The ATEN Framework for Creating the Realistic Synthetic Electronic Health Record

Scott McLachlan[1], Kudakwashe Dube[2], Thomas Gallagher[3], Bridget Daley[4] and Jason Walonoski[5]
[1]*School of Electronic Engineering and Computer Science, Queen Mary University of London, London, U.K.*
[2]*School of Engineering and Advanced Technology, Massey University, Palmerston North, New Zealand*
[3]*Applied Computing and Engineering Technology, University of Montana, Missoula, U.S.A.*
[4]*Western Sydney Local Health District, NSW Health, Australia*
[5]*The MITRE Corporation, Massachusetts, U.S.A.*

Keywords: Synthetic Data, Synthetic Healthcare Record, Knowledge Discovery, Data Mining, Electronic Health Records, Computer Simulation, ATEN Framework, Validation, RS-EHR.

Abstract: Realistic synthetic data are increasingly being recognized as solutions to lack of data or privacy concerns in healthcare and other domains, yet little effort has been expended in establishing a generic framework for characterizing, achieving and validating realism in Synthetic Data Generation (SDG). The objectives of this paper are to: (1) present a characterization of the concept of realism as it applies to synthetic data; and (2) present and demonstrate application of the generic ATEN Framework for achieving and validating realism for SDG. The characterization of realism is developed through insights obtained from analysis of the literature on SDG. The development of the generic methods for achieving and validating realism for synthetic data was achieved by using knowledge discovery in databases (KDD), data mining enhanced with concept analysis and identification of characteristic, and classification rules. Application of this framework is demonstrated by using the synthetic Electronic Healthcare Record (EHR) for the domain of midwifery. The knowledge discovery process improves and expedites the generation process; having a more complex and complete understanding of the knowledge required to create the synthetic data significantly reduce the number of generation iterations. The validation process shows similar efficiencies through using the knowledge discovered as the elements for assessing the generated synthetic data. Successful validation supports claims of success and resolves whether the synthetic data is a sufficient replacement for real data. The ATEN Framework supports the researcher in identifying the knowledge elements that need to be synthesized, as well as supporting claims of sufficient realism through the use of that knowledge in a structured approach to validation. When used for SDG, the ATEN Framework enables a complete analysis of source data for knowledge necessary for correct generation. The ATEN Framework ensures the researcher that the synthetic data being created is realistic enough for the replacement of real data for a given use-case.

1 INTRODUCTION

Rapid adoption of the Electronic Healthcare Record (EHR) continues to increase demand for secondary use of patient records. Synthetic EHRs have been suggested as a privacy protecting solution for producing available data sets, yet without the property of realism these synthetic data will have limited secondary use. The Realistic Synthetic EHR (RS-EHR) has been suggested as a privacy protecting technology for producing data for secondary use (Dube and Gallagher, 2014). Early prototypes, such as CoMSER (McLachlan et al., 2016), have demonstrated the important characteristic of realism when examined by clinicians. A continued challenge in developing the RS-EHR is the lack of consistent methods for validating and verifying realism.

Sourcing or developing experimental datasets can be costly and often presents an insurmountable challenge (Bozkurt and Harman, 2011, Whiting et al., 2008, Williams et al., 2007). The use of synthetic data has developed from concept to a key solution routinely employed (Weston et al., 2015). Synthetic

McLachlan, S., Dube, K., Gallagher, T., Daley, B. and Walonoski, J.
The ATEN Framework for Creating the Realistic Synthetic Electronic Health Record.
DOI: 10.5220/0006677602200230
In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) - Volume 5: HEALTHINF*, pages 220-230
ISBN: 978-989-758-281-3

data may be used in research and experimentation, software and systems testing or for training and demonstration purposes (Whiting et al., 2008, Houkjaer et al., 2006, Mouza et al., 2010). The literature describes an array of models and methods; from those that simply remove or replace personally identifiable information (Mouza et al., 2010) through to those that entirely eschew using any personally identifiable or confidential records (McLachlan et al., 2016).

While generating synthetic data may sound simple, generating good synthetic data proves to be extremely difficult (Whiting et al., 2008). Articles published across every domain propose new methods for generating synthetic data. In the healthcare domain synthetic data may take the form of health records (McLachlan et al., 2016, Mwogi et al., 2014), or models of gene expression (Van den Bulcke et al., 2006) and nerve cells (Ascoli et al., 2001). An often stated primary goal in many articles is that synthetic data must comprise a suitable replacement for real data (Wu et al., 2003, Stratigopoulos et al., 2009). Some go so far as to prescribe that synthetic data should possess the property of realism; that it should be a realistic substitute for existing data (Bozkurt and Harman, 2011, Jaderberg et al., 2014, Brinkhoff, 2003).

Very few SDG solutions perform true validation, and for those that do it is only entertained once development concludes and the generation process is complete. This makes it difficult to ascertain where issues identified in the synthetic dataset occurred (Lydiard, 1992). It is almost impossible to establish whether issues are an artefact of the seed information, a coding or algorithmic miscalculation in the generation process, an output error, or a combination of all of these. These situations can be identified and resolved through use of the validation method described in this research.

