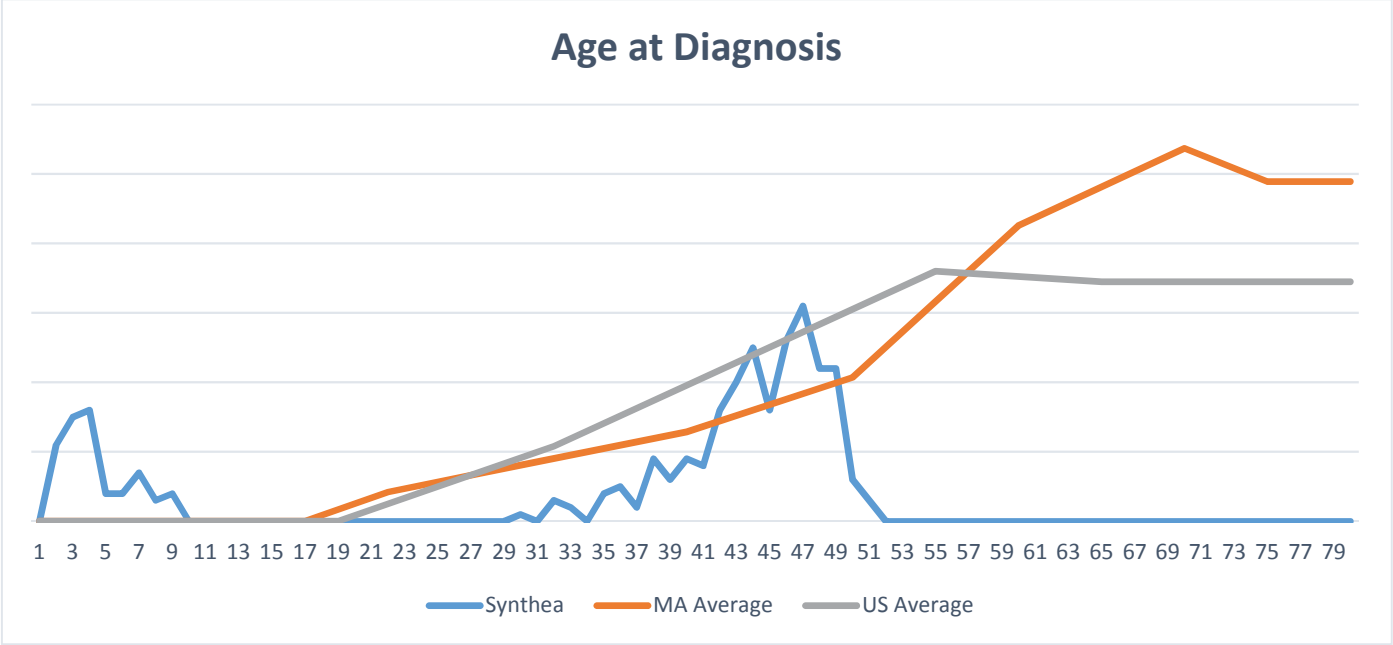


Figure 24: Age at Diagnosis of Type-2 Diabetes Mellitus

The first issue is that no Synthea patient over the age of 52 is ever diagnosed with type-2 diabetes. Neither the state nor national statistics show a reduction in diagnoses for this age group. In both cases we see increases and peaks in the overall number of diagnoses beyond the point where Synthea has ceased generating diagnoses.

The second issue is similar in that the 11-28 age group is also underrepresented by Synthea.

The third issue comes from the overrepresentation of diagnoses in the <10 age group, and more specifically, in those patients who are ≤5. Given that we have already seen that diagnosis of type-2 diabetes is quite rare in children, more so in the very young, it is difficult to understand Synthea's propensity for generating almost a quarter of all type 2 diagnoses in this age group.

The model Synthea presently uses for generating diagnoses requires additional training in order to correctly represent across the age distribution seen in publicly available health statistics.

Qualitative Items

Leaving aside the more obvious quantitative issues already discussed, there were a number of obvious qualitative issues that should be addressed. **Firstly**; it was noted that diabetic patients were only seen by a physician in the outpatient clinic of a hospital. These outpatient attendances always last for exactly fifteen minutes, even when the patient is displaying some challenging or complicated symptomology that in a real patient would necessitate urgent medical admission and treatment. **Secondly**; in many of these outpatient appointments potentially life threatening observations and blood tests go ignored. In

many cases the blood pressure generated for the synthetic patient was at levels suggesting hypertensive crisis, a condition that when left untreated rapidly results in heart and kidney failure, stroke and death. **Thirdly**; it seems incredible that every diabetic patient appears to predominately attend the outpatient clinic, most often late at night and very early in the morning, to receive immunisations. This occurred in every outpatient interaction except in some rare **fourth and final** instances where the patient somehow undergoes the amputation of a foot or hand during their fifteen-minute outpatient visit. This amputation is conducted with neither the need for admission or anaesthetics. No post-amputation follow-up and rehabilitation is recorded in the patient history and rather than being given the gravity and due concern that such a disfiguring procedure warrants, is treated by Synthea in much the same way as one might record prescribing creams to treat foot fungus.

Conclusion

While qualitative and quantitative analysis shows that the Synthea synthetic patient data bears very little similarity to publicly available knowledge and statistics, it should be noted that the version reviewed here was an early otherwise untested version. The prevalence of type 2 diabetes in the Synthea cohort is around 70% greater than the Massachusetts average. Amputations occur at a rate that is almost 4000 times greater than the state average, and Synthea patients are almost 1000 times more likely to undergo a foot amputation than the national average. Synthea patients are 110 times more likely to be diagnosed with kidney failure, never receive dialysis and more often live two to three times longer than diabetic patients with kidney failure observed in published studies. The distribution of diagnosis age shows that Synthea tends towards two peaks, one in infant children and the other in middle aged adults. Only the peak in 46-year-old adults is supported by the literature. The spread of diagnosis ages completely disregards two very large population groups that are significant in the observed statistics, the 10-29 year olds and those over 52 years of age. Being as the Synthea type-2 diabetes model fails the first two validation steps, the model should be reviewed and updated prior to continuing validation using the knowledge discovery approach described in this research.

Epilogue

The MITRE team used the findings of this review to refine the early sample of the Synthea Diabetes module. The refinement included the incorporation of more recent statistics, the sourcing of statistics to ensure correct probability distribution of a number of comorbid conditions that the developers had left to random chance and the correction of a number of assumptions that the developers had made regarding progression of the diabetes disease.

The diabetes module's greatly improved realism is the subject of a manuscript submitted to JAMIA entitled "*Synthea: An approach, method and software mechanism for generating synthetic patients and the synthetic electronic healthcare record*".