Approach and Method for Generating Realistic Synthetic Electronic Healthcare Records for Secondary Use

Kudakwashe Dube[1] and Thomas Gallagher[2]

[1] School of Engineering and Advanced Technology, Massey University, New Zealand
[2] Applied Computing and Electronics, University of Montana, Missoula, USA
K.Dube@Massey.ac.nz, Thomas.Gallagher@UMontana.edu
http://www.massey.ac.nz, http://ace.cte.umt.edu

Abstract. This position paper presents research work involving the development of a publicly available *Realistic Synthetic Electronic Healthcare Record* (RS-EHR). The paper presents PADARSER, a novel approach in which the real *Electronic Healthcare Record* (EHR) and neither authorization nor anonymisation are required in generating the synthetic EHR data sets. The *GRiSER method* is presented for use in PADARSER to allow the RS-EHR to be synthesized for statistically significant localised synthetic patients with statistically prevalent *medical conditions* based upon information found from publicly available data sources. In treating the synthetic patient within the GRiSER method, clinical workflow or *careflows* (Cfs) are derived from *Clinical Practice Guidelines* (CPGs) and the standard local practices of clinicians. The Cfs generated are used together with health statistics, CPGs, medical coding and terminology systems to generate coded synthetic RS-EHR entries from statistically significant observations, treatments, tests, and procedures. The RS-EHR is thus populated with a complete medical history describing the resulting events from treating the medical conditions. The strength of the PADARSER approach is its use of publicly available information. The strengths of the GRiSER method are that (1) it does not require the use of the real EHR for generating the coded RS-EHR entries; and (2) the generic components for obtaining careflow from CPGs and for generating coded RS-EHR entries are applicable in other areas such as knowledge transfer and EHR user interfaces respectively.

Keywords: synthetic data, healthcare statistics, clinical guidelines, clinical workflow, electronic health record, knowledge modeling, medical terminology, ICD10, SNOMED-CT.

1 Introduction

As the healthcare industry continues its transition to the *Electronic Health Record (EHR),* the rigorous process of obtaining patient records remains a deterrent for clinical trainers, researchers, software system developers and testers. The EHR is a cradle-to-grave systematic collection of a patients health information

J. Gibbons and W. MacCaull (Eds.): FHIES 2013, LNCS 8315, pp. 69–86, 2014.

that is distributed across locations and computing platforms and includes a wide range of disparate data types including demographics, medical history, medication, allergies, immunisation status, laboratory test results, radiology images, vital signs, personal measurements (height, weight, age) and billing information, all of which are sharable across healthcare settings [9]. A complete EHR paints a holistic picture of a patient′s overall medical history and provides a chronological description of an individual′s medical conditions, procedures, tests, and medications.

The problem of generating synthetic data has been widely investigated in many domains [23] [16] [11][3]. Despite the chronic limitations on access to the EHR for secondary use due to privacy concerns, there are very few research efforts to-date directed on developing promising and low cost approaches to generating synthetic EHRs for secondary uses. Only about 5 major works have appeared in literature during the past 12 years [5][15][2][14][21] compared to more than 60 works in other domains. This scenario is not expected in a domain that is characterised by the highly sensitive nature of the information. The work of Buczak et al is of particular significance to this paper due to its comprehensiveness and the ingenuity in the method of generating synthetic EHRs. The method uses clinical care patterns to create a care model that guides the generation of synthetic EHR entries that have realistic characteristics [2]. The major weaknesses of Buczak et al′s method are: (1) the use of the real EHR, whose access is still subject to limitations; (2) at a very high-level, the method amounts to the anonymisation of the real EHR still raise concerns from advanced data mining techniques; and (3) the possibility of the existence of an inverse method to re-identify the real EHR used in the method, which could lead to potential privacy breaches.

This paper presents an approach and method that has been developed as part of early work in the investigation of the problem of generating a *Realistic Synthetic Electronic Health Record* (RS-EHR). A key aspect of this problem is the emphasis on no access to the real EHR as well as the use of freely and publicly available health statistics, *clinical practice guidelines* (CPGs) and protocols practiced by clinicians, and medical coding and terminology systems and standards. *Clinical workflow,* also known as *Careflow* (Cf) are the *workflows* (Wf) of a health unit. The Cfs involve steps and processes that a patient goes through in either one clinician-patient encounter or a series of these encounters in the process of disease management and patient care [7]. A CPG is a systematically developed set of statements that guides the clinician and his patient in making decisions and performing tasks as part of managing the patient′s health problem [19]. Cfs and CPGs are closely related in that clinical workflow, and hence Cfs, can be derived from CPGs [7]. RS-EHR data that is realistic could be generated from clinically realistic Cf. The hypothesis of our approach and method is that codified RS-EHR skeletons can be generated from publicly available health care statistics while the codified and usually non-codified textual content of these EHR can be derived from the Cf that can be extracted from CPG that is used to manage the disease or clinical problem.

The novelty of the PADARSER approach for generating the RS-EHR presented in this paper is based upon the use of publicly available information. The GRiSER method adopted in this approach involves the creation of a synthetic patient of demographic significance to a region. The generated synthetic patient is iteratively injected with a statistically significant medical condition that is associated with a relevant clinical guideline or protocol. The patient RS-EHR is populated with entries based upon the standardized clinician careflow extracted from the relevant guideline or protocol used in treating the injected medical condition. From the careflow based on clinical guideline or protocols, an associated workflow of treatments, procedures, and lab tests will be generated in order to populate the RS-EHR. The anticipated outcome of PADARSER approach and GRiSER method is the coded RS-EHR that any practising clinician from the location of statistical relevance would deem to be realistic as would be confirmed by using our assessment rubric. Following completion of the generation of the RS-EHR, we have defined a rubric for assessing the realistic characteristics of the RS-EHR that includes comparing the characteristics the clinician would find in an actual EHR.

