







Realistic Synthetic Data Generation: The ATEN Framework

Scott McLachlan¹ , Kudakwashe Dube² ,
Thomas Gallagher³ , Jennifer A. Simmonds⁴,
and Norman Fenton¹ 

¹ Queen Mary, University of London, London, UK
s.mclachlan@qmul.ac.uk

² Massey University, Palmerston North, New Zealand

³ Missoula College, University of Montana, Missoula, MT, USA

⁴ NSW Health, Sydney, NSW, Australia

Abstract. Getting access to real medical data for research is notoriously difficult. Even when data exist they are usually incomplete and subject to restrictions due to confidentiality and privacy. *Synthetic data* (SD) are best replacements for real data but must be *verifiably realistic*. There is little or no investigation into systematically achieving *realism* in SD. This work investigates this problem, and contributes the *ATEN framework*, which incorporates three component approaches: (1) THOTH for synthetic data generation (SDG); (2) RA for characterising realism in SD, and (3) HORUS for validating realism in SD. The framework is found promising after its use in generating the *realistic synthetic EHR* (RS-EHR) for labour and birth. This framework is significant in guaranteeing realism in SDG projects. Future efforts focus on further validation of ATEN in a controlled multi-stream SDG process.

Keywords: Synthetic data generation · Knowledge discovery

1 Introduction

The McGraw-Hill dictionary of Scientific and Technical Terms describes Synthetic Data as *any production data applicable to a given situation that are not obtained by direct measurement* [1]. Prior to [2] the domain of statistics, especially population statistics, primarily viewed synthetic data to be larger datasets that result from merging two or more smaller datasets [3, 4]. The earliest direct reference to synthetic data is a 1971 article describing creation of tables of synthetic data for use in testing, modifying, and solving problems with marketing data [5]. Other works present methods for creating fully synthetic data based on observed statistics [6, 7]; predicting and testing observational outcomes [8]; and generation driven by probability models for use in simulations [9]; and forecasting [10]. The *reasons for generating synthetic data* include software testing [11–14], population synthesis [15], hypotheses testing or generation of seed data for simulations [16, 17]. Recently, the major reason for generating synthetic data is limiting the release of confidential or personally identifiable information inherent to the use of real data sources [13, 18–20]. Some *synthetic data generation*

(SDG) approaches use *real data* either directly, or as seed data in their SDG methods [11, 21, 22]. Caution should be used prior to release of such synthetic datasets as a poorly designed or inappropriate model can still carry the risk of exposing confidential or personally identifiable information. Most contemporary research works have focused heavily on data *anonymization*, that is, isolating and replacing personally identifiable data with the concomitant goal of maintaining integrity of the data that an organisation may wish, or be required, to publish [23]. Anonymization has been dogged by modern methods for *re-identification of anonymised data* using a person's linkages to publicly available personal information sources, such as the electoral roll and newspaper articles [24–26]. As a result, some SDG methods also risk suffering inverse methods and re-identification attacks that ultimately breach personal privacy.

It is not enough to generate random data and hope it will be suitable to the purpose for which it will be used [27]. The data values may be required to fall within a defined set of constraints. For example, the *heart rate* should be a numerical value that falls within healthy resting (60–100), exercising (100–160) or disease state (40–60 or 160+) ranges. Some projects require increasingly more complicated datasets where not only the values of single attributes must be valid, but all values and interrelationships must be indistinguishable from observed data [28, 29]. *This is where the problem of realism becomes imperative*, yet it remains unexplored in current SDG literature [30]. The common sense implication of the term *realistic* is as [31] succinctly puts it: synthetic data that becomes “*sufficient to replace real data*”. The property of realism brings a greater degree of accuracy, reliability, effectiveness, credibility and validity [22]. Most researchers recognise the need for realism [18, 22, 31], however many leave realism unexplored in their works with only two authors giving some attention to it [18, 19]. In both cases this was vague and limited only to hinting that the aim of realism was that the synthetic data should be a representative replacement for real data [19], and comparably correct in size and distribution [18]. Neither handled validation of realism in the synthetic data they created. The lack of research attention makes it difficult to imbue realism into SDG methods, and to verify success in doing so. *Realism should only be asserted if it has been verified* [32, 33]. Scientific endeavours should always be concerned with testing and verification, yet few published approaches present systematic ways for validation [34, 35]. We find many SDG methods that claim success in the absence of a systematic ways of scientific validation [12, 36–38]. Some form of validation is necessary to support claims for realism in resulting synthetic data [32, 38, 39]. Otherwise, reliability of the approach must be questioned [40]. This work addressed these challenges and hereby presents the ATEN framework that allows realism to be inherent in SDG methods while also incorporating validation of realism in the resulting synthetic data.

The rest of this chapter is organised as follows: First, a review of related works focusing on SDG methods and realism is presented. Second, the ATEN framework and its component approaches, namely, THOTH, RA and HORUS, are covered in detail. Third, the ATEN framework is evaluated by applying it to the case of generating the synthetic electronic healthcare record (EHR) for labour and births. Fourth and finally, the chapter is concluded and summarised.

2 Related Works

A literature search was conducted to identify works describing methods or approaches for synthetic data generation ($n = 7,746$). This collection was reduced to works that also used the terms realistic ($n = 290$) or realism ($n = 6$) in describing either the need or purpose for synthetic data, their method, or the resulting synthetic dataset. The resulting collection included works that identified realism as a primary concern in the generation of synthetic data generally [12, 22, 41], or that discussed developing synthetic data that would be sufficient to replace, or be representative of, real data [13, 19, 31, 42]. Due to the low number of works that identified realism as a factor in synthetic data, a random selection of excluded works was included. This review found that one third of SDG articles focused on common goals, namely, *authenticity* [11], *accuracy* with respect to real structures [21], and the *replacement of real data* [43]. A key observation is the conspicuous absence, in the literature, of an investigation of **realism** for synthetic data, along with the lack of rigorous explanation of the approaches used to produce what authors claim to have been realistic datasets. *In the absence of a clear definition and framework for realism in the context of SDG, any process seeking to verify and validate realism in synthetic data is severely challenged.*

Works in the literature present common narrative for describing their SDG problem justification, operational method, and claimed results. This narrative consists of a common sequence of themes, each presented with two components. The themes are presented in Table 1. For the justification theme, research challenges include limited available data [44, 45] and privacy protection [37, 43]. Uses include testing of learning algorithms [45], enabling release of data [43], and prediction [37]. The operation theme includes SDG inputs such as network structures [45], observational statistics [44], and configuration files [37]. Methods ranged from random selection [45] and change behaviour modelling [37], to stochastic simulation using Markov models [44]. The result theme covers actions such as the use of benchmark and performance test simulation [45], comparative graphs [44], and performance analysis [46] used to assess published SDG methods. Resemblance to real networks [45], model advantages and capabilities [44] and likeness of the synthetic data to the synthetic scenario [37] were all reasons claimed by authors for claiming their SDG method was promising or successful.

Table 1. The common SDG narrative.

Narrative themes	Narrative components
Justification	It is difficult because of [<i>some difficulty</i>] to get real data for [<i>some use</i>], so we developed a new method to generate synthetic data for this purpose
Operation	Our method uses [<i>some input</i>] to generate the synthetic data using [<i>some method</i>]
Result	We performed [<i>some action</i>] and believe that the synthetic data created by our method is promising for [<i>some reason</i>]

SDG approaches set the goal of simply producing synthetic data that is a suitable replacement for real data. The focus is heavily weighted toward the outcome, the synthetic data. **Validation** of realistic aspects of synthetic data tended to be absent or singular or simplistic, ranging from direct comparisons between either the entire dataset or fields within the synthetic data to observations drawn from the real data [22], or graphical and statistical comparisons between the two [21, 44, 47]. The majority did not discuss validation at all [36, 48, 49]. Disclosure of the validation approach in research work completes and improves understandability of their work. It would also allow researchers to adequately assess whether or not a project met its goal; and the success claimed is truly justified [50]. This characteristic ensures that SDG experiments can be independently verified to the same standard as other scientific endeavours.

