Digital Healthcare Empowering Europeans
R. Cornet et al. (Eds.)

doi:10.3233/978-1-61499-512-8-35

Simulating Realistic Enough Patient Records

James CUNNINGHAM[a,1] and John AINSWORTH[a]
[a] *Institute of Population Health, University of Manchester*

Abstract Information systems for storing, managing and manipulating electronic medical records must place an emphasis on maintaining the privacy and security of those records. Though the design, development and testing of such systems also requires the use of data, the developers of these systems, rarely also their final end users, are unlikely to have ethical or governance approval to use real data. Alternative test data is commonly either randomly produced or taken from carefully anonymised subsets of records. In both cases there are potential shortcomings that can impact on the quality of the product being developed. We have addressed these shortcomings with a tool and methodology for efficiently simulating large amounts of realistic enough electronic patient records which can underpin the development of data-centric electronic healthcare systems.

Keywords. patient simulation, software tool, computerized medical record

Introduction

The rise of 'Big Data' has lead to an increasing prevalence of information systems that need to process, analyse and manipulate large volumes of data [1, 2]. The development of such data-centric software applications naturally requires the use of data for development and testing, with the structure and semantics of the data that these systems deal with impacting on their architectural design. Aspects of design that impact on the efficiency of processing data, user interface and user experience are necessarily, at least in part, driven by the type and form of the data that the finished system will deal with. It can be the case that the data that such a system will deal with is not available to the developers of such software at design time; this can either be due to the exact data that will be processed by the software not being known at design time, or access to that data being limited to the software developers due to sensitivity or confidentiality issues. This second point is particularly true in the case of systems that deal with sensitive medical data.

Historically medical records have been paper-based, with the UK has recently moved towards an electronic representation [3], mirroring efforts in Europe, North America and Australia amongst others [4, 5]. In turn this has lead to increasingly strict legislative and regulatory frameworks being put in place to protect individuals medical

[1] Corresponding Author: james.a.cunningham@manchester.ac.uk

data and constraining the use of data to medical or research needs [6]. Since the developers of these applications are rarely also their end users they are then precluded from accessing the same data that such applications will process in their final form.

Test data that falls outside these regulations is generally either in the form of anonymised or purely random data (as rarely will developers gain the necessary ethical approval to access real data). Anonymised data sets take time and effort to produce and require originating source data that will be subject to those same ethical and governance requirements [7]. Purely random data on the other hand, whilst easy to produce, will have no structural or semantic integrity when interpreted as medical data. i.e. there will be no sense to the data produced. This can then lead to situations where system behaviour based on the semantics of the data being processed are not taken into account in the design of the system.

We address these issues of obtaining useable medical data for system development through the development of a tool for simulating what can be described as 'realistic enough' patient records. These capture the structural essence of patient records, for example precluding inconsistencies in the timings of medical events, or the interrelations between types of medical event, whilst not needing such computational power to run as to exclude the production of large amounts of data (for example into the millions of patient records). This work is outlined in detail in the remaining sections of this paper.

1. Methods

1.1. Simulation Model

The method we use to simulate patients is based on producing a description of all possible life events for a patient and selecting a particular instantiation of this description each patient simulated. The form of this description is a 'Lifeline' object that consists of an ordered series of Events. Each Event carries with it a template for outputting a textual Journal Entry and a list of child Events.

As the simulation is run the next Event is selected from the Lifeline of a given patient, a Journal Entry is produced and appended to the medical history of the patient and zero or more of the child events are selected and inserted into the future of the Lifeline for the patient. Two models exist for choosing child Events to be placed into the future of a Lifeline — the selection of a single child Event using a weighted choice from the list of children (for n child Events each child ei is assigned a weight wi and the probability of a given child Event being selected is wi / $\sum$(1 to j)wj) or the independent possible selection of each child Event with a given probability (each child ei is assigned a probability pi < 1 of being inserted into the lifeline).