This paper contributes a new approach and method for generating RS-EHR that incorporates novel techniques for ensuring that the RS-EHR generated is realistic and less costly to produce. The novel techniques are: (1) the statistical generation of the patients and the disease they suffer from; and (2) the generation of RS-EHR entries from clinical guideline-based careflow and medical terminology and coding systems and standards. The statistical basis of the approach takes into account the disease prevalences and probabilities from publicly available information. The benefits of our approach are: (1) no real EHR will be used since content is generated from statistics and clinical guidelines; (2) no anonymization techniques are required since no real EHR is involved; (3) no authorization or obligatory consent are required as no personal patient information will be used; and (4) all acquired information will be obtained through freely and publicly available statistical health data and standardized clinician guidelines, medical codes and terminologies.

This paper begins by examining our motivation for generating RS-EHR data sets. Related works are discussed followed by a high-level analysis of the data sets incorporated for the RS-EHR. We describe our approach in generating RS-EHRs in which we have named PADARSER. GRiSER is introduced as our method in populating RS-EHRs and the iterative algorithm used by GRiSER is documented. Derivation of careflow from CPGs are examined to populate the RS-EHR through coding and textual events. Lastly, before concluding the paper we present our evaluation process and detail the evaluation rubric that include four criteria areas each with multiple components for assessing the realistic aspect of the RS-EHR with the help of domain experts.

2 Background, Problem and Motivation

The rapid adoption of the EHR is taking place throughout the world as one component of a healthcare information technology initiative aimed at improving

patient care while reducing costs. In the United States, EHR adoption is being driven by the *Health Information Technology for Economic and Clinical Health* (HITECH) Act (2009) with oversight from the *Office of the National Coordinator* (ONC) for Health Information Technology, while in New Zealand implementation is taking place through the National Health IT Plan with oversight from the IT Health Board and the Ministry of Health. Privacy and confidentiality concerns limit access to actual EHRs for secondary use. Legal protection for patient privacy is addressed in the United States through the *Health Insurance Portability and Accountability Act* (HIPPA) (1996), while in New Zealand patient privacy rights are covered by the *Health Information Privacy Code (1995).* Although secondary use of the patient EHR is permitted, obtaining authorization to access actual records is rigorous and the anonymisation of records can be expensive. Our motivation in generating the RS-EHR is based upon the current challenges associated with obtaining EHR for secondary use, especially research work in *Health Informatics.* These challenges have continued unabated even in the presence of *advances in anonymisation techniques.*

Privacy concerns for the individual restrict the availability of EHRs, limiting secondary use. Addressing the problem of creating synthetic EHRs has been recognised to be the best solution to this limitation. Most works in the literature that address the problem of generating synthetic EHRs use approaches and methods that require access to the real EHR during the process of generating the synthetic EHRs and hence cyclically suffer from the same limitation that they seek to address.

For clinician trainers, the secondary use of the EHR serves as an invaluable teaching tool. The EHR describes all the elements of a successful treatment encounter. As patient careflow takes place in the clinic, the EHR documents details of treatment procedures, tests, and medications. The EHR is a patient case study demonstrating the unique careflow provided to the individual. The EHR can also serve as a knowledge transfer tool. In developed regions of the world, the best practices in careflow are regularly demonstrated by expert clinicians. These careflow practices are documented in the EHR. Sharing the EHR as a training tool to developing regions of the world where expertise is limited serves as a knowledge transfer tool. The knowledge of clinical best practices derived from expertise found in developed regions can be shared through the EHR.

Developing a synthetic EHR with the characteristics of the realistic patient careflow will function as a tool to meet the demands of clinician trainers and serve as a knowledge transfer tool to help developing regions of the world. The synthetic characteristic addresses all privacy concerns associated in obtaining confidential patient records without the expense of anonymisation processing.

3 Related Works

The problem of generating synthetic data for various purposes has been widely recognised and investigated in a wide variety of domains [23][16][11][3]. However, there is the glaring lack of the adequate investigation into addressing this

problem for EHRs despite the chronic limitations of access to the EHR due to privacy concerns. For example, following a literature search, limited to four main digital libraries, PubMed, ACM, IEEE and ScienceDirect, for works on synthetic data resulted in the review of 62 published articles of interest which were refined to a list of 42 articles specifically mentioning research involving synthetic data generation, yet only five articles published during the period 2000 to 2012 were identified to have completed work in the EHR domain. From the five articles identified in our search of relevance to synthetic EHR generation, Esteller mentions the use of synthetic and real data in comparing electro-encephalograms (EEG) waveform [5]. Macjewski et al describe a data set developed from emergency room patient data where temporal disease outbreaks can be injected based upon seasonal trends [15]. Lee et al describe exploratory work using synthetic data and real EHRs for use in analytic systems [14]. Raza and Clyde describe synthetic data from work done in developing a test-data creation tool for health data [21]. Buczak et al, in their work on modelling a flu-like epidemic, present two comprehensive methods: one for generating their synthetic background EHR; and another for generating their synthetic epidemic EHR based on the synthetic background EHR [2].