The rest of this paper is organised as follows: First, the paper presents the realism as an unresolved issue in the SDG domain. Second, a review of literature on realism and its validation in SDG approach is presented. Third, realism is then characterised and an approach for establishing the elements of realism SDG is proposed. Fourth, an approach for realism validation is presented with its demonstration and evaluation using EHR examples from the midwifery domain. Fifth, a discussion of the application and benefits of the approaches to the characterisation and validation of realism is presented. Finally, future work is followed by a summary and conclusion.

2 THE RESEARCH PROBLEM

The concept of realism remains largely unaddressed in current SDG literature. In the case of the RS-EHR, realism means synthetic patients should not just statistically and structurally mirror real counterparts, but that the concepts, symptomatology, treatments and language used should also be consistent. Certainly, any clinician accessing the record should also find it to be realistic. A definition and description for realism would enable the reader and any future users to contextualize elements and relate them to the SDG author's desired outcomes. This paper seeks such a definition, as well as a method for achieving and validating results in SDG. Comprehensive and systematic classification and validation is necessary to facilitate the repeating of experiments and validate whether the authors actually achieved their objectives (Crawford and Stucki, 1990, Creswell, 2003).

This paper investigates the property of realism in synthetic data within the healthcare domain. This research provides a framework that can be used in SDG to ensure: (a) realism is a properly considered component in the development and testing of the algorithm, generation method and outputs; (b) quality of documentation, and; (c) a basis for validating and substantiating the claim that synthetic data is a realistic substitute for the real data it is intended to replace.

3 RELATED WORKS

In order to understand the concept of realism, its utility and validation in SDG, this section reviews works that; (a) discussed realism and/or a need for synthetic data to be realistic, and; (b) provided some discourse on validation approaches for synthetic data. A search was conducted to identify works concerning *synthetic data generation* (n = 7,746). The search was narrowed further to works that also used the terms *realistic* (n = 290) or *realism* (n = 6). This included works that identified realism as a primary concern in the generation of any synthetic data (Bozkurt and Harman, 2011, Houkjaer et al., 2006, Tsvetovat and Carley, 2005) or that discussed developing synthetic data that would be realistic, sufficient to replace or be representative of real data (Mouza et al., 2010, Jaderberg et al., 2014, Richardson et al., 2008, Sperotto et al., 2009). Given the low number of works that specifically mentioned the terms realistic and realism, a selection of the excluded works was further reviewed. This review found that around one third of SDG articles use functionally similar terminology

such as *authentic* (Barse et al., 2003) and *accurate to real structures* (Ascoli et al., 2001).

3.1 Realism in SDG

In order to be an accurate replacement for real data, synthetic data must be realistic (Sperotto et al., 2009). The presence of realism should only be asserted if it is verified (Penduff et al., 2006, Putnam, 1977). The property of realism is seen to bring a greater degree of accuracy, reliability, effectiveness, credibility and validity (Bozkurt and Harman, 2011). Realism is therefore desirable; an important property needed of most synthetic data (Bozkurt and Harman, 2011, Mouza et al., 2010, Alessandrini et al., 2015, Bolon-Canedo et al., 2013). Researchers should ascertain the elements of realism in the data they seek to synthesize (Laymon, 1984, McMullin, 1984).

Few authors provide a definition or foundation for understanding realism (n = 2). In both cases this is limited and vague, implying only that the aim of realism was to be that the synthetic data needed to be a representative replacement for real data (Sperotto et al., 2009) and comparably correct in size and distribution (Killourhy and Maxion, 2007). Neither article discussed validating for realism in the synthetic data they created.

3.2 Validation of Synthetic Data

The focus for the majority of SDG articles captured in this review was on the synthetic data, rather than on thoroughly documenting the generation method or validation. The limited validation methods observed consisted of comparisons between either the entire dataset or attributes selected from it and minor observations from the real data (Bozkurt and Harman, 2011) or provide imperfect graphical and statistical comparisons which the authors admit only provide indirect visual or structural comparison to a low probability, and not true algorithm validation (Ascoli et al., 2001, Efstratiadis et al., 2014, Gafurov et al., 2015). Some focused on simply reporting the performance of their method or discussing efficiency of algorithms used to create synthetic data (Agrawal et al., 2015). The majority of authors often included no discussion of validation at all (Brinkhoff, 2003, Brissette et al., 2007, Giannotti et al., 2005).

The term *validation* is representative of acts which seek to prove accuracy and comparability, but its meaning and use are often misunderstood (Carley, 1996, Oreskes et al., 1994, Oxford, 2016). The way validation is used in contemporary literature more correctly describes two related but separate concepts; *validation* and *verification* (Oreskes et al., 1994). Validation assesses functional correctness and internal consistency while verification seeks comparability of the synthetic data with observation (Oreskes et al., 1994). A number of validation approaches are regularly used and documented within related fields such as Computational Modelling (Carley, 1996), however no complete, applicable or reusable validation approaches were observed across the SDG literature reviewed.