Event objects are further specialised in sub-classes. 'Repeating Events', which, along with the selection of child Events also inserts a clone of itself into the Lifeline. Repeating Events repeat either regularly or at specified random intervals and can do so either an indefinite or limited number of times. 'Terminating Events' erase all future events from the Lifeline, hence ending the current simulation (the canonical example being a 'Death Event'). Finally 'Scrubbing Events' erase all future Events of a given type from the Lifeline. These are used typically to model 'curative' events which preclude future disease related events from occurring.

1.2. Producing Simulations

The above model describes a general method for describing rich simulations in a relatively simple form. The requisite realism of the produced simulations relies though on the details of the simulation modelled. In sourcing and creating actual simulations based on this model we have taken two approaches outlined below.

One of the drivers behind the design of the above model was the notion of 'Care pathways'. Care pathways are a map of the anticipated care that a patient will receive, based on time frame and the current medical state of the patient [8]. In essence a care pathway describes the ideal actions a medical practitioner would undertake in deciding on the course of care for a patient. The development of the simulation model was informed by this notion of care pathways, and there is a natural representation of care pathways within the simulation model, with each event on the care pathway being represented by an Event within the model, and future actions described in the pathway being modelled in the children of the Event. It only remains here to assign probabilities to events with multiple possible outcomes, where experimenting with different outputs is possible.

The underlying structure of the simulation model, that of an Event triggering future Events based on some probability suggests an amenability to representing some form of Markov model [9] where future outcomes are based solely on the current state of the world. A relatively simple analysis of real patient data, if available, can be used to calculate the probability of an item in a patients medical record appearing given the occurrence of a previous item [10]. These chains of probabilities can be translated into the event-based model of the simulation.

Both the techniques for constructing simulations described above can be used by engineers with little or no medical knowledge to create simulations. This is a benefit in that it overcomes the gap that often exists between technical and domain expert. A medical domain expert could of course though create valid simulation models based purely on domain expertise.

2. Results

The development of this tool was driven by the need for data for the development of a suite of eLab tools, for socially oriented analysis and manipulation of medical data [11]. These tools were designed to expose data from an array of heterogeneous medical data sources covering a range of disease areas [12]. Whilst the development of the eLab tools required data for testing and design purposes the inherent sensitivity of medical data precluded the use of real data. A series of simulations were produced using a mix of the methods described in section 1.2. These included, Type I and II diabetes, hypertension, asthma and general health issues, with the asthma model illustrated in Figure 1.

The aim of this model is to produce patient records similar to the records of patients who suffer from asthma. The core features of this simulation are the presence of an asthma Diagnosis Event that is triggered at some point between 4 and 60 years of age. This is triggers an asthma Management Event which repeats every 12 months. In turn this Event triggers of subsequent events, shown in Figure 1, leading to the construction of the simulated record. These cover the prescription of various drugs,

journal entries relating to an annual review of the patient, possible suits of hospital tests and a small chance of a death event (a 'Scrubbing Event' that ends the simulation). The probabilities of events occuring or following other events were derived from analysis of an anonymised extract of data from a similar underlying population as that which would be accessed by users of the eLab. Using the tool large amounts of simulated data were produced which were used to inform the successful development of the eLab software.