Among the five works that present methods of synthetic EHRs, Buczak et al′s work provide the most comprehensive method [2]. By using the real EHR, Buczak et al propose an algorithm for generating EHRs that include a strategy to inject a disease into a synthetic patient who already has an existing background EHR. Buczak et al later realised that the background EHRs also need to be synthetic due to further privacy limitations. Instead of avoiding privacy limitation problem, Buczak et al made their method to require access to the real EHR for deriving the clinical care patterns. A key aspect of Buczak et al′s method is the strategy that involves the injection of a disease into the synthetic patient after which care patterns are used to generate the synthetic EHR entries. The care patterns that are associated with the synthetic patient are derived from the real EHR belonging to the patient that has the least similarity or distance measure from the synthetic patient. At a more abstract level, what Buczak et al present for generating their background and epidemic synthetic EHR is another method for the anonymisation of the real EHR, which has not led to the abatement of privacy concerns and resulting limitations to accessing EHRs. The weaknesses of the work of Buczak et al are: (1) the assumption that the real EHR will be accessible for the purpose of generating the synthetic EHR; and (2) the possibility of the existence of an inverse algorithm that uses their similarity or distance measure in reverse to partially or completely re-identify the real EHR and hence the patient using the information in the synthetic EHR. This inverse algorithm could potentially compromise the privacy of the real patient. The ingenuity of the Buczak et al′s method is the inherent assurance that the resulting synthetic EHR has realistic characteristics by using clinical care patterns to generate the synthetic entries for the synthetic EHR. This ingenuity in Buczak et al′s method inspired the GRiSER method′s use of clinical workflow or careflow. The major difference being that the GRiSER method derives the careflow from CPGs from

which the RS-EHR entries are generated instead of from the *real EHR.* The PADARSER approach and GRiSER method presented in this paper differ from the work of Buczak et al in that (1) they do not make use of any actual patient data, rather they derive clinical workflow from CPGs; (2) they do not focus on a single disease injection but use health statistics to create a synthetic EHR that has statistically prevalent set of diseases or medical conditions, which could be more than one for each synthetic patient, for instance, diabetes patients may also statistically suffer from renal complications; and (3) as no actual patient data is ever used, the approach used in the GRiSER method is uniquely distinct.

Research works that examine CPGs and clinical workflow have focused mainly on expressing or specifying and executing the process logic in CPGs by using workflow and process languages and technologies [20][8]. Other works have recognised the common process-oriented characteristics and conceptualisations of CPGs with a view towards developing some form of unification of concepts and computational models and formalisms [7]. There is the wide recognition that CPGs can be expressed, at least partially, as clinical workflow [12][17] and that care pathways are goal- and process-oriented and so define the workflow of activities as well as roles and sequencing of activities while also providing a framework for generating data for the EHR [7]. The work presented in this paper seeks to exploit the process nature of CPGs to create the necessary context for the realistic nature of the generated RS-EHR entries.

Healthcare statistics are drawn for the coded aspects of the real EHR [18]. The health statistics contain information on diseases, medical procedures performed, medications and laboratory tests performed. The statistics also usually include the relevant medical codes and terminologies for systems and standards especially the *International classification of Diseases* (ICD) [25] and *Systematized Nomenclature Of Medicine Clinical Terms* (SNOMED-CT) [10]. Therefore, it is possible to extract coded diseases, laboratory tests, observations and medications from statistical information. While CPGs may make use of medical or clinical terms, they are not systematically coded using using medical coding systems such as the ICD-10 [25]. Furthermore, CPGs may also not systematically make use the appropriate terminologies from SMOMED-CT. One of the objectives of this work is to generate coded RS-EHR entries that also systematically make use of terminology systems.

4 Generating Realistic Synthetic E-Healthcare Records

Secondary uses of the EHR such as training clinicians and testing of software under development do not require access to the real EHR. For such forms of secondary use of the EHR, a realistic synthetic EHR would be enough. *Synthetic data is data that is created to simulate real data in the application domain of interest. The synthetic EHR is synthetic data that is generated to simulate the real EHR.* The synthetic EHR can be used in place of the real EHR in scenarios that involve the appropriate forms of the secondary use of the EHR. To be usable, the synthetic EHR must be realistic, which is the single most important

characteristic of it whose test would be that a clinician examining the record would not be able to tell that the EHR is synthetic. This work investigates the problem of creating the realistic synthetic EHR (RS-EHR) and this paper presents the approach and method developed in this work for generating the realistic synthetic EHRs (RS-EHRs).The novelty of the approach and method is in (1) the use of publicly available data such as public health statistics, clinical guidelines and protocols, and medical coding and terminology standards and systems, and (2) their assumption that there is no real EHR that would be available for access during the generation of entries that would make up the resulting RS-EHR. Figure 1 presents the relevant aspects that are important in the problem of generating realistic synthetic EHRs.

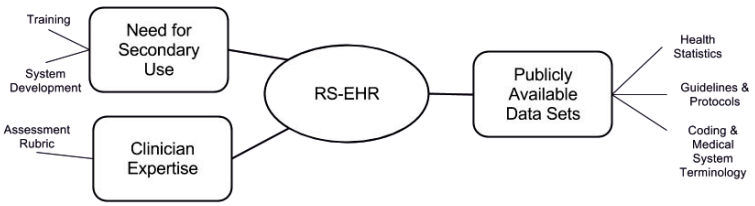

Fig. 1. Relevance diagram presenting the relevant aspects for the creation of the realistics synthetic EHR as investigated in this paper

Publicly available health information and knowledge are important sources that could contribute towards extrapolations that would be used in the generation of data for synthesising the realistic synthetic EHR. For example, health statistics would contribute information for generating patient demographics and disease prevalences as well as medical procedures, medications and laboratory tests. Clinical guidelines and protocols (CPGs) would assist by suggesting the careflow involved in the management of patient problems. Furthermore, the CPGs, when used in an appropriate method to be developed in this work combined with medical coding and terminology systems, would provide the material basis for generating coded textual narrative aspects of the synthetic EHR. Medical experts or clinicians are of significance to this work at the step for the evaluation of the resulting RS-EHR. As pointed out earlier in this section, the resulting RS-EHR must pass the test that establishes the fact that when a practising clinician from the region of statistical significance examines the RS-EHR, its should be that the RS-EHR is indistinguishable from, and have the typical characteristics of, a real EHR that the practising clinician would normally encounter in his or her daily work routine. It would appear from this discussion that the RS-EHR for some secondary use purposes could be generated from publicly available information such that practising clinician could deem the resulting RS-EHR to be realistic.