3 ATEN: The Framework for Realistic Synthetic Data Generation

It is common to see methodologies with multiple separate, combined, or sequential components presented as a framework [51]. This section presents the ATEN framework shown in Fig. 1. The ATEN framework is a synthesis of three interdependent component approaches, THOTH, RA, and HORUS which, when used together infuse realism into synthetic data. Each component of the ATEN framework seeks to answer the related questions in Table 2. The sections that follow describe in detail each of the components of ATEN.

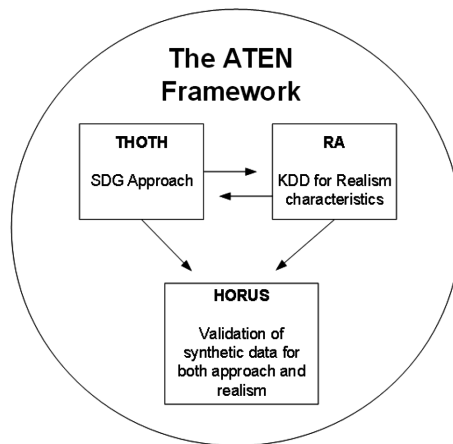


Fig. 1. The ATEN framework [52].

Table 2. ATEN component aims.

THOTH	How should we generate this synthetic data?
RA	What knowledge is necessary to achieve realism in the synthetic data?
HORUS	When THOTH operates using RA’s knowledge, was realism achieved in the resulting synthetic data?

3.1 THOTH: The Enhanced Generic Approach to SDG

A review of the way authors described data generation approaches yielded a generic four-step SDG approach, which incorporates the minimum common structural elements shared by all SDG methods. The approach is presented as a waterfall model, primarily due to its cumulative and sequential nature. Thus, the next phase is undertaken solely through completion of the previous [53]. Verification, a required step of any scientific endeavour but one rarely seen in the context of SDG, can only occur during limited opportunities at the end of each step of the approach [53] and after the SDG operation is complete. The following paragraphs present the four-step SDG approach.

1. *Identify the need for synthetic data:* This step involves recognising both the need and justification, or reason, for creating synthetic data. The most commonly expressed justification across the contemporary literature was that the synthetic data being created was necessary to replace real data containing personally identifiable, sensitive or confidential information.
2. *Knowledge gathering:* This step 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 generated, identifying such things as necessary fields to be generated, the scope, and any constraints or rules to be imposed.
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 has drawn significant focused attention or is considered more reliable to producing a particular outcome. Many of these algorithms have operational steps or processes requiring 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 relationship to the quality of the output.
4. *Generate the synthetic data:* The process of generation involves presenting any seed data, conditional requirements, rules, and constraints to the generation algorithm that will perform the processes that output synthetic data.

This four-step approach represents a simple method, which are favoured due to its usefulness, reduced complexity, and experiment time; all of which reduces cost [54–57]. However, the approach suffers the waterfall model weakness; flowing unidirectionally, lacking flexibility, meaning any change in requirements or issues identified necessitate expensive and time consuming redevelopment and retesting [58]. For this reason, a more adaptable and agile approach to SDG development should be encouraged. Pre-planning and preparation may mitigate the weaknesses of the generic SDG waterfall model. This is where THOTH will assist. THOTH encourages the synthetic

data creator to perform decisive steps prior to engaging in the generation process. THOTH begins with characterisation, that is, identifying the level of synthetic-ness desired in the data to be generated. The synthetic-ness required of generated data can range from anonymisation of personally identifiable components in real data, through to truly synthetic data relying on no personally identifiable information during the creation process. The five primary characterisation types are shown in Table 3.

The characterisation level provides an element that aids in the second step, selection of the classification, or generation model, from the following five categories of synthetic generation methods: (i) data masking models that replace personally identifiable data fields with generated, constrained synthetic data [13, 43, 59], (ii) those that embed synthetic target data into recorded user data in a method known as Signal and Noise [11, 18, 60], (iii) Network Generation approaches that deliver relational or structured data [21, 41, 45], (iv) truly random data generation approaches like the Music Box Model [61], and (v) probability weighted random generation models like the Monte Carlo [12], Markov chain [61], and Walkers Alias methods [62].

Table 3. Characteristics of synthetic data.

Truly synthetic data	Data generated where no confidential or sensitive data has been directly used. Approaches may rely on algorithms that populate a dataset with generic seed data based on statistical probability, or acute randomness. An example of Truly Synthetic Data can be seen in CoMSER [62]
Fully synthetic data	Data generated using real data in the knowledge discovery (pre-generation) phase, but where no real data carries across into the synthetic dataset. Examples include capturing and breaking up real-world data into elemental components, rebuilding these into entirely new rows of data. Another uses the real data to construct a database architecture, populating that database with synthetic data based on observation [12]
Partially synthetic data	Datasets containing some form of synthetic data intermixed or aggregated with unaltered real data. An example is the Outbreak-Detection system using simulated ‘signals’ superimposed on real background ‘noise’ [63]
Anonymised-only data	Projects that identify and replace, remove, or scramble sensitive fields within a dataset, leaving the remaining fields unchanged
Real data	Real or observed data in which no attempt has been made to anonymise, conceal or synthesise any values

When combined with the generic SDG approach discussed earlier, the resulting THOTH-enhanced generic approach is shown in Fig. 2. With these steps complete, the synthetic data creator engages the remaining steps from the generic SDG approach described previously. However, they are beginning with an additional level of wisdom that comes from knowing where they are going (the level of synthetic-ness required of their efforts) and the framework for how they are going to get there (the informed selection of a generation model).

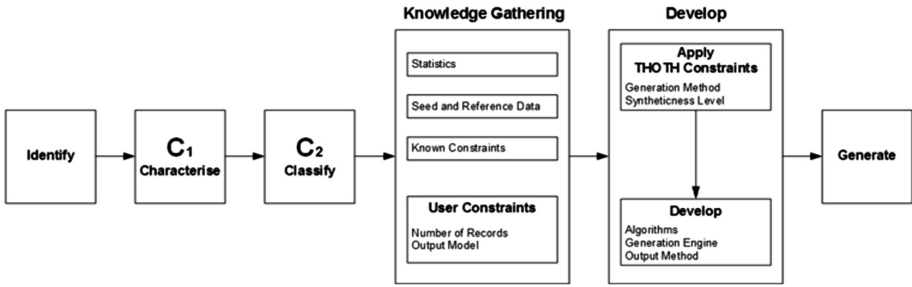


Fig. 2. THOTH integrated into the generic approach to SDG [52].

Summary of THOTH: We found a generic four-step waterfall approach is common to most SDG methods. This approach moves through identifying a need for synthetic data, gathering knowledge necessary to its generation, developing or customising an algorithm or generation method common to their domain or solution needs, before generating the synthetic data. Incorporation of THOTH benefits the researcher, providing greater awareness of their requirements and guiding the direction of the overall synthetic data generation approach.

3.2 RA: Characterising Realism for SDG

RA provides a structured approach to identifying and characterising realism elements, or knowledge, for use in SDG. The RA process, including the steps of enhanced knowledge discovery, are shown in Fig. 3 and described in Table 4. RA identifies extrinsic and intrinsic knowledge following a logical progression of steps, with increased focus on elements drawn from [64–67]. The following subsections present the processes used within the KDD data mining in Step 5 of Table 4.

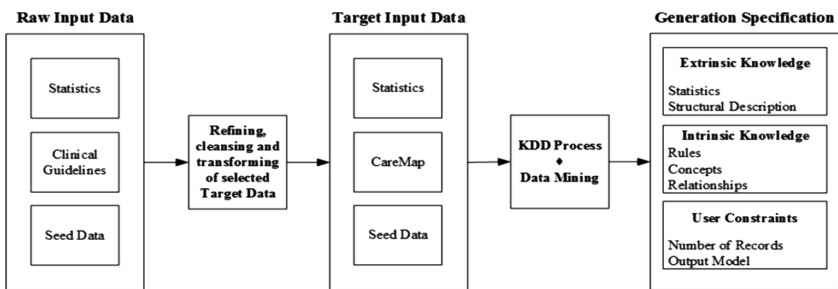


Fig. 3. Overview of the RA approach to realism in SDG [52].