4 MATERIALS AND METHODS

In the case of synthetic data, realism can be identified as the sum of two levels of knowledge. ***The first level*** is the *extrinsic* or overt knowledge; the structure. The data fields and general statistical values easily realised from observational data or other inputs being used in the generation process: that is, the observable quantitative and qualitative aspects of the input data. ***The second level*** is *intrinsic* or covert knowledge. The relationships, concepts, rules and representations drawn out from within the input data. The method proposed in this research utilizes established Knowledge Discovery in Database (KDD) processes, extended through application of Human Computer Interaction (HCI)-KDD principles, Concept Hierarchies (CH), Formal Concept Analysis (FCA) and identification of Characteristic and Classification rules that inherently describe the data.

Ensuring realism can mean different things for data synthesis in the domain of healthcare. Strict realism requires adherence to structural, statistical and conceptual qualities of patient records. In the healthcare context, there is an ***additional realism element*** in the manner that information within the synthetic patient record is recorded using medical language, terminology and logic, especially in the narrative of clinicians within the patient notes. We recognize that this realism element is difficult to reproduce and it is accordingly not covered in the focus of this paper. Our solution to this problem is appropriate synthetic patient narrative scripted by practicing clinicians for application to particular medical events (McLachlan et al., 2016).

5 THE ATEN FRAMEWORK

The ATEN Framework (Figure 1) is an SDG lifecycle that provides researchers with a structured approach for SDG projects including identifying and validating

realism. The ATEN Framework is composed of the THOTH, RA and HORUS approaches that respectively provide for: (a) an approach to generating synthetic data; (b) an approach to defining and describing the elements of realism, and; (c) an approach for validating synthetic data. ATEN was a syncretized deity from the thirteenth century BC, elevated for a time as the single central god of Egyptian religion (Gore, 2001). ATEN, whose name was abbreviated from Ra-Horakhty or Ra-Horus-Aten (meaning Ra, who is Horus of the two horizons) represented the synthesis of a number of gods from Egyptian mythology (Gunn, 1923, Wilkinson, 2008).

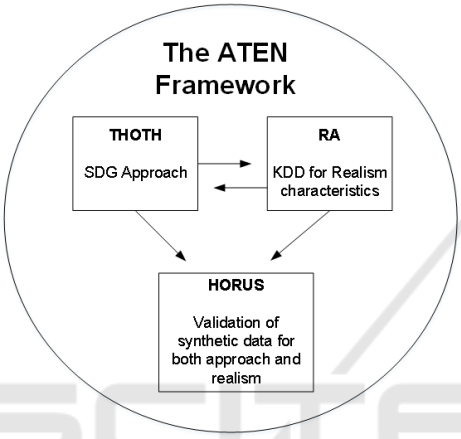

Figure 1: The ATEN Framework.

The THOTH approach describes the overarching SDG process from conceptualisation to generation. The primary focus of RA is knowledge discovery to classify and characterize the realistic elements of the input data including: structures, statistics, rules and concepts. These elements are then made available to the generation process (THOTH) with the intention of ensuring that the synthetic data can be generated consistent with the knowledge discovered. The HORUS approach uses this knowledge to validate the realism in the resulting synthetic data. In this way, the RA approach identifies knowledge necessary for achieving realism in synthetic data, while HORUS considers whether realism has been achieved. The next three sections present the THOTH, RA, and HORUS approaches in greater detail.

5.1 THOTH – The Enhanced Generic SDG Approach

THOTH describes the four-step generic SDG approach observed in much of the SDG literature. The four steps shown in Figure 2 include (1) Identifying the need for synthetic data; (2) Gathering the required raw information, datasets and knowledge required to generate the synthetic data (and also input to RA); (3) Developing the algorithm or method that will perform the generation process, and; (4) The generation process.

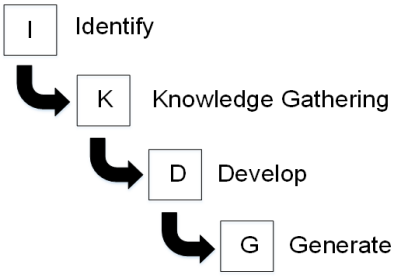

Figure 2: The four-step generic SDG approach.

The enhanced THOTH approach incorporates additional steps shown in Figure 3 which *characterise* the synthetic-ness of the data required, and *classify* the generation method best suited to the task. The synthetic-ness of generated data can range from real data on which anonymisation processes may be performed, through to truly synthetic data which relied on no personally identifiable information during creation. The generation method is drawn from one of the five primary classification types identified in our research, shown in Table 1.

Table 1: SDG Classification Models with examples from the literature.

SDG Classification Model	Example
Data Masking	Mouza et al, 2010
Signal and Noise	Whiting et al, 2008
Network Generation	Ascoli et al, 2001
Random Generation Models	Mwogi et al, 2014
Probability Weighted Random Models	Mwogi et al, 2014 Houkjaer et al, 2006 McLachlan et al, 2016

5.2 The RA Approach – The Generic Approach to Realism in SDG

RA provides a structured approach to discovering, classifying and characterizing knowledge and realism elements for use in SDG. The process of RA, including the steps of enhanced knowledge discovery are both shown in Figure 4 and further elaborated in Table 2. While encompassing the entire KDD process, the structured enhancements prescribed by RA are contained within the Data Mining step.