5 The PADARSER Aproach: *Generating Realistic Synthetic E-Healthcare Records*

The *Publicly Available Data Approach to the Realistic Synthetic EHR* (PADARSER), is the approach presented in this section, whose goal is to exploit publicly available data and information, without access to real EHRs, in generating the RS-EHRs. In particular, the three main types of publicly available data and information that have been selected in the approach for such exploitation are illustrated in Figure 2.

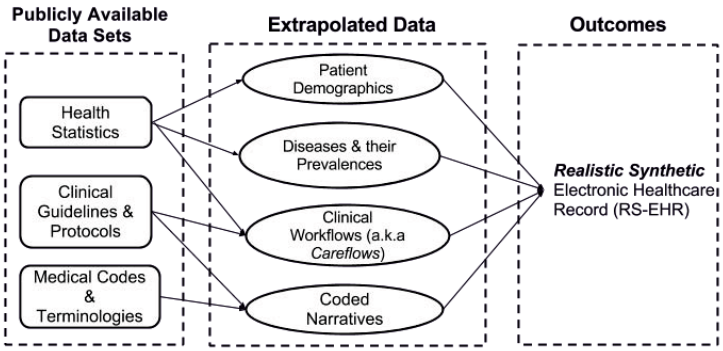

Fig. 2. The PADARSER approach to generating the realistic synthetic EHR from publicly available data and information

As can be seen in Figure 2, the three types of publicly available information that has been selected in this work for use in creating the RS-EHR are: (1) health statistics; (2) clinical guidelines and protocols (CPGs); and (3) medical coding and terminology systems and standards. Figure 2 also illustrates the sub-types of information that can be extrapolated from each of these three main types of information. Figure 2 further elaborates and presents the conceptual approach, from an data and information derivation and extrapolation perspective, for building the RS-EHR. In other words, Figure 2 illustrates the synthetic data for the RS-EHR that would be derived or extrapolated from the publicly available information sources in the approach presented in this paper.

Health statistics for a country or a region within a country provide detailed statistical health information that covers the following aspects of the synthetic EHR: patient demographics, encountered diseases occurring in the population in focus and their prevalences, medications administered to patients, laboratory tests and medical procedures that were performed on patients during the timeframe in focus. Clinical practice guidelines (CPGs) and protocols for specific disease and health problems are useful in the process of generating actual RS-EHR data entries because clinical workflow (Wf) or careflow (Cf) can be derived from these CPGs [7]. The generated Cf contains the appropriate clinical and healthcare events, tasks and procedures that then become the basis for

generating the realistic data that goes into the RS-EHR. Synthetic textual narratives to describe the patients clinical conditions, observations and procedures can also be extracted from CPGs for entry into the RS-EHR. Medical codes are important in that most health statistics are based on coded aspects of real EHRs and in that the ideal EHR is one that has extensively coded data [1]. Furthermore, medical coding systems are closely related to, or make use of medical terminologies, which are also implicitly or explicitly used in CPGs. Of particular interest to this paper is the generation of (1) coded synthetic EHR entries and (2) coded synthetic textual narratives in these entries, both of which will make use of medical coding standards, medical terminology systems and CPGs as key data sources in the PADARSER approach.

6 The GRiSER Method for Generating the RS-EHR

This section presents, at a conceptual level, the method for generating the RS-EHR from public available data and information sources. As already pointed out earlier in this paper, all the data that would be required to synthesise the RS-EHR are derived or extrapolated from public health statistics (PHS), clinical guidelines and protocols (CPGs) and medical coding and terminology systems and standards (MCTSS).

6.1 Generating the RS-EHR

The GRiSER Method presented in this section is the method adopted for the PADARSER Approach to generating the RS-EHR. GRiSER stands for Generating the *Realistic Synthetic EHR* (RS-EHR). The GRiSER Method aims at using data from the publicly available data that has been presented in the previous sections to synthesise the RS-EHR. The resulting RS-EHR should be typical of the region from which the publicly available information in obtained. The GRiSER Method seeks to attain this aim without using any information or knowledge from examining the real EHR, which is assumed to be unavailable. The expected outcome of the GRiSER Method is a RS-EHR that would be deemed to be realistic by a practising clinician. Figure 4 presents the conceptual illustration of the method for generating the RS-EHR from publicly available information. In the GRiSER Method, synthetic patients are generated from statistical information. Patient demographics are generated from the statistics and added to the RS-EHR. Each synthetic patient is iteratively injected with a disease extrapolated from publicly available data sources including disease prevalences in the population of interest. Each disease is associated with a specific guideline or protocol from which the clinical workflow for managing the disease is derived. Careflow based on CPGs provide a realistic context for clinical events, which, in turn provide a realistic context for generating synthetic entries for the RS-EHR. Therefore, in the GRiSER Method, CPGs and the careflow that is generated from them are critical in rendering the generated realistics. For each RS-EHR entry synthesised, appropriate codes from the standardised medical coding system are

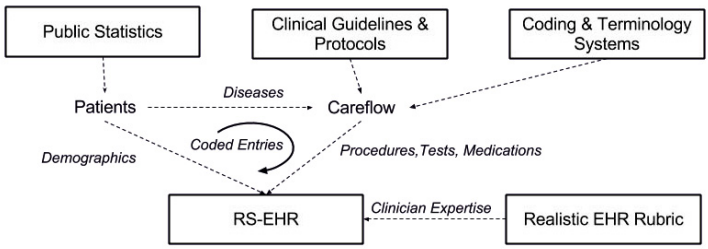

Fig. 3. Conceptual architecture of the GRiSER Method for generating the RS-EHR from publicly available information

attached. The synthetic entrys textual narratives are generated from appropriate textual narratives in the CPG. These narratives are improved by using the appropriate medical terminologies from medical terminology systems and are also coded appropriately. The various types of entries that would be synthesised from careflow events include clinical observations, procedures performed, laboratory test results and medications prescribed. Once the whole RS-EHR for the synthetic patient is fully generated, a RS-EHR rubric for assessing its validity could be applied by a practised clinician to establish that the RS-EHR is indeed realistic.