Table 4. Enhanced KDD process following the RA approach [52].

Step	Activity	Tasks
1	Develop and document information (overlaps with THOTH)	Relevant prior knowledge Understanding of application domain, and Goal(s) of KDD process
2	Collect raw data (overlaps with THOTH)	Selecting relevant datasets on which discovery is to be performed
3	Refining and cleansing of raw data	Cleanse and pre-process data to eliminate noise, and Remove incomplete or inconsistent data
4	Create target data	Integrate data from multiple sources Transform raw data Project data by identifying useful features for representing the data, and Reduce variables to those that are necessary for KDD process
5	KDD and data mining	Identify data mining method to search for patterns within the target data (summarisation, classification, regression, clustering, web mining and others as described in Fayyad et al. [64]) Perform concept hierarchy analysis, formal concept analysis, rule identification methods used in HORUS
6	Interpret and evaluate mined patterns	Identify truly interesting and useful patterns
7	Presentation	Make knowledge available for use in synthetic data generation

RA: Extrinsic Knowledge

Extrinsic knowledge is the *sum of quantitative and qualitative properties* found in the real data to be synthesised. To be a suitable replacement, the synthetic data will need to adhere to these properties.

Quantitative Characteristics: Real or observed data may in itself be statistical, and therefore quantitative, such as patient demographic data shown later in Fig. 10. Even if it is not, it is often possible to identify quantitative knowledge, for example; consider generating a synthetic version of a spreadsheet of people who voted at a selection of polling booths, 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: (a) how many people of each genealogical nationality voted in (b) each hour, (c) the percentage that were male, (d) the percentage of the overall population as found in census data voted in each polling booth, and so on.

Qualitative Characteristics: The qualitative characteristics of real or observational data should be identified and documented for any SDG project, but especially for those

projects seeking realistic synthetic data. One example of qualitative characteristics may be to identify and describe the database schema. The database schema explains how the data is structured [68]. 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 [68].

Summary: Extrinsic Knowledge: These quantitative and qualitative observations of real data, once identified and documented, represent the characteristics that should be created and validated in synthetic data. This is especially true if authors present that there is a requirement for, or claim of, realism.

RA: Intrinsic Knowledge

Knowledge Discovery in Databases: 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 [64, 66, 67]. Modern approaches, where the human is accentuated by machine learning or neural network algorithms are considered more expedient for realising insights from today's extremely large datasets [64, 66, 67].

Concept Hierarchies: Concept Hierarchies (CH) are a deduction of attribute-oriented quantitative rules drawn from large to very large datasets [69]. CH allow the researcher to infer general rules from a taxonomy, structured as general-to-specific hierarchical trees of relevant terms and phrases. For example: "bed in ward in hospital in health provider in health district" [67, 69, 70]. Developing a concept hierarchy involves organizing levels of concepts identified within the data into a structured taxonomy, reducing candidate rules to formulas with a particular vocabulary [69]. CH are used by RA to identify an entity type, the instances of that entity and how they relate to each other; they help to ensure identification of important relationships in the data that can be used to synthesise meaningful results [71].

Once the concept hierarchy tree is identified, a second pass across the source data is performed to provide an occurrence count for each concept. This second pass allows the researcher to enhance the concept hierarchy with statistical knowledge to improve accuracy of the generation model.

Formal Concept Analysis: Formal Concept Analysis (FCA) is a method of representing information that allows the researcher to easily realise concepts observed recognised from instances of relationships between objects and attributes. For example: occurrences of different nosocomial infections across the wards of a hospital. FCA starts with a formal context represented as a triple, where an object {G} and attribute {M} are shown with their incidence or relationship {I} [72]. A table is created displaying instances where a relationship exists between the object and its corresponding attribute(s).

Concept creation, represented as rules, occurs from the context table. For example, one might seek to identify the smallest or largest concept structures containing one particular object.

The second step to FCA involves creating the concept lattice. A concept lattice is a mapping of the formal context, or intersections of objects and attributes. The concept

lattice allows easy identification of sets of objects with common attributes as well as the order of specialisation of objects with respect to their attributes [73].

Characteristic and Classification Rules: [69] provides a set of strategies that can be used to learn characteristic and classification rules from within a dataset. These rules can be applied as constraints during generation, and later as tools to compare against the resulting synthetic data to validate its accuracy and realism.

Characterisation Rules: The development of characteristic rules entails three steps. First, data relevant to the learning process is collected. All non-primitive data should be mapped to the primitive data using the concept hierarchy trees as shown in Fig. 5 (e.g. Forceps would be mapped to Assisted, Elective would map to Caesarean and so on). Second, generalization should be performed on components to minimize the number of concepts and attributes to only those necessary for the rule we are working to create. In this way, the *name* attribute on a patient record would be considered too general and not characteristic to a set of data from which we could make rules about the treatment outcomes for a particular ethnicity. The final step transforms the resulting generalization into a logical formula that identifies rules within the data. These rules are the sum of four elements, where if the values of any three of those elements are found to be consistent to the rule for a given instance in the dataset, the fourth element will always be true.

Classification Rules: Classification knowledge discovery discriminates the concepts of a target class from those of a contrasting class. This provides weightings for the occurrence of a set of attributes for the target class in the source dataset, and accounts for occurrences of attributes that 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. Attributes specific to a target class form the basis of classification rules.

RA: Summary

The RA enhanced and extended KDD method identifies realistic properties from real data, providing improved input data quality and constraints that improve the output of generation algorithms used to create synthetic data. An obvious benefit is that generation methods using this knowledge should deliver data that is an accurate replacement for real data. Another benefit is a set of knowledge and conditions that can be used in validation of realism in the data created. Its use for this last purpose is discussed in the next section.

3.3 HORUS: An Approach to Validating Realism

One of ancient Egypt's earliest precursor national gods, Horus, was revered as the god of the sky; that which contains both the sun and the moon. In the same way, the Horus approach to realism validation draws on both THOTH's enhanced generic SDG and RA's enhanced KDD approaches, effectively containing both the sun and moon as a means to validate for realism in synthetic data.

The validation approach incorporates five steps that analyse separate elements of the SDG method and resulting synthetic data. These steps are identified as the smaller square boxes in Fig. 4, with their descriptions below. Collectively, the five steps provide the information necessary for confirmation of whether synthetic data is consistent with and compares realistically to real data that the SDG model seeks to emulate.

Input Validation: Input validation concerns itself only with that knowledge coming from the generation specification in the form of data tables and statistics. The input validation process verifies each item, confirming that the right input data in the correct form is being presented to the generation engine, thus ensuring smooth operation of the data synthesis process [74]. Input validation is intended to prevent corruption of the SDG process [75].

Realism Validation 1: The first realism validation process verifies concepts and rules derived from the HCI-KDD process, along with any statistical knowledge that has been applied. Realism validation reviews and tests both the premise and accuracy of each rule to ensure consistency with the semantics of any data or guidelines used in their creation.

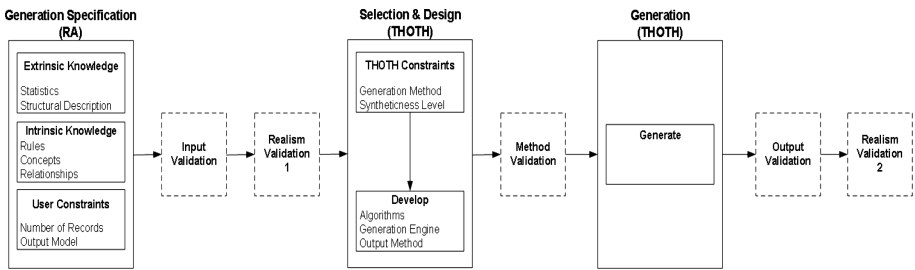


Fig. 4. HORUS approach to realism validation, showing touch points with THOTH and RA [52].