RA is built on the generic KDD process that identifies both the extrinsic and intrinsic realism elements. It follows a logical progression of steps that

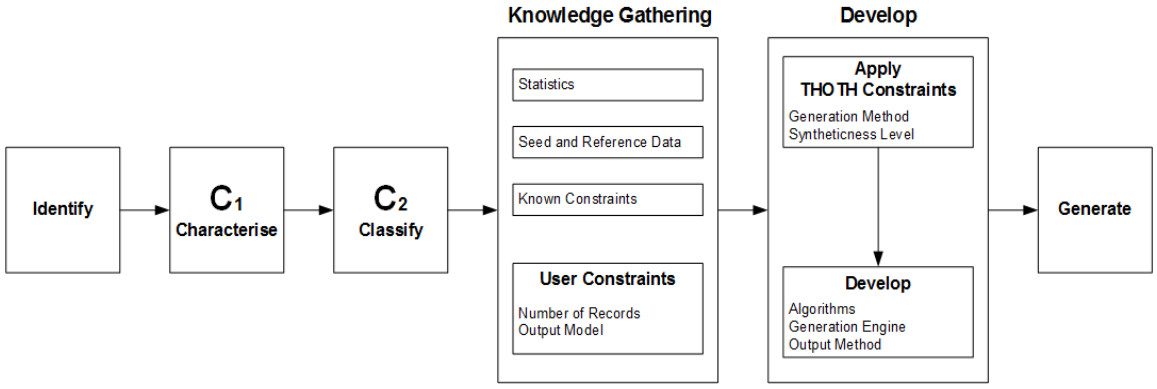

Figure 3: The Enhanced THOTH Approach.

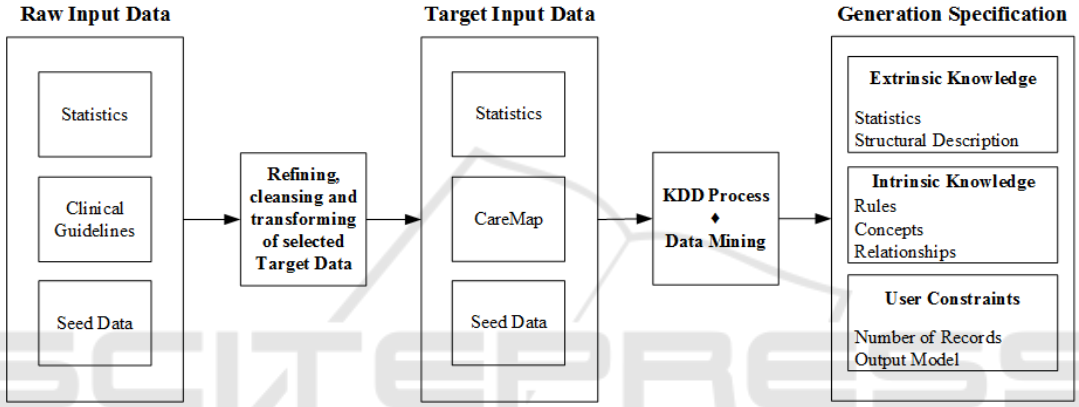

Figure 4: Overview of the RA Approach to Realism in SDG.

Table 2: Enhanced Knowledge Discovery Process further describing the RA Approach to realism in SDG as presented in Figure 4.

Step	Activity	Tasks
1	Develop and document information (overlaps with THOTH)	Relevant prior knowledge; An understanding of the application domain, and; The goal(s) of the 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 from the target data pool.
4	Create target data	Integrate data from multiple sources; Transform raw data, for example: Clinical guidelines and health incidence statistics into consistent input data sets; Project data by identifying useful features for representing the data, for example: as CareMaps, and; Reduce the number of 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, 1996). Perform concept hierarchy analysis, formal concept analysis, rule identification of the methods used in HORUS
6	Interpret and evaluate mined patterns	Identify the truly interesting and useful patterns.
7	Presentation	Make the knowledge available for use in synthetic data generation

are observable within the literature (Fayyad et al., 1996, Fernandez-Arteaga et al., 2016, Holzinger et al., 2014, Mitra et al., 2002). The following subsections present the processes used as part of the KDD data mining in Step 5 of Table 2.

5.2.1 Extrinsic: Identifying Quantitative and Qualitative Properties

Identifying and documenting the quantitative and qualitative properties of real data is the first step of knowledge discovery. The synthetic data generated must possess these properties to be a suitable replacement. Examples of quantitative and qualitative knowledge from the demographic portion of the EHR are shown in Tables 3 and 4.

Table 3: An example of quantitative knowledge - Patient Ethnicity data.