6.2 Description of the GRiSER Method

Figure 4 presents the activity diagram that illustrates the GRiSER method. The Activity Diagram of Figure 4 illustrates that the GRiSER Method starts with the preparation of the publicly available local health statistics before using the statistics to create synthetic patient demographics, which is then used to create the patient′s synthetic EHR. The patient′s synthetic EHR holds only the patient demographics at this stage.The demographics are then used to generate a set of health problems that will be reflected in the synthetic EHR for the patient. Figure 4 illustrates a loop to iterate over this set of complaints in order to generate coded synthetic EHR entries for each complaint with the entries covering the same period as that covered by the health statistics. This time period may cover from one year to ten years. The number of complaints to be considered could be limited subject to computation challenges. A model, as illustrated in Figure 4, is generated to guide the temporal aspects of the generated synthetic EHR entries. A part of the Activity Diagram illustrated the selection of a CPG for each health complaint and the generation of the careflow based on the CPG. The careflow then determines the set of events for whose synthetic entries need to be added to the RS-EHR. The activity of generating coded synthetic entries is iterative over all events and makes use of medical terminology and coding systems. Listing 1 presents a high-level algorithm to illustrate the method described here.

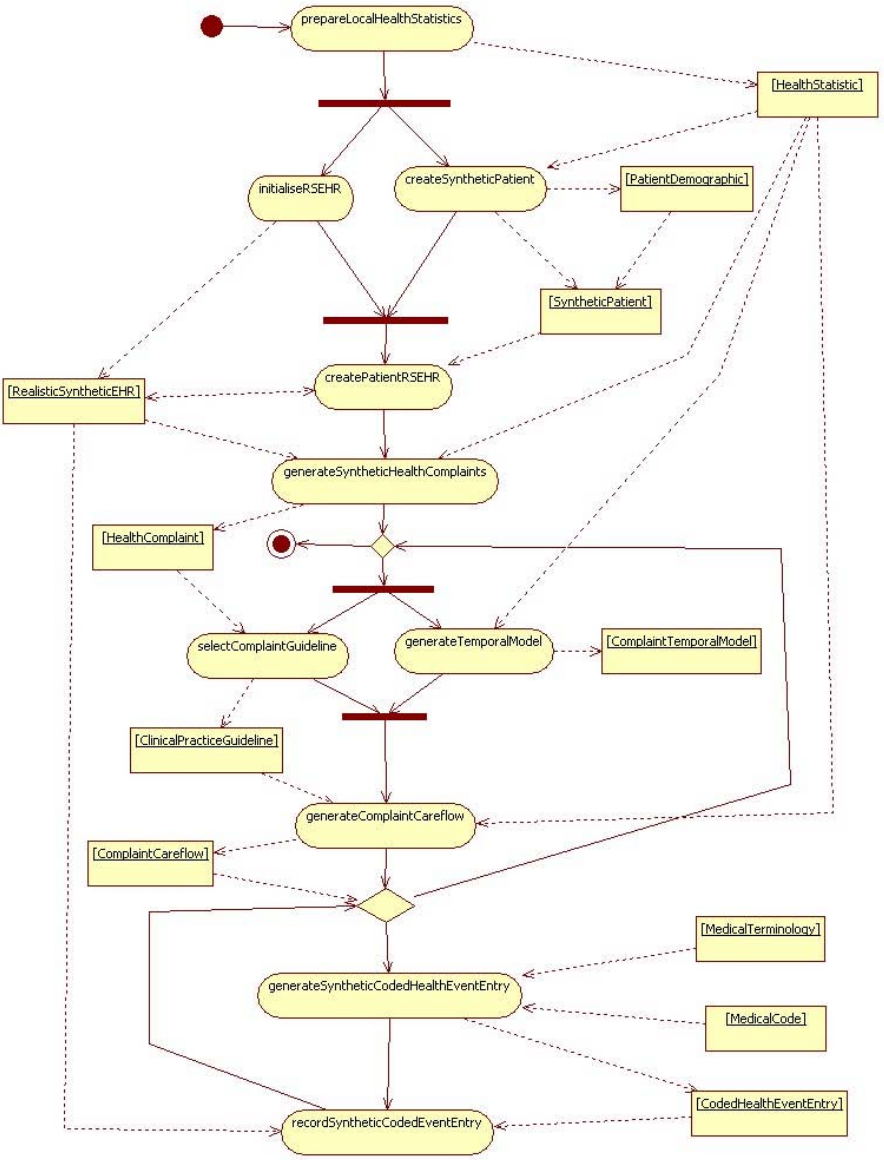

Fig. 4. The UML Activity Diagram illustrating the main activities in the GRiSER method for generating the realistic synthetic EHR from publicly available information.

Listing 1. The GRiSER Algorithm - the high-level algorithm for generating the RS-EHR from publicly available information in the GRiSER Method

```
*0  GRiSER(healthStats, cpgLibrary, medCodes, medTerms, complaintsLimit)
1  // GRiSER - Generating the Realistic Synthetic E-healthcare Record
1  // output: patientRSEHR
2  // input parameters: healthStats, cpgLibrary, medCodes, medTerms, complaintsLimit
3  patientRSEHR ⟵ initRSEHR();
4  patient ⟵ initPatient();
*6  patient.demographics ⟵ genDemographics(healthStats);
7  patientRSEHR.patient ⟵ patient;
8  complaintsCount ⟵ 0
+9  A: repeat until (complaintCount ≤ complaintsLimit)
10  // iteratively inject patient with statistically significant clinical
11  // complaint or disease
*12    clinicalComplaint ⟵ genComplaint(healthStats,patient);
*13    cpg ⟵ selectComplaintCPG(clinicalComplaint, patient, cpgLibrary);
*14    temporalModel ⟵ initTemporalModel(healthStats);
*15    careflow ⟵ genClinicalWorkflow(healthStats, cpg, patient, temporalModel);
16    cfEvents ⟵ careflow.getEvents();
+17    B: repeat until (cfEvents.iter.hasMore()  ==  false)
18  // iteratively generate coded RS-EHR entries from careflow events:
19  // coded textual narratives are generated from CPGs, medical codes and
20  // terminologies
21  cfEvent ⟵ cfEvents.iter.next();
*22 codedEntries ⟵ genCodedEventEntries(cfEvent, cpg, medCodes, medTerms);
23  patientRSEHR.entries.add(codedEntries);
+24    B: end
25  complaintsCount ⟵ complaintsCount + 1;
+26 A: end
27 return patientRSEHR;
```