Method Validation: Method validation reviews the efforts of others inside and outside of the research domain. Attention is drawn to methodological approaches common for that domain, as well as methods other domains have employed for similar types of SDG. Evaluating the entire scope of method application ensures that which is chosen should be the most appropriate for the particular need and solution. Method validation also seeks to verify that the algorithm being used is correctly and completely constructed, and free of obvious defect [76].

Validation is not a search for absolute truth, more correctly, and in this instance, it is a search to establish legitimacy [76]. Table 5 provides the six key questions that should be asked of any SDG methodology the researcher may propose to use.

Table 5. Method validation questions [77].

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?

Output Validation: Output validation evaluates the output data and verifies its basic statistical content. This step demonstrates the difference between the two terms: validation and verification. Validation ensures the model is free from known or detectable flaws and is internally consistent [76]. Verification establishes whether the output, or predictions, of the SDG model are consistent with observational data. The output validation step ensures that the synthetically generated data conforms to the quantitative and qualitative aspects derived during the knowledge discovery phase.

Realism Validation 2: The second realism validation process undertakes the same tests as the first, except that tests are now performed against the synthetic dataset. This test aims to ensure synthetic data is consistent with the knowledge (rules, constraints and concepts) previously derived from the input data and used in creation of the synthetic data. The second realism validation step is the most important for establishing, and justifying, any claim that this synthetic data presents as a realistic and proper substitute for the real data it was created to replace.

3.4 Summary: Benefits of the ATEN Framework

There are a number of ways that ATEN benefits those engaging in SDG. First, it is a complete SDG lifecycle that considers every element before, during and after data generation. Second, it encourages more complete level of self-documentation than most presented in the SDG literature. The third benefit is cumulative from the first two, in that when applied during an SDG project, THOTH and RA provide the necessary knowledge to validate realism using HORUS. ATEN supports claims of success, realism, and enables repeatability. All of which are fundamental to the scientific method. Works found in the literature do not conform to the ATEN Framework, as significant gaps are evident in most SDG literature. The framework provides for additional knowledge discovery and documentation processes, which could be automated. However, this is dependent on the type of data being analysed, generation method, synthetic data sought, and the use to which that data will be applied. The knowledge discovery component leads to greater accuracy and help to support validation of realism.

4 Evaluating the ATEN Framework: The Labour and Birth EHR

This section evaluates ATEN by applying it to the **domain of midwifery**. While ATEN is intended to be generally applicable for use with any defined group of patients and chosen health problem or disease that has a Caremap, for the purposes of evaluating the ATEN framework, this work now focuses on the problem of generating the RS-EHR for only the delivery episodes for female patients who are giving birth in the Counties Manukau District Health Board (CMDHB) catchment area of Auckland in New Zealand. The practical advantages, to the authors, of focusing on delivery episodes for the purpose of this evaluation only are that: (1) deliveries take relatively short periods of time; (2) comprehensive statistics are readily available that cover a long period of time; (3) clinical guidelines as well as locally specified midwifery practice protocols derived from localisation of international clinical practice guidelines are widely available; (4) the delivery episode can range from being very simple to being very complex with a wide variety of complicating factors that include the health of the mother and that of the baby; and, (5) the authors had ready access to midwives on a regular basis throughout this research work. The rest of this section presents the prototype system, results of evaluation, and discussion of these results.

The labour and birth EHR contains a record of the labour and birth events starting at onset of labour and ending when delivery is complete and the new child is presented to her parents. To generate the labour and birth EHR in such a way that realism is achieved we apply the ATEN framework's components: THOTH, RA and HORUS. The next sections present this application, which leads to the synthetic labour and birth EHR that has the realistic properties that are guaranteed by the ATEN Framework.

THOTH is a combination of the generic method for SDG, combine with the pre-planning elements that characterise and classify the synthetic data being sought, in this case, the synthetic labour and birth EHR. Table 6 summarises the application of THOTH to the labour and birth scenario leading to the ingredients, method and context for the generation of the synthetic labour and birth EHR. In the context of the labour and birth EHR, the characterisation (truly synthetic data) was selected to meet with the ideal that we do not rely on access to the real EHR in the context of our generation approach. This ensures the highest degree of patient privacy as, unlike most other methods in this domain, no real patient records are necessary to this generation approach.

Analysis of SDG literature demonstrated that a probability weighted random generation approach was more likely to generate the synthetic records required. Also, other methods including the data masking and the signal and noise models required access to some amount of real (seed) EHR data, which discounted their use in this example.

RA is the knowledge discovery and characterisation approach seeking to identify realistic elements of the data gathered during THOTH. Application of RA specifically to the Labour and Birth problem required identification of the care process (Caremap) for labour and birth, as well as its concepts and contexts.

Table 6. Application of THOTH in the context of midwifery EHR generation.

Aspect of THOTH	Application to Labour & Birth context
Identify	Midwifery EHR in the context of the Labour and Birth event
Characterisation	Truly Synthetic Data
Classification (<i>method/algorithm</i>)	Probability Weighted Random Generation
Knowledge Gathering (<i>used in data/knowledge-driven generation algorithm</i>)	Clinical Practice Guidelines & organisational caremaps Ministry of Health (MoH) Labour and Birth statistics Expert Clinical Knowledge from Midwives and Obstetricians Population (census) demographic data Clinical Vocabulary Clinical Notes Library (authored by midwives)

Extrinsic Knowledge Quantitative Properties: The quantitative properties in the domain of midwifery included a range of demographic statistics regarding the mother and baby. Essentially they were not as simple as looking at the examples in blue contained in Fig. 10, presented later in this section, and saying that 22% of mothers were European, or that 24% of mothers were aged between 20–24 years. There were inter-relationships between these values that also needed to be modelled, including that of the 24% of mothers between 20 and 24 years of age, only 8% were identified as European. Other statistics included how many mothers delivered naturally versus by caesarean section, and the spread of clinical interventions across ethnicity, age, and gestation.

Qualitative Properties: A range of qualitative properties were assessed within the knowledge gathered for generating midwifery EHRs. These included the structure of the source data being used, as well as the structure and appearance of how the synthetic data should be presented on generation. A truncated example of how demographic data was structured in one midwifery EHR system is shown in Table 7. Other qualitative aspects might include: (a) logical internal consistence between the dates reported in different fields (last menstrual period, estimated due date, and so on), (b) whether fields have been misappropriated as placeholders for other data types, and (c) the completeness of fields within the dataset.

Table 7. Application of THOTH in the context of midwifery EHR generation [52].

Patient		
PK	patientID	INT
	title	TEXT(10)
	lastName	TEXT(30)
	firstName	TEXT(30)
	dateOfBirth	DATE/TIME
	gender	CHAR(10)
	ethnicity	CHAR(20)
	primaryLanguage	VARCHAR(100)

Intrinsic Knowledge Concept Hierarchy for Labour and Birth Domain: An extract focusing on child birth from the concept hierarchy (CH) developed for the labour and birth domain is presented in Fig. 5. The general term Childbirth 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 with the addition of quantitative statistics (in brackets) identified from the Ministry of Health (New Zealand) source data.

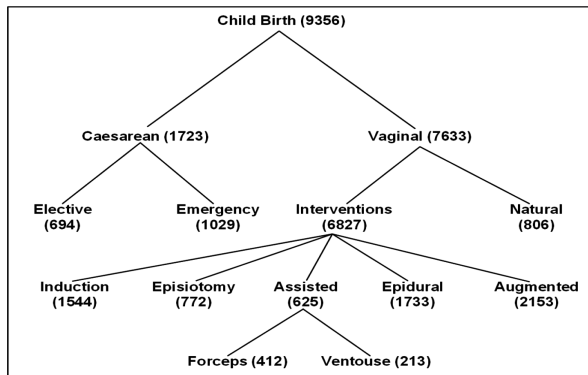


Fig. 5. Concept hierarchy enhanced with statistics [52].

The CH provides structural understanding of primary or significant concepts, from most general to most specific, within the domain being modelled. In RA, is also used to provide statistical understanding of the incidence of each concept. The CH provides constraints, or weights, that are applied during the generation phase, as well as forming one component used to verify statistical accuracy, and in turn realism, within the resulting synthetic data.