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

5.2.2 Intrinsic: Concept Hierarchies

Concept Hierarchies (CH) are a deduction of attribute-oriented quantitative rules drawn from large and very large datasets (Han et al., 1993). 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” (Han et al., 1993, Mitra et al., 2002, Sanderson and Croft, 1999). Developing a concept hierarchy involves organizing levels of concepts identified within the data into a taxonomy, reducing candidate rules to formulas with a particular vocabulary (Han et al., 1993). CH are used in RA to identify an entity type, the instances of that entity and how they relate to each other; they help to ensure the identification of important relationships in the data that can be used to synthesise meaningful results (Barnes, 1990).

Table 4: An example of qualitative knowledge - The structure and field definitions from the Patient Demographics table.

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

Once the concept hierarchy tree is identified, another pass across the source data should occur to count the occurrence of each of the specific terms. This second pass allows the researcher to enhance the concept hierarchy with statistics that can be used as knowledge to improve accuracy in the generation model. An example of a concept hierarchy enhanced with statistics is shown in Figure 5.

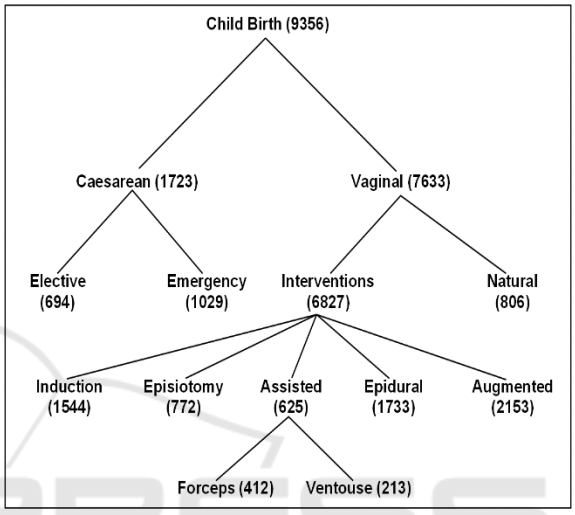

Figure 5: Concept Hierarchy enhanced with statistics.

5.2.3 Intrinsic: Formal Concept Analysis

Formal Concept Analysis (FCA) is a method of representing information that allows the researcher to easily realise relationships between instances of an object and occurrences of a concept; for example: occurrences of various nosocomial infections across the different 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} (Ganter and Willie, 1997). A table is created displaying instances where a relationship exists between an object and 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 of the objects.

The second step to FCA involves creating a 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 (Rodriguez-Jiminez et al., 2016).

5.2.4 Intrinsic: Characterization and Classification Rules

Characterization and classification rules come from within the target dataset and are used as constraints during the generation process. When applied in the generation process, these rules help to improve overall accuracy.

Characterization 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 Figure 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. In the domain of midwifery, we might find that while only those pregnancies clinically described as low risk would receive intermittent fetal heart monitoring, 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 Figure 6 containing the values Sex:Female, Pregnant:Yes, Pregnancy Status:Low Risk, Fetal Heart Monitoring:Intermittent in Labour.

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

Figure 6: Example of a Characteristic rule conditional formula from the domain of midwifery.

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. Figure 7 demonstrates an example of a classification rule showing that 100% of patients will undergo a caesarean procedure for the current birth if two or more of their previous births have also been by caesarean section.

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

Figure 7: Example of a Classification rule conditional formula from the domain of midwifery.

5.2.5 Summary and Conclusion

The RA enhanced and extended KDD method identifies realistic properties from real data, providing improved input data quality, constraints and generation algorithms used to generate 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 to validate realism in the generated data. Its use is discussed in the next section.

5.3 The HORUS Approach to Validation of the Realism of Synthetic Data

The validation approach incorporates five steps that analyse separate elements of the SDG method and resulting synthetic data. These steps are identified in the small boxes in Figure 8 and described in Table 5. 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.

5.3.1 Input Validation

Input validation concerns itself only with that knowledge presented in the form of data tables and statistics that come from the generation specification. The input validation process verifies each item, confirming that correct input data is being presented, ensuring smooth operation of the data synthesis process (Bex et al., 2006). Input validation is intended

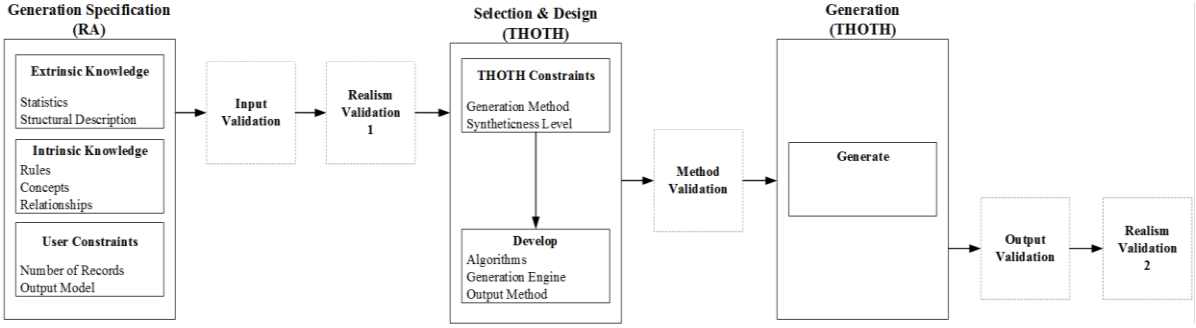

Figure 8: The HORUS Approach to Realism validation showing touch points with THOTH and RA.