The GRiSER algorithm presented in Listing 1 creates the RS-EHR for a single synthetic patient. The algorithms returns a RS-EHR accepting the following parameters:

- `healthStats` - the complete set of publicly available health statistics that are relevant for generating RS-EHR;
- `cpgLibrary` - the library of clinical practice guidelines and protocols for diseases that are statistically prevalent in the population;
- `medCodes` - the codes from a standardised medical coding system, e.g., ICD10;
- `medTerms` - the terminology from a standardised medical terminology system, e.g. SNOMED; and
- `complaintsLimit` - ensures that the synthetic patient is injected with a realistic number of clinical complaints.

In Listing 1, the lines marked with the asterisk (*) are the important steps in the GRiSER algorithm while the lines marked with the plus (+) indicate the main iterations. In Line 6, the function, `genDemographics(healthStats)`, generates synthetic patient demographics. Line 9 presents the starting point for the iteration over the patients clinical complaints. In Line 9, the synthetic clinical complaint is generated for the synthetic patient based on statistical information. In Line 13, the clinical practice guideline, cpg, is selected for managing the patients clinical complaint. The GRiSER algorithm, as presented here, assumes

that there exists a CPG for every clinical complaint that could arise statistically. This assumption is presented here as a simplification that may not appear in the implementation. In Line 14, a temporal model for each clinical complaint is initialised based on the temporal issues, such as seasonal considerations, that could be deduced from health statistical data. A temporal model that is derived initially from statistical data and enhanced with CPG information for each clinical complaint is important to enhance the realistic characteristics of the generated careflow. In Line 15, the clinical workflow or careflow is generated from the CPG form the synthetic patient taking into account statistical information as well as temporal issues. The careflow will provide the clinical events that would guide the generation of RS-EHR entries. In Line 17 an inner loop is initiated to iterate over the clinical events from the careflow to create all the RS-EHR entries that arise from the event.. In Line 22, coded entries for the RS-EHR are synthesised by using CPG narratives, medical codes and terminologies.

There are two major challenges within the GRiSER algorithm that are very significant from the modelling and computational hardness point of view, and of wider applications. The complete and detailed analysis of modelling and computational analysis of these challenges are subjects of currently on-going work and will not be presented in this paper. These challenges are: (1) the derivation of clinical workflow or careflow from clinical practice guidelines and protocols handled by the function, `genClinicalWorkflow(healthStats, cpg, patient, temporalModel)` in the GRiSER algorithm; and (2) the generation of coded synthetic RS-EHR entries from events, CPGs and medical coding and terminology system, handled by the function, `genCodedEventEntries(cfEvent, cpg, medCodes, medTerms)`. The next two subsections, Section 6.2 and 6.3, present and elaborate further on these two challenges without going into the the analysis of the modelling and computational complexity, which will not be presented in this paper as pointed out earlier.

6.3 Deriving Clinical Workflow or Careflow from Clinical Guidelines

Relationship between clinical practice guidelines (CPG) and clinical workflow, also known as careflow, has been identified in the area of CPG computerisation. CPGs structure best practice and clinical workflow while also assisting clinicians in diagnosis and treatment [4]. Laleci and Dogac [13] claims that computerised CPGs could be used to drive computer-based clinical workflow because they communicate with external applications to retrieve patient data and then initiates medical actions in the clinical workflows. Workflow technologies have also been used to computerise CPGs [24][20][8]. It has also been recognised that clinical workflows have demonstrated to be an effective approach to partially model CPGs [12]. As a result of the need to customise CPGs to suite a patient at a specific location,CPGs are known not to completely define clinical workflow. Hence, Juarez et al [12] proposed a workflow fulfilment function to determine the degree of completeness of careflow derived fromm CPGs. More recently, Gonzalez-ferrer et al [6] demonstrated the translation of CPGs into Temporal Hierarchical Task Networks (THTN), which facilitates the automatic generation of time-annotated

and resource-based clinical workflow. Thus, it has been shown from the literature that it is feasible to obtain clinical workflow or careflow from CPGs.

Generating clinical workflow or careflow from CPGs is a key feature of the PADARSER approach and the GRiSER method. The derivation of careflow from clinical guidelines is an important part of the GRiSER method that distinguishes it from other methods of generating RS-EHRs found in the literature. For example, in the work of Buczak et al [2], care patterns for generating synthetic is derived from inferences that result from examining entries recorded in the real EHR, which does not apply where there is absolutely no access to the real EHR.

Besides exploiting the process nature of CPGs in generating synthetic entries for the RS-EHR, the GRiSER method for creating careflow from CPGs also has wider applications. For example, Figure 5 details the proposed process of using CPGs to generate careflow and it's proposed application as a knowledge transfer tool.