Constraining Rules: Characteristic Rule: Fetal heart monitoring is used in midwifery to assess the health, and stress being suffered, by the baby. In the domain of midwifery, we found that only those pregnancies clinically described as low risk receive intermittent fetal heart monitoring. However, clinical practice guidelines (CPGs) necessitate continuous monitoring for a higher risk pregnancy. Properties of this rule would be expressed as the sum of the four elements. The characteristic rule expressed in the conditional formula is shown in Fig. 6 containing the values: *Sex: Female, Pregnant: Yes; Pregnancy Status: Low Risk; Fetal Heart Monitoring: Intermittent in Labour.* This rule was validated against, and found to be consistent with, the CPGs for several hospital birthing facilities in New Zealand.

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

Fig. 6. Example of a characteristic rule [52].

Classification Rule: The CPGs for Labour and Birth provide that where an expectant mother has had a previous caesarean birth, she may elect in this subsequent birth to still (safely) attempt a vaginal birth (known in medical terms as a VBAC - vaginal birth after caesarean). However, where she has had two or more previous caesarean births the obstetric team will counsel her to only have a caesarean birth due to considerations of risk and safety for both mother and baby that result from the previous caesarean scars and potential stress on the uterus. Figure 7 provides an example of a classification rule showing that 100% of patients undergo a caesarean procedure for the current birth if two or more of their previous births have also been by caesarean section. This rule was successfully validated against the MoH Labour and Birth statistics, with the finding that it was true in operation across all births that occurred in New Zealand for that year.

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

Fig. 7. Example of a classification rule [52]

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 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. These rules are used to constrain generation, ensuring consistency between real-world and the synthetic. They are used during validation to identify instances where synthetic records may be inconsistent, for example, if the midwifery patient being generated was male.

Formalisation of Labour and Birth CPG into Labour and Birth Caremaps

The core set of constraints in the CoMSER Method [62] are CPGs, Health Incidence Statistics (HIS), patient demographic statistics and the Caremap, all formalised in an integrated way into the state transition machine (STM) following the process shown in Fig. 8 [51]. The STM is the constraint enforcement formalism for generating the RS-EHR entries satisfying the constraints.

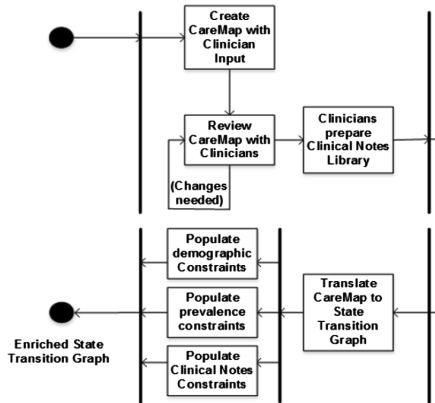


Fig. 8. UML activity diagram: process of creating and integrating constraints into State Transition Machine for the midwifery Caremap [62].

Figure 9 presents the UML State Diagram (USD) for the State Transition Machine (STM) that integrates the core constraints for generating the RS-EHR for delivery episodes within the Counties Manukau District Health Board (CMDHB) of Auckland, New Zealand (NZ).

The transition from one state to the next is determined by the pseudo-random selection of one state in the STM in which is stored the health incident prevalence constraint that is formally specified as the 2-tuple, $\langle P, O \rangle$, such that P is the total number of patients who are known to enter the state according to statistics within the CMDHB catchment area, and O is the number of patients expressed as a percentage of the immediately preceding parent state. The caremap formalised by the STM in Fig. 9, covers the midwifery delivery event, which is also referred to, in this work, as the delivery episode. The caremap begins temporally at the point where the pregnant patient is established as ‘in labour’. It follows the sequence of possible states, that is, clinical events or decisions or both, consistent with the locations, interventions and outcomes that are currently available to the patient or her treating clinicians until the birth process concludes in one of the possible outcomes. Thus, the Caremap and hence its STM form the basis of the integrated constraint framework and also the basis for the algorithm for the RS-EHR generation.

In validating the midwifery RS-EHR, **HORUS** was applied, adhering to the steps as presented in Fig. 4. The following subsections describe the results observed.

Input Validation: In creating the midwifery EHRs for the Labour and Birth event we used CPGs along with treatment and outcome statistics. Input validation necessitated ensuring statistics could be located or extracted that correctly applied to each part of the processes described in the CPGs. Also, cross-validation of those statistics was performed through comparison against more than one source. Where any difference in terminology existed between input datasets, clinicians were involved to ensure these data were correctly linked together [62].

Realism Validation 1: The first realism validation process verified both the premise and accuracy of each rule, ensuring consistency with the semantics of knowledge used in their creation, such as the CPGs discussed in the Input Validation example above.

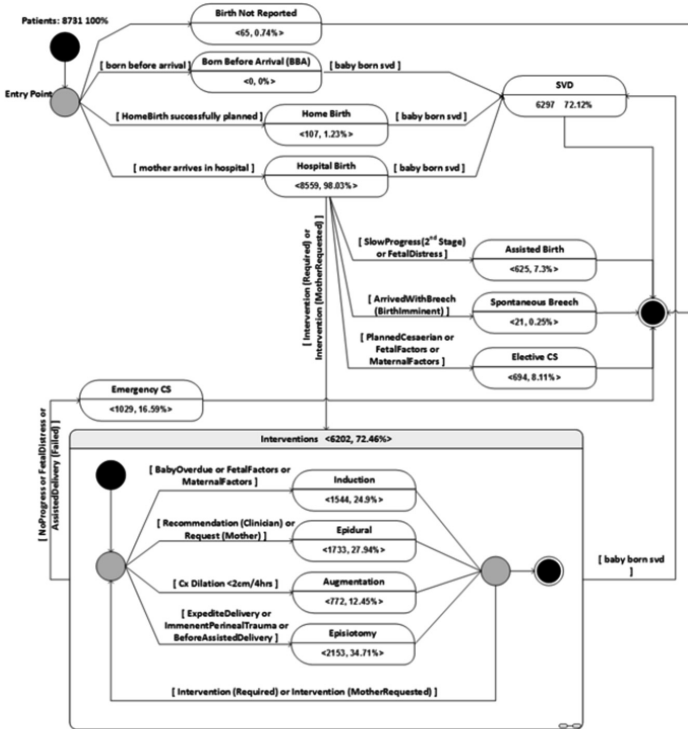


Fig. 9. UML State Diagram that integrates constraints for generating the RS-EHR for delivery episodes within the CMDHB catchment area of NZ [62].

They were tested in real circumstances to ensure they were not rendered irrelevant through interaction with the original source or observed data. Where any knowledge is at issue, the researcher should return to the knowledge discovery phase.

Method Validation: Method validation for these midwifery EHRs concluded that the use of caremaps extended with descriptive rules and statistics, presented as State Transition Machines, and a probability weighted generation model were appropriate given the available input knowledge, purpose and output data required of the CoMSER model.

Output Validation: As one example of output validation, statistical values from within the synthetic data were validated and verified against those identified in the knowledge gathered prior to generation. This comparison is shown as the orange line in Fig. 10, demonstrating that the values contained in the CoMSER synthetic midwifery records were consistent with the MoH statistics used in their production.

Realism Validation 2: In the example of RS-EHR, if a synthetic patient were to be treated in a manner contradictory to the principles or application of a CPG, this could invalidate the entire dataset. In the same instance, if seeking validation by clinicians, it may be necessary to present the synthetic EHR in a clinician-familiar manner.