Table 5: Activities and tasks in the HORUS Approach to synthetic data validation - a further elaboration of the HORUS Approach shown in Figure 8.

Step	Activity	Task
1	Input Validation	Verify each piece of input data or information; Confirm correctness & validity of input data & information
2	Realism Validation – I (RV1)	Verify concepts & rules derived from the KDD process & health statistical information applied; Review & test premise & accuracy of each rule to ensure consistency with domain semantics Tests rules and semantics in real circumstances to eliminate irrelevancy due to interaction with observed data
3	Method Validation	Review method and compare with others found in literature; Ensures chosen method is appropriate for generating the synthetic data; and Verify that the algorithm for the method to be used has been correctly and completely constructed
4	Output Validation	Establish that output of the SDG model are consistent with observational data; and Ensure that synthetically generated data conforms to qualitative and quantitative aspects derived during the knowledge discovery phase.
5	Realism Validation – II (RV2)	Perform the same tasks as for Realism Validation – I (RV1)

to prevent corruption of the SDG process (Laranjeiro et al., 2009).

For example, CoMSER used CPGs along with treatment and outcome statistics. Input validation necessitated ensuring that statistics could be located or extracted that correctly applied to each part of the process described by the CPGs. Cross-validation of those statistics was performed through comparison to more than one source. Clinicians were involved to ensure that where any difference in terminology existed between input datasets, correct linkage had been applied (McLachlan et al., 2016).

5.3.2 Realism Validation 1

The first realism validation process verifies the concepts and rules derived from the HCI-KDD process, along with any statistical knowledge applied. Realism validation reviews and tests the premise and accuracy of each rule to ensure consistency with the semantics of any data or guidelines used in their creation, such as CPGs. Where required it tests them in real circumstances to ensure they are not rendered irrelevant through interaction with observed data. Where any knowledge is at issue, the researcher should return to the knowledge discovery phase.

5.3.3 Method Validation

Method validation reviews the efforts of others inside and outside of the research domain. Attention should be paid to methodological approaches common for the domain, as well as methods other domains have used for similar types of SDG. Assessing these methods ensures the chosen method is appropriate for generating the synthetic data. Method validation also verifies that the algorithm to be used has been correctly and completely constructed.

5.3.4 Output Validation

Output validation validates the output data and verifies its basic statistical content. This step demonstrates the difference between the terms validation and verification. Oreskes describes validation as ensuring the model is free from known or detectable flaws and is internally consistent (Oreskes et al., 1994). Verification seeks to establish that 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 qualitative and quantitative aspects derived during the knowledge discovery phase.

5.3.5 Realism Validation 2

The second realism validation process performs all of the same tests as the first except that the tests are performed against the synthetic dataset. This ensures the synthetic data is consistent with the knowledge (rules, constraints and concepts) that were derived and used for its creation. The second realism validation step is the most important for establishing and justifying any claim that the synthetic data is a realistic and proper substitute for the real data it was created to replace. If a synthetic patient was treated in a manner that contradicted the principles or application of a CPG, this could invalidate the entire dataset. It is necessary for clinician validation to present the synthetic EHR in a clinician-familiar manner, as shown in Figure 9 from the CoMSER midwifery case study used in this work.

Figure 9: Sample realistic synthetic EHR.

5.3.6 Summary and Conclusion

The HORUS validation approach establishes and justifies claims of realism in synthetic data. The efficiency of HORUS comes from utilising the knowledge extracted in the RA process. It is this approach which supports realism claims, as it is one which directly compares the extrinsic and intrinsic elements of synthetic data with knowledge and generation specifications previously learned from the real data.

6 DISCUSSION

Realism is a collection of two levels of knowledge: (1) the extrinsic knowledge, and (2) the intrinsic knowledge. The extrinsic knowledge includes the overt structural and readily observable statistical information, while the intrinsic knowledge provides a detailed picture of the covert concepts, relationships, and interdependencies contained within the input materials. The ATEN framework presented in this paper ensures a complete analysis of requirements, source data and the SDG method. More information can only improve the generation approach, and a better generation approach delivers better synthetic data. The knowledge that is extracted and documented also provides a solid base with which to validate the synthetic data that has been created, ensuring that it is an adequate replacement for real data. The elements of realism can be identified through the engagement of an extended knowledge discovery in databases (KDD) approach. This approach first establishes the quantitative and qualitative aspects of the input data (the extrinsic), and follows it with in-depth and structured investigation of the concepts, relationships and rules that exist within the data (the intrinsic). The intrinsic results are then integrated with statistics from the extrinsic phase for the purpose of realism in the synthetic data generation phase.