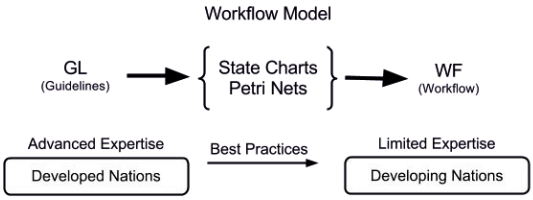

Fig. 5. Careflow knowledge transfer from clinical guidelines

The top part of Figure 5 presents a conceptualisation of the knowledge and technological transfer that is often a key goal for international bodies such as The United Nations and their constituent agencies. Methods, approaches and techniques are being sought by such bodies to facilitate the transfer of advanced expertise from developed regions of the world in forms that encapsulate best practices to developed region of the world where limited expertise is in acute shortage. This conceptualisation of knowledge and technology transfer could be exemplified or instantiated by the computer-assisted extraction of clinical workflow from the text of best practice guidelines as illustrated in the bottom part of Figure 6. Thus, best healthcare practice in the form of CPGs could be captured by using formal models that underlie workflow technology to create clinical workflow that could be customised to suit the conditions of the target developing world.

6.4 Generating Coded Textual Narrative from CPGs and Medical Coding and Terminology Systems and Standard

GRiSER method aims at incorporating a special strategy to extract relevant textual narratives from CPGs and then code them appropriately as well as apply terms from SNOMED-CT. Generating coded entries for the RS-EHR including

coded textual narrative is an important component of the GRiSER method that has other applications especially within the area of EHR user interfaces (UIs). Figure 6 illustrates the concept of using codes and coded textual narrative suggestions in EHR user interfaces.

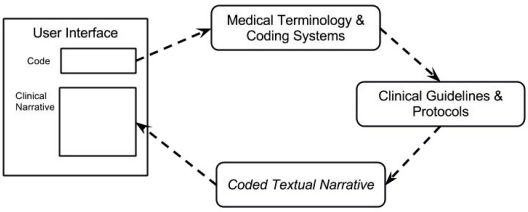

Fig. 6. Coded Textual Narrative for UI

The coded textual narratives are generated from CPGs using medical codes and terminology systems and presented as suggestions to the user, who could be allowed to accept or reject the suggested entries. The suggested coded entries would help the user to enter more accurate information as well as increase the use of medical codes within the EHR entries.

7 Evaluation, Outcomes and Applications

A rubric assessing the realistic aspects of the synthetic EHR has been developed. a high-level summary of this rubric is presented as the table in Table 1. Criteria elements include: (1) patient representation; (2) disease representation; (3) careflow; and (4) clinician acceptance. The record of the synthetic patient will be assessed to match the statistical accuracy of the demographic characteristic found in the regional area of interest. As an example, the synthetic patient in Africa would, in most cases, be of African ethnicity rather than European descent. A limitation on the number of disease injections will be restricted to a maximum value for any given patient. Injecting a patient with an unreasonable number of consecutive diseases will distort the natural prevalence normally found in real patients. Diseases will be limited to those which are reasonably prevalent to a particular region. For instance, the prevalence of malaria would not be considered common to patients in Canada thus unusual to include in the synthetic EHR from this region. The seasonal changes in weather can lead to variation in the prevalence of diseases and will be included in the assessment.

Careflow will be measured for both logical and temporal organization of clinical encounters. The overall completeness of the expected events following diagnosis of a particular disease will be included in the assessment. Lastly the expertise of a clinician reviewer will be used in our assessment. Examination by the expert will be used to determine whether the expected clinical encounters

Table 1. RS-EHR Assessment Rubric

No.	Criteria Area	Assessment Component
1	*Patient Representation*	(1) Regional demographic; and (2) Limit maximum injections
2	*Disease Representation*	(1) Regional disease; and (2) Consistent with seasonal changes
3	*Careflow*	(1) Logical event organization; (2) Temporal event placement; and (3) Comprehensive event inclusion
4	*Clinician Acceptance*	(1) Verification - EHR events follow guidelines and protocols; and (2) Verification - RS-EHR as a whole is realistic

found in the RS-EHR match those experienced in an actual patient EHR. Upon reviewing the RS-EHR, the clinician should be able to state that the events and encounters found in the record mimic those expected to be experienced by a real patient. The components of the assessment rubric are presented in Table 1.

An alternative approach in providing assessment for the realistic aspect of the synthetic EHR involves the use of automata learned from the real EHR. Models from automata processes constructed learned or mined from either clinical guidelines [22] or the real EHR provide novelty to assessing the realism characteristic of the RS-EHR. Computer Interpretable Guidelines (CIG) created through machine learning systems are capable of developing patterns of careflow [22]. Using CIG patterns crafted from the actual EHR, a further innovation for assessing the realistic aspect of the synthetic EHR will be worthy of consideration in this on-going work.

8 Summary, Future Work and Conclusion

This paper has presented the results of our early efforts in developing a framework in creating a realistic synthetic EHR for secondary usage. By incorporating publicly available data sets and clinical careflows, this paper has presented a novel strategy in generating EHRs. The RS-EHR alleviates confidentiality concerns and eliminates the use of anonymisation techniques. Proposed application for RS-EHRs include system developers and clinician trainers. The comprehensive medical history found in the patient EHR serves as a new format of teaching tool. It contains the clinical details for delivering rich case studies of patient treatment for training future clinicians.

Adoption of the EHR will continue to drive the desire for secondary use of patient information. Although the associated benefits from secondary use are well documented, the implementation will be slow due to patient privacy concerns, expensive anonymisation techniques, and limited data sets. Our work in developing the RS-EHR can provide highly available data sets of realistic patient information. In addition to standard methodologies employed in education, a new digital medium for case studies will be available in the form of a publicly

available EHR. The publicly available EHR data set provides a new medium for training the next generation of clinicians.

This paper provides a general blueprint of our investigative work in developing the RS-EHR. We envision several sub-projects to evolve in our future work. Some examples of future projects include: (1) data extrapolation from public data sources for supporting the PADARSER approach; (2) temporal data population of EHR encounters derived from careflow events; (3) implementation and further refinement of the GRiSER algorithm for RS-EHR; (4) further conceptualisation and implementation of the method for generating careflow from published clinical practice guidelines; and (5) generating coded textual narratives from CPGs and medical coding and terminology systems.