Using the caremap STM in Fig. 9, the prototype system was used to generate midwifery RS-EHR for 1000 synthetic patients. Figure 11 presents a sample RS-EHR that has been generated by the CoMSER Method prototype. It should be noted that the

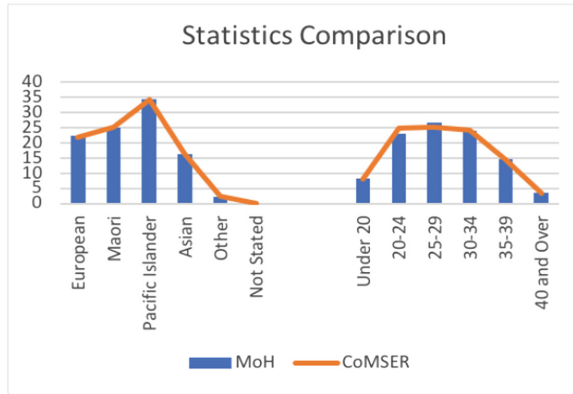


Fig. 10. A comparative quantitative example using patient demographics from the Ministry of Health (NZ) statistics with output validation from our prototype RS-EHR. (Color figure online)

Brianna Allen

Gender: Female
 Ethnicity: European
 DOB: 15 May 1978
 NHI: XX1234

Clinical Records View

Date	Time	Node	Clinical Note	Clinician
18 November 2014	9:05 AM	Start	G4P3	A.Midwife
18 November 2014	9:50 AM	>	Pt presenting to Labour and Birthing. Reports SRM 2 days ago, now reports brownish coloured fluid draining. CTG commenced. FH 188bpm. IV luer sited, bloods sent. IV plasmalyte commenced 500mL/hr. Mec liquor seen on pad, SRM confirmed. T 37.9, P 110 bpm, BP 102/66. Contracting 3.10, strong to palpate. VE: Cx 8 cm, St +1. Plan: commence IV fluids, for IV Abs for PROM, awaiting Obstetric review.	A.Midwife
18 November 2014	9:50 AM	Hospital Birth	Pt presenting to Labour and Birthing.	A.Midwife
18 November 2014	10:50 AM	>	BP 160/98. Call to obstetric consultant. Will attend shortly, meanwhile requesting that epidural be sited. CTG commenced, call to anaesthetist, who will attend shortly.	A.Midwife
18 November 2014	11:20 AM	Epidural	Epidural sited, BP following test dose 106/66. Epidural appears effective, pt now comfortable with contractions. BP stable.	A.Midwife
18 November 2014	2:20 PM	>	VE: Fully dilated, St +1, OA Clear liquor, normal CTG. Epidural remains effective. Plan: allow 1 hour for passive descent, then begin pushing.	A.Midwife
18 November 2014	3:50 PM	SVD	Spont delivery, live baby in poor condition. Cord clamped and cut, emergency bell rung for assistance. Baby to Resuscitaire. 1mL Syntometrine to left thigh. Placenta and membranes delivered CCT, appear complete. Labial lacerations sutured. 4.0 vicryl. Fundus firm and central, EBL 300mL.	A.Midwife

Fig. 11. Sample realistic synthetic EHR generated by CoMSER.

column in the screenshot entitled “Node” indicates either the state or the transition in the STM from which the synthetic entry has been generated. The column has been inserted only for debugging purposes and may or may not be meaningful to the clinician.

A convenience survey of clinicians from New Zealand’s midwifery discipline was conducted to assess the realistic characteristic of synthetic records generated using the CoMSER prototype application. The survey instrument used a forced choice Likert scale in which the clinician examined clinical and temporal notes independently and jointly. The realism survey questions posed to midwife clinicians is found in Table 8.

Table 8. Realism survey [62].

Ref	Survey response prompt	Aspect evaluated
Q1	After reviewing the record of a randomly selected patient, <i>I find the <u>clinical notes</u> for the record identical to the notes a clinician would expect to find in an actual patient EHR</i>	Realistic property for clinical notes
Q2	After reviewing the record of a randomly selected patient, <i>I find the <u>temporal (day/time) information</u> identical to what the clinician would expect to find in the actual patient EHR</i>	Realistic property for the temporal model
Q3	After reviewing the record of a randomly selected patient, <i>I find the <u>clinical notes and temporal (day/time) information</u>, when <u>read together</u>, has neither conflicts nor inconsistencies as would be expected in the actual patient EHR</i>	Realistic property for the entire RS-EHR and hence for clinical logic flow (<i>all constraints taken together</i>)

A total of n = 45 randomly selected records were examined (15 records each by 3 clinician experts) in answering whether the synthetic EHR possessed the same qualities as the clinician would expect to find in actual EHR. The results of this survey demonstrate that clinical and temporal notes, when examined independently, were identical in 93% (Q1) and 93% (Q2) of the records respectively, while 87% (Q3) of the records were identical when examined jointly. In assessing inter-rater reliability among the experts, inconsistencies between the RS-EHR and the actual EHR were identified in 0% (Expert 1), 7% (Expert 2), and 33% (Expert 3) of the records. This survey, involving practicing midwife clinicians, indicates that realism is found in the majority of clinical notes and temporal information, when examined independently and jointly, in synthetic EHRs produced by the CoMSER prototype application. This analysis substantiates our claim that the characteristic of realism does exist in the majority of RS-EHRs developed through the CoMSER method, thus demonstrating the promising usefulness for secondary use.

Summary of Application of ATEN Framework

This section has presented application of the ATEN framework to the generation of the synthetic EHR for the labour and birth domain. The most significant challenge in RS-EHR generation is ensuring that the generated RS-EHR is realistic. The prototype system for generating RS-EHR for midwifery uses an integrated constraints framework, which is formalised using the State Transition Machine (STM). The guideline-based Caremap for the labour and birth domain for which the RS-EHR is to be generated is embodied within the STM. Computations that makeup the RS-EHR are driven by STM execution using pseudo-random transition selection within defined frequency distributions based on local HIS. The quality of the generated RS-EHR is guaranteed by recognition and use of direct interaction with experienced and practising midwives. The development of methods and techniques for measuring the extent of realistic properties of the generated RS-EHR was necessary. Generating RS-EHRs using publicly available health statistics and CPGs ensures patient privacy and confidentiality while also benefiting many uses including: research, software development and training. The ATEN framework provided a structured approach that ensured procedural steps and documentation were not overlooked, and that validation was a consideration from inception through prototype production to evaluation of the resulting synthetic EHR.

While all random number generation methods apply statistics and therefore can be considered as applying the statistics in generation that the researcher intends to find in the result, most still have some variation from true. Many set only one or two parameters (for example, heads or tails), which simplifies their models and limits potential variation in the expected result. Ours set a large number of constraints that all had to be within statistical limits, such as: age, ethnicity, age at pregnancy, age at pregnancy by ethnicity, the type of birth, incidence of each node in the caremap, and the overall patient outcomes. There were more than 15 variables, some interrelated, being handled by the SDG algorithm. Each to be statistically similar at the end of the generation cycle. Validation using HORUS has shown that the prototype system designed with THOTH and RA has achieved the realism that the overall ATEN framework sets out to produce.

5 Discussion and Future Work

ATEN provides a comprehensive way to achieve realistic synthetic data through three inter-dependent approaches, THOTH, RA, and HORUS, that respectively cover (1) a generic approach to SDG with enhancement (THOTH); (2) knowledge discovery (RA); and (3) validation of realism in the resulting synthetic data (HORUS). To the best of our knowledge, no other work in this domain has produced a generic model for SDG, a framework for realism, or a unified approach to validation of synthetic data.

The main *benefit* of **THOTH** is the guarantee of a best plan for the generation method, as well as ensuring best preparation of the knowledge elements and techniques to be used in creation of synthetic data. The THOTH approach is easily implemented and comes with little resource 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.

With adherence to THOTH, **RA** *benefits* through assurance as to the quality of synthetic data being created. This is achieved through establishment of elements and characteristics that define realism for the generation project; the extrinsic quantitative and qualitative properties, and intrinsic knowledge aspects that inhabit the input data. Another benefit of RA is that as additional items of input or seed data are introduced, the statistics, knowledge, constraints, and rules become further refined, increasing the potential accuracy and realism of resulting synthetic data. A *limitation* that arises is that it is presently conducted manual, requiring the researcher to possess an eye for detail along with sound logic, analytical, and problem solving skills.