The process of validating realism requires that each of the identified realism elements can be found in its correct form within the synthetic data. Validation follows the same flow as knowledge discovery, in that the quantitative and qualitative aspects are first assessed. If accurately established, each element of component knowledge that was established is verified. In this way the validation process is greatly simplified through the benefits gained by having identified the realistic elements of knowledge prior to generation. At any step of the validation process a return to one of the previous steps to address any identified issues is possible. If successful, validation supports claims of correctness of the SDG model, synthetic data and realism that exists within the data. As discussed, many authors do not describe validation of their SDG method and results. Validation of SDG methods would go some way to support the claims of realism that are encountered in the literature. Use of the complete realism identification and validation approach presented here will not only provide guidance throughout the SDG lifecycle, but also allow for demonstration of faithful adherence to established scientific methods. Finally, it empowers future SDG creators with the requisite knowledge to utilise the SDG method being described.

7 FUTURE WORK

The most pressing direction for future work is further validation of the ATEN framework in different domains. A new SDG project should be commissioned and undertaken as two streams: the first would follow the common approach to SDG without benefit of the ATEN framework. The second would use the same raw input data but adhere to the ATEN framework's approach. The resulting synthetic datasets should be comparatively validated to assess which is more successful through its proximity to real or observed data and its accuracy to being realistic.

8 SUMMARY AND CONCLUSION

The attainment and validation of realism in modern synthetically generated datasets represents a complicated challenge that is largely unaddressed in the literature. Many authors claim to have created realistic synthetic data yet few even discuss simple validation of their realism proposition. The ATEN framework presented in this paper is a new, complete, and comprehensive realism characterization and validation solution. If SDG approaches and methods in the literature had used and presented any one validation method, then their claims of having created realistic synthetic data could be given more credence. The approach presented is simple and not overly burdensome. Many of the component steps are activities that data synthesizers may already be undertaking, albeit unstructured or in an unconsidered way that is not deliberately aimed at attaining high levels of realism.

ACKNOWLEDGEMENTS

SM acknowledges support from the EPSRC under project EP/P009964/1: PAMBAYESIAN: Patient Managed decision-support using Bayes Networks. For Danika, Thomas, Liam and James.

REFERENCES

Agrawal, D., Butt, A., Doshi, K., Larriba-Pey, J., Li, M., Reiss, 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.

Alessandrini, M., De Craene, M., Bernard, O., Giffard-Roisin, S., Allain, P., Waecheter-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.

Ascoli, G., Krichmar, J., Nasuto, S. & Senft, S. 2001. Generation, Description And Storage Of Dendritic Morphology Data. *Philosophical Transactions Of The Royal Society Of London*, 365**,** 1131-1145.

Barnes, C. A. 1990. Concepts Hierarchies For Extensible Databases. Monterey, Ca: Naval Postgraduate School.

Barse, E., Kvarnstrom, H. & Jonsson, E. Synthesizing Test Data For Fraud Detection Systems. 19th Annual Computer Security Applications Conference, 2003. 384-395.

Bex, G., Neven, F., Schwentick, T. & Tuyls, K. Inference of Concise Dtds From Xml Data. 32nd Int. Conference on Very Large Databases, 2006. 115-126.

Bolon-Canedo, V., Sanchez-Marono, N. & Alonso-Betanzos, A. 2013. A Review of Feature Selection Methods on Synthetic Data. *Knowledge Information Systems*, 34.

Bozkurt, M. & Harman, M. 2011. Automatically Generating Realistic Test Input From Web Services. *6th International Symposium on Service Oriented System Engineering*. IEEE.

Brinkhoff, T. 2003. Generating Traffic Data. *IEEE Data Engineering Bulletin*, 26**,** 19-25.

Brissette, F. P., Khalili, M. & Leconte, R. 2007. Efficient Stochastic Generation of Multi-Site Synthetic Precipitation Data. *Journal of Hydrology*, 345**,** 121-133.

Carley, K. 1996. Validating Computational Models. Carnegie Mellon University.

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.

Dube, K. & Gallagher, T. Approach and Method For Generating Realistic Synthetic Electronic Healthcare Records For Secondary Use. International Symposium on Foundations of Health Informatics Engineering and Systems, 2014 Berlin, Heidelberg. Springer, 69-86.

Efstratiadis, A., Dialynas, Y., 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.

Fayyad, U. M., Piatetsky-Shapiro, G. & Smyth, P. 1996. Knowledge Discovery and Data Mining: Towards A Unifying Framework. *Kdd*, 96**,** 82-88.

Fernandez-Arteaga, V., Tovilla-Zarate, C., Fresan, A., Gonzalez-Castro, T., Juarez-Rojop, I., Lopez, I., 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.

Gafurov, T., Usaola, J. & Prodanovic, M. 2015. Incorporating Spatial Correlation Into Stochastic Generation Of Solar Radiation Data. *Solar Energy*, 115**,** 74-84.

Ganter, B. & Willie, R. 1997. Applied Lattice Theory: Formal Concept Analysis. *In:* Gratzer, G. (Ed.) *In General Latice Theory.* Birkhauser.