References

1. American Medical Association (AMA). CPT coding, medical billing and insurance (2013), http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance.page?
2. Buczak, A., Babin, S., Moniz, L.: Data-driven approach for creating synthetic electronic medical records. BMC Medical Informatics and Decision Making 10(1), 59 (2010)
3. Daud, H., Razali, R., Asirvadam, V.: Sea Bed Logging Aapplications: ANOVA analysis 2 for Synthetic Data From Electromagnetic (EM) Simulator. In: 2012 IEEE Asia-Pacific Conference on Applied Electromagnetics (APACE), pp. 110–115 (2012)
4. De Backere, F., Moens, H., Steurbaut, K., De Turck, F., Colpaert, K., Danneels, C., Decruyenaere, J.: Automated generation and deployment of clinical guidelines in the ICU. In: 2010 IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS), pp. 197–202 (2010)
5. Esteller, R., Vachtsevanos, G., Echauz, J., Lilt, B.: A comparison of fractal dimension algorithms using synthetic and experimental data. In: Proceedings of the 1999 IEEE International Symposium on Circuits and Systems, ISCAS 1999, vol. 3, pp. 199–202 (1999)
6. GonzáLez-Ferrer, A., Teije, A.T., Fdez-Olivares, J., Milian, K.: Automated generation of patient-tailored electronic care pathways by translating computer-interpretable guidelines into hierarchical task networks. Artif. Intell. Med. 57(2), 91–109 (2013)
7. Gooch, P., Roudsari, A.: Computerization of workflows, guidelines, and care pathways: a review of implementation challenges for process-oriented health information systems. J. Am. Med. Inform. Assoc. 18, 738–748 (2011)
8. Grando, M.A., Glasspool, D., Boxwala, A.: Argumentation logic for the flexible enactment of goal-based medical guidelines. J. of Biomedical Informatics 45(5), 938–949 (2012)
9. Grimson, J.: Delivering the electronic healthcare record for the 21st century. International Journal of Medical Informatics 64(2-3), 111–127 (2001)
10. International Health Standards Development Organization (IHSDO). Systematized nomenclature of medicine clinical terms, SNOMed-CT (2013), http://www.ihtsdo.org/snomed-ct/

11. Jeske, D.R., Lin, P.J., Rendon, C.: Rui Xiao, and B. Samadi. Synthetic data generation capabilities for testing data mining tools. In: IEEE Military Communications Conference, MILCOM 2006, pp. 1–6 (2006)
12. Juarez, J.M., Martinez, P., Campos, M., Palma, J.: Step-Guided Clinical Workflow Fulfilment Measure for Clinical Guidelines. In: Moreno-Díaz, R., Pichler, F., Quesada-Arencibia, A. (eds.) EUROCAST 2009. LNCS, vol. 5717, pp. 255–262. Springer, Heidelberg (2009)
13. Laleci, G.B., Dogac, A.: A semantically enriched clinical guideline model enabling deployment in heterogeneous healthcare environments. IEEE Transactions on Information Technology in Biomedicine 13(2), 263–273 (2009)
14. Lee, N., Laine, A.F.: Mining electronic medical records to explore the linkage between healthcare resource utilization and disease severity in diabetic patients. In: 2011 First IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology (HISB), pp. 250–257 (2011)
15. Maciejewski, R., Hafen, R., Rudolph, S., Tebbetts, G., Cleveland, W.S., Grannis, S.J., Ebert, D.S.: Generating synthetic syndromic-surveillance data for evaluating visual-analytics techniques. IEEE Computer Graphics and Applications 29(3), 18–28 (2008)
16. Margner, V., Pechwitz, M.: Synthetic Data for Arabic OCR System Development. In: Proceedings of the Sixth International Conference on Document Analysis and Recognition, pp. 1159–1163 (2001)
17. Milla-Millán, G., Fdez-Olivares, J., Sánchez-Garzón, I., Prior, D., Castillo, L.: Knowledge-driven adaptive execution of care pathways based on continuous planning techniques. In: Lenz, R., Miksch, S., Peleg, M., Reichert, M., Riaño, D., ten Teije, A. (eds.) ProHealth 2012 and KR4HC 2012. LNCS, vol. 7738, pp. 42–55. Springer, Heidelberg (2013)
18. New Zealand Ministry of Health (NZ-MoH). New Zealand Health Statistics: Classificiation and Terminology (2011), `http://www.health.govt.nz/nz-health-statistics/classification-and-terminology` (accessed: May 21, 2013)
19. Institute of Medicine (IOM). Guidelines for Clinical Practice: From Development to Use. National Academy Press, Washington DC (1992)
20. Peleg, M., Tu, S.W.: Design patterns for clinical guidelines. Artif. Intell. Med. 47(1), 1–24 (2009)
21. Raza, A., Clyde, S.: Testing health-care integrated systems with anonymized test-data extracted from production systems. In: 2012 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), pp. 457–464 (2012)
22. Riaño, D.: Ordered Time-Independent CIG Learning. In: Barreiro, J.M., Martín-Sánchez, F., Maojo, V., Sanz, F. (eds.) ISBMDA 2004. LNCS, vol. 3337, pp. 117–128. Springer, Heidelberg (2004)
23. Stark, E., Eltoft, T., Braathen, B.: Performance of Vegetation Classification Methods Using Synthetic Multi-Sspectral Satellite Data. In: International Geoscience and Remote Sensing Symposium (IGARSS 1995). Quantitative Remote Sensing for Science and Applications, vol. 2, pp. 1276–1278 (1995)
24. Tsai, A., Kuo, P.-H., Lee, G., Lin, M.-S.: Electronic clinical guidelines for intensive care unit. In: 2007 9th International Conference on e-Health Networking, Application and Services, pp. 117–124 (2007)
25. World Health Organisation (WHO). International Classificiation of Diseases (ICD), Web, `http://www.who.int/classifications/icd/en/` (accessed: May 21, 2013)