HORUS *benefits* through being an inherently straight-forward model for validation and verification of synthetic data. HORUS identifies rules, constraints, or data that may be causing issues; reducing the accuracy, realism, and utility of synthetic data being delivered. It is possible that fewer SDG iterations may be required, significantly reducing the time taken to produce accurate and realistic synthetic data. No directly comparable works were located during this research. However, the closest relatable work encountered was that of [35], whose work presented four separate approaches to validation of synthetic data produced in the domain of computational modelling. Each of these approaches appears, even in that authors' own summation, not to be representative of a single validation solution. The strength of HORUS is that it represents a single operational validation solution. HORUS has a significant *limitation* in that it is wholly dependent on having already engaged RA to identify the statistics, knowledge, and rules that will be significant in assuring that the synthetic data is suitably representative. Another limiting issue is that the case study conducted in this work 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 those extrinsic issues are resolved.

There are a number of avenues open for **future work**, including use of ATEN during the entire lifecycle of a significant real-life SDG project. This would necessitate the considered operation of a new SDG project where every element was documented rigorously, and where two streams are conducted concurrently. This new project due to the incompleteness of every SDG project reviewed during this research. 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 Fig. 2. 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 same input materials and documentation from the first and use them to follow the complete and validated SDG approach described in this work. The second researcher would ameliorate his input materials and generation method through operation of ATEN. Both synthetic datasets could then be validated using HORUS. Another avenue for future work would be development of machine learning models to automate some or all of the KDD and validation.

6 Summary and Conclusion

This chapter has presented and demonstrated the ATEN framework, a triangle of three interdependent approaches: THOTH, RA and HORUS. The triangle is one of the strongest structures seen in engineering and nature. The components communicate with their adjacent neighbours; each enhanced through interaction with and engagement of its counterparts. 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 approach proposed in this work, first, draws on, expands and enhances established methods to result in a complete end-to-end validation solution. This ensures a complete analysis of the source data leading to useful knowledge, which greatly improves the generation method leading to better realism in synthetic data. Second, the knowledge gathered prior to synthetic data generation provides a solid base with which to validate the synthetic data, ensuring its ability to actually replace real data. Third, the approach presented here is simple, intuitive 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

1. McGraw-Hill: McGraw-Hill Dictionary of Scientific and Technical Terms, 6th edn. McGraw-Hill, London (2003)
2. Rubin, D.: Discussion: statistical disclosure limitation. *J. Off. Stat.* **9**, 461–468 (1993)
3. Alter, H.: Creation of a synthetic data set by linking records of the Canadian survey of consumer finances with the family expenditure survey. *Ann. Econ. Soc. Meas.* **3**(2), 373–397 (1994)
4. Wolff, E.: Estimates of the 1969 size distribution of household wealth in the US from a synthetic data base Trans.). In: Smith, J. (ed.) *Modelling the Distribution and Intergenerational Transmission of Wealth*. University of Chicago Press, Chicago (1980)
5. Green, P.E., Rao, V.R.: Conjoint measurement for quantifying judgmental data. *J. Mark. Res.* **8**(3), 355–363 (1971)
6. Birkin, M., Clarke, M.: SYNTHESIS – a synthetic spatial information system for urban and regional analysis: methods and examples. *Environ. Plan.* **20**(1), 1645–1671 (1998)
7. Stedinger, J., Taylor, M.: Synthetic streamflow generation: model verification and validation. *Water Resour. Res.* **18**(4), 909–918 (1982)
8. Geweke, J., Porter-Hudak, S.: The estimation and application of long memory series models. *J. Time Ser. Anal.* **4**(4), 221–238 (1983)
9. Graham, V.A., Hollands, K., Unny, T.E.: A time series model for Kt with application to global synthetic weather generation. *Sol. Energy* **40**(2), 83–92 (1988)
10. Delleur, J., Kavvas, M.: Stochastic models for monthly rainfall forecasting and synthetic generation. *J. Appl. Meteorol.* **17**, 1528–1536 (1978)
11. Barse, E., Kvarnstrom, H., Jonsson, E.: Synthesizing test data for fraud detection systems. Paper presented at the 19th Annual Computer Security Applications Conference (2003)

12. Houkjaer, K., Torp, K., Wind, R.: Simple and realistic data generation. Paper presented at the VLDB 2006 (2006)
13. Mouza, C., et al.: Towards an automatic detection of sensitive information in a database. Paper presented at the 2nd International Conference on Advances in Database Knowledge and Database Applications (2010)
14. Whiting, M., Haack, J., Varley, C.: Creating realistic, scenario-based synthetic data for test and evaluation of information analytics software. Paper presented at the 2008 Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualisation (BELIV 2008) (2008)
15. Gargiulo, F., Ternes, S., Huet, S., Deffuant, G.: An iterative approach for generating statistically realistic populations of households. *PLOS ONE* **5**(1), e8828 (2010)
16. Srikanthan, R.M.T.: Stochastic generation of annual, monthly and daily climate data: a review. *Hydrol. Earth Syst. Sci. Discuss.* **5**(4), 653–670 (2001)
17. Wan, L., Zhu, J., Bertino, L., Wang, H.: Initial ensemble generation and validation for ocean data assimilation using HYCOM in the Pacific. *Ocean Dyn.* **58**, 81 (2008)
18. Killourhy, K., Maxion, R.: Toward realistic and artefact-free insider-threat data. Paper presented at the 23rd Annual Computer Security Applications Conference (CSAC) (2007)
19. Sperotto, A., Sadre, R., Van Vliet, F., Pras, A.: A labelled data set for flow-based intrusion detection. Paper presented at the 9th IEEE International Workshop on IP Operations and Management (IPOM 2009) (2009)
20. Zanero, S.: Flaws and frauds in the evaluation of IDS/IPS technologies. Paper presented at the Forum of Incident Response and Security Teams (FIRST 2007) (2007)
21. Ascoli, G., Krichmar, J., Nasuto, S., Senft, S.: Generation, description and storage of dendritic morphology data. *Philos. Trans. R. Soc. Lond.* **365**, 1131–1145 (2001)
22. Bozkurt, M., Harman, M.: Automatically generating realistic test input from web services. Paper presented at the 6th International Symposium on Service Oriented System Engineering (2011)
23. Drechsler, J., Reiter, J.: An empirical evaluation of easily implemented, non-parametric methods for generating synthetic datasets. *Comput. Stat. Data Anal.* **55**(12), 3232–3243 (2011)
24. Gymrek, M., McGuire, A., Golan, D., Halperin, E., Erlich, Y.: Identifying personal genomes by surname. *Science* **339**(6117), 321–324 (2013). <https://doi.org/10.1126/science.1229566>
25. Ohm, P.: Broken promises of privacy: responding to the surprising failure of anonymisation. *UCLA Law Rev.* **57**, 1701 (2010)
26. Sweeney, L., Abu, A., Winn, J.: Identifying Participants in the Personal Genome Project by Name. Data Privacy Lab, Harvard University (2013)
27. Lundin, E., Kvarnström, H., Jonsson, E.: A synthetic fraud data generation methodology. In: Deng, R., Bao, F., Zhou, J., Qing, S. (eds.) ICICS 2002. LNCS, vol. 2513, pp. 265–277. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-36159-6_23
28. Stratigopoulos, H., Mir, S., Makris, Y.: Enrichment of limited training sets in machine-learning-based analog/RF test. Paper presented at the DATE 2009 (2009)
29. Wu, X., Wang, Y., Zheng, Y.: Privacy preserving database application testing. Paper presented at the WPES 2003 (2003)
30. McLachlan, S., et al.: Learning health systems: the research community awareness challenge. *BCS J. Innov. Health Inform.* **25**(1), 038–040 (2018)
31. Jaderberg, M., K. Simonyan, A. Vedaldi and A. Zisserman. (2014). Synthetic data and artificial neural networks for natural scene text recognition. [arXiv:1406.2227](https://arxiv.org/abs/1406.2227)
32. Penduff, T., Barnier, B., Molines, J., Madec, G.: On the use of current meter data to assess the realism of ocean model simulations. *Ocean Model.* **11**(3), 399–416 (2006)