Giannotti, F., Mazzoni, A., Puntoni, S. & Renso, C. Synthetic Generation Of Cellular Network Positioning Data. 13th Annual Acm International Workshop on Geographic Information Systems 2005. Acm, 12-20.

Gore, R. 2001. Pharaohs of The Sun. *National Geographic*, 199**,** 35-57.

Gunn, B. 1923. Notes on The Aten And His Names. *The Journal of Egyptian Archaeology*, 9**,** 168-176.

Han, J., Cai, Y. & Cercone, N. 1993. Data-Driven Discovery of Quantitative Rules In Relational Databases. *IEEE Transactions on Knowledge And Data Engineering*, 5.

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.

Houkjaer, K., Torp, K. & Wind, R. 2006. Simple and Realistic Data Generation. *Vldb '06.*

Jaderberg, M., Simonyan, K., Vedaldi, A. & Zisserman, A. 2014. Synthetic Data and Artificial Neural Networks For Natural Scene Text Recognition. *Arxiv:1406.2227.*

Killourhy, K. & Maxion, R. Toward Realistic And Artefact-Free Insider-Threat Data. 23rd Annual Computer Security Applications Conference (Csac), 2007.

Laranjeiro, N., Vieira, M. & Madeira, H. Improving Web Services Robustness. IEEE Int. Conference on Web Services Icws'09, 2009. 397-404.

Laymon, R. 1984. The Path From Data To Theory. *In:* Leplin, J. (Ed.) *Scientific Realism.* University Of California Press.

Lydiard, T. 1992. Overview of The Current Practice And Research Initiatives For The Verification And Validation Of Kbs. *The Knowledge Engineering Review*, 7.

Mclachlan, S., Dube, K. & Gallagher, T. 2016. Using Caremaps and Health Statistics For Generating The Realistic Synthetic Electronic Healthcare Record. *International Conference on Healthcare Informatics (Ichi'16).* Chicago, USA: IEEE.

Mcmullin, E. 1984. A Case For Scientific Realism. *In:* Leplin, J. (Ed.) *Scientific Realism* University of California Press.

Mitra, S., Pal, S. & Mitra, P. 2002. Data Mining In Soft Computing Framework: A Survey. *IEEE Transactions on Neural Networks*, 13.

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.*

Mwogi, T., Biondich, P. & Grannis, S. An Evaluation of Two Methods For Generating Synthetic Hl7 Segments Reflecting Real-World Health Information Exchange Transactions. Amia Annu Symp Proc, 2014.

Oreskes, N., Shrader-Frechette, K. & Belitz, K. 1994. Verification, Validation and Confirmation of Numerical Models In The Earth Sciences. *Science* 263.

Oxford. 2016. *Definition of Validation In English* [Online]. Available: Https://En.Oxforddictionaries.Com/Definition/Us/Validation [Accessed].

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.

Putnam, H. 1977. Realism and Reason. *Proceedings and Addresses of The American Philosophical Assoc*, 50**,** 483-498.

Richardson, I., Thomson, M. & Infield, D. 2008. A High-Resolution Domestic Building Occupancy Model For Energy Demand Simulations. *Energy and Buildings*, 40**,** 1560-1566.

Rodriguez-Jiminez, J., Cordero, P., Enciso, M. & Rudolph, S. 2016. Concept Lattices with Negative Information: A Characterisation Theorem. *Information Sciences*, 369.

Sanderson, M. & Croft, B. Deriving Concept Hierarchies From Text. 22nd Annual International Acm Sigir Conference on Research And Development In Information Retrieval, 1999. Acm, 206-213.

Sperotto, A., Sadre, R., Van Vliet, F. & Pras, A. A Labelled Data Set For Flow-Based Intrusion Detection. 9th IEEE International Workshop on IP Operations And Management (Ipom '09), 2009. 39-50.

Stratigopoulos, H., Mir, S. & Makris, Y. 2009. Enrichment Of Limited Training Sets In Machine-Learning-Based Analog/Rf Test. *Date '09.*

Tsvetovat, M. & Carley, K. 2005. Generation of Realistic Social Network Datasets For Testing of Analysis And Simulation Tools. DTIC.

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.

Weston, J., Bordes, A., Chopra, S., Rush, A., Merrienboer, B., Joulin, A. & Mikolov, T. 2015. Toward AI Complete Question Answering: A Set Of Prerequisite Toy Tasks. Under Review as A Conference Paper At ICLR.

Whiting, M., Haack, J. & Varley, C. Creating Realistic, Scenario-Based Synthetic Data For Test And Evaluation Of Information Analytics Software. 2008 Workshop On Beyond Time And Errors: Novel Evaluation Methods For Information Visualisation (Beliv'08), 2008.

Wilkinson, R. 2008. Anthropomorphic Deities. *Ucla Encyclopedia of Egyptology*, 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.

Wu, X., Wang, Y. & Zheng, Y. 2003. Privacy Preserving Database Application Testing. *WPES '03.*