33. Putnam, H.: Realism and reason. In: *Proceedings and Addresses of the American Philosophical Association*, vol. 50, no. 6, pp. 483–498 (1977)
34. Barlas, Y.: Formal aspects of model validity and validation in system dynamics. *Syst. Dyn. Rev.* **12**(3), 183–210 (1996)
35. Carley, K.: *Validating Computational Models*. Carnegie Mellon University, Cambridge (1996)
36. Brinkhoff, T.: Generating traffic data. *IEEE Data Eng. Bull.* **26**(2), 19–25 (2003)
37. Giannotti, F., Mazzoni, A., Puntoni, S., Renso, C.: Synthetic generation of cellular network positioning data. Paper presented at the 13th Annual ACM International Workshop on Geographic Information Systems (2005)
38. Stodden, V.: The scientific method in practice: reproducibility in the computational sciences. SSRN Paper 1550193. MIT Sloan School of Management (2010)
39. Collins, H.: *Changing Order: Replication and Induction in Scientific Practice*. University of Chicago Press, Chicago (1992)
40. Moss, P.: Can there be validity without reliability? *Educ. Res.* **23**(2), 5–12 (1994)
41. Tsvetovat, M., Carley, K.: Generation of realistic social network datasets for testing of analysis and simulation tools. Technical report 9. DTIC (2005)
42. Richardson, I., Thomson, M., Infield, D.: A high-resolution domestic building occupancy model for energy demand simulations. *Energy Build.* **40**(8), 1560–1566 (2008)
43. Domingo-Ferrer, J.: Marginality: a numerical mapping for enhanced exploitation of taxonomic attributes. In: Torra, V., Narukawa, Y., López, B., Villaret, M. (eds.) *MDAI 2012. LNCS (LNAI)*, vol. 7647, pp. 367–381. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-34620-0_33
44. Efstratiadis, A., Dialynas, Y., Kozanis, S., Koutsoyiannis, D.: A multivariate stochastic model for the generation of synthetic time series at multiple time scales reproducing long-term persistence. *Environ. Model. Softw.* **62**, 139–152 (2014)
45. Van den Bulcke, T., et al.: SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinform.* **7**(1), 43 (2006)
46. Mateo-Sanz, J.M., Martínez-Ballesté, A., Domingo-Ferrer, J.: Fast generation of accurate synthetic microdata. In: Domingo-Ferrer, J., Torra, V. (eds.) *PSD 2004. LNCS*, vol. 3050, pp. 298–306. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-25955-8_24
47. Gafurov, T., Usaola, J., Prodanovic, M.: Incorporating spatial correlation into stochastic generation of solar radiation data. *Sol. Energy* **115**, 74–84 (2015)
48. Brissette, F.P., Khalili, M., Leconte, R.: Efficient stochastic generation of multi-site synthetic precipitation data. *J. Hydrol.* **345**(3), 121–133 (2007)
49. Gainotti, S., et al.: Improving the informed consent process in international collaborative rare disease research: effective consent for effective research. *Eur. J. Hum. Genet.* **24**, 1248 (2016)
50. Arifin, S.M.N., Madey, G.R.: Verification, validation, and replication methods for agent-based modeling and simulation: lessons learned the hard way! In: Yilmaz, L. (ed.) *Concepts and Methodologies for Modeling and Simulation. SFMA*, pp. 217–242. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-15096-3_10
51. Greene, J.C., Caracelli, V., Graham, W.F.: Toward a conceptual framework for mixed-method evaluation designs. *Educ. Eval. Policy Anal.* **11**(3), 255–274 (1989)
52. McLachlan, S., Dube, K., Gallagher, T., Daley, B., Walonoski, J.: The ATEN framework for creating the realistic synthetic electronic health record. Paper presented at the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018), Madiera, Portugal (2018)
53. Lydiard, T.: Overview of the current practice and research initiatives for the verification and validation of KBS. *Knowl. Eng. Rev.* **7**(2), 101–113 (1992)

54. Ishigami, M., Cumings, J., Zetti, A., Chen, S.: A simple method for the continuous production of carbon nanotubes. *Chem. Phys. Lett.* **319**(5), 457–459 (2000)
55. Mahmoud, E.: Accuracy in forecasting: a survey. *J. Forecast.* **3**(2), 139–159 (1984)
56. Nicoletti, I., Migliorati, G., Pagliacci, M., Grignani, F., Riccardi, C.: A rapid and simple method for measuring thymocyte apoptosis by propidium iodide staining and flow cytometry. *J. Immunol. Methods* **139**(2), 271–279 (1991)
57. Rosevear, A.: Immobilised biocatalysts – a critical review. *J. Chem. Technol. Biotechnol.* **34**(3), 127–150 (1984)
58. Parnas, D., Clements, P.: A rational design process: how and why to fake it. *IEEE Trans. Softw. Eng.* **2**, 251–257 (1986)
59. Winkler, W.E.: Masking and re-identification methods for public-use microdata: overview and research problems. In: Domingo-Ferrer, J., Torra, V. (eds.) *PSD 2004*. LNCS, vol. 3050, pp. 231–246. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-25955-8_18
60. Andoulsi, I., Wilson, P.: Understanding liability in eHealth: towards greater clarity at European Union level. In: George, C., Whitehouse, D., Duquenoy, P. (eds.) *eHealth: Legal, ethical and governance challenges*, pp. 165–180. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-22474-4_7
61. Mwogi, T., Biondich, P., Grannis, S.: An evaluation of two methods for generating synthetic HL7 segments reflecting real-world health information exchange transactions. Paper presented at the AMIA Annual Symposium Proceedings (2014)
62. McLachlan, S., Dube, K., Gallagher, T.: Using CareMaps and health statistics for generating the realistic synthetic electronic healthcare record. Paper presented at the International Conference on Healthcare Informatics (ICHI 2016), Chicago, USA (2016)
63. Cassa, C., Olson, K., Mandl, K.: System to generate semisynthetic data sets of outbreak clusters for evaluation of outbreak-detection performance. *Morb. Mortal. Wkly Rep. (MMWR)* **53**, 231 (2004)
64. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: Knowledge discovery and data mining: towards a unifying framework. *KDD* **96**, 82–88 (1996)
65. Fernandez-Arteaga, V., et al.: Association between completed suicide and environmental temperature in a Mexican population, using the KDD approach. *Comput. Methods Programs Biomed.* **135**, 219–224 (2016)
66. Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge discovery and interactive data mining in Bopinformatics: state-of-the-art, future challenges and research directions. *BMC Bioinform.* **15**(6), II (2014)
67. Mitra, S., Pal, S., Mitra, P.: Data mining in soft computing framework: a survey. *IEEE Trans. Neural Netw.* **13**(1), 3–14 (2002)
68. Nijssen, G.M., Halpin, T.A.: *Conceptual Schema and Relational Database Design: A Fact Oriented Approach*. Prentice Hall Inc., Upper Saddle River (1989)
69. Han, J., Cai, Y., Cercone, N.: Data-driven discovery of quantitative rules in relational databases. *IEEE Trans. Knowl. Data Eng.* **5**(1), 29–40 (1993)
70. Sanderson, M., Croft, B.: Deriving concept hierarchies from text. Paper presented at the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1999)
71. Barnes, C.A.: *Concepts Hierarchies for Extensible Databases*. Naval Postgraduate School, Monterey (1990)
72. Ganter, B., Willie, R.: Applied lattice theory: formal concept analysis. In: *General Lattice Theory*. Birkhauser, Basel (1997)
73. Rodriguez-Jimenez, J., Cordero, P., Enciso, M., Rudolph, S.: Concept lattices with negative information: a characterisation theorem. *Inf. Sci.* **369**(51), 51–62 (2016)

74. Bex, G., Neven, F., Schwentick, T., Tuyls, K.: Inference of concise DTDs from XML data. Paper presented at the 32nd International Conference on Very Large Databases (2006)
75. Laranjeiro, N., Vieira, M., Madeira, H.: Improving web services robustness. Paper presented at the IEEE International Conference on Web Services ICWS 2009 (2009)
76. Oreskes, N., Shrader-Frechette, K., Belitz, K.: Verification, validation and confirmation of numerical models in the earth sciences. *Science* **263**(5147), 641–646 (1994)
77. McLachlan, S.: Realism in synthetic data generation. Master of Philosophy in Science MPhil, Massey University, Palmerston North, New Zealand (2017). Available from database