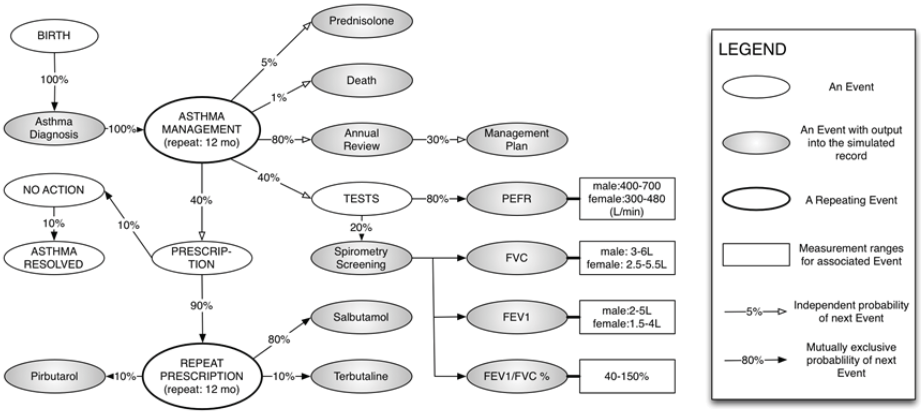

Figure 1. An asthma simulation model.

3. Discussion

Previous approaches to simulating medical data have however tended to be used for either educational tooling [13], where the results of the simulation are used to mimic real life scenarios and are then used as to train medical professionals in an environment that removes the need for direct involvement with either the patient or their health record, or statistical research, where the results of the simulation are used to generate valid scientific hypotheses about the underlying population being simulated [14]. The method and tool presented here do not attempt to simulate patients in a way that would be useful to either medical students or practitioners or in a way that would provide serious scientific output in the area of epidemiological or statistical research, instead addressing a previously overlooked need for data for use in development and testing.

The tool described in the preceding sections has been designed primarily as an aide for use in the development and testing of medical information systems. It deliberately avoids an attempt to either simulate realistic medical processes in depth, or to produce medical records that have pedagogical value. As such it is not possible to formally validate the tool, either from the perspective of measuring the 'realism' of the simulated records or via an analysis of the impact of using such simulated records rather than randomised or anonymised data. However for practical purposes the tool has proven useful and additionally the methodology described, that of representing potential outputs of medical records as a self-modifying series of events represents a novel approach that can be carried forward in other simulation models or adapted in

other areas. We have identified a need for simulation tools that address the problem of producing realistic data for use in the development and testing of medical information systems, and have outlined a methodology and tool that begins to address this need.

References

[1] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, *Big data: The next frontier for innovation, competition, and productivity*, McKinsey Global Institute, 2011.

[2] V. Mayer-Schonberger and K. Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt, Boston, 2013.

[3] House of Commons Health Committee, *The electronic patient record*, *[cited: 15 Feb. 2015]* Available From: http://www.parliament.the-stationery-office.co.uk/pa/cm200607/cmselect/cmhealth/422/422.pdf.

[4] A. Cornwall, Electronic *health record*s: An international perspective, *Health Issues*, **73** (2002)_.

[5] HIMSS Enterprise Systems Steering Committee, *Electronic health records: A global perspective (2010) [cited: 15 Feb. 2015]*, Available From: http://www.himss.org/files/HIMSSorg/content/files/Globalpt1-edited%20final.pdf.

[6] J. G. Anderson, Social, ethical and legal barriers to e-health. *I. J. Medical Informatics*, **76** (2007), 480–483.

[7] Roberto J. Bayardo and Rakesh Agrawal. 2005. Data Privacy through Optimal k-Anonymization. In Proceedings of the 21st International Conference on Data Engineering (ICDE '05). IEEE Computer Society, Washington, DC, USA

[8] Karen Zander. Integrated care pathways: eleven international trends. *Journal of Integrated Care Pathways*, **6** (2002), 101-107.

[9] L. E. Baum and T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains, *Annals of Mathematical Statistics*, **37** (1966), 1554–1563.

[10] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, **57** (1970) 97–109

[11] J. D. Ainsworth and I. E. Buchan, e-labs and work objects: towards digital health economies *I Communications Infrastructure. Systems and Applications*, **16** (2009), 205–216.

[12] J. Ainsworth, J. Cunningham, and I. Buchan, eLab: bringing together people, data and methods to enhance knowledge discovery in healthcare settings, *Stud Health Technol Inform*, **175**, (2012), 39–48.

[13] M. L. Good, Patient simulation for training basic and advanced clinical skills, *Med Educ*, **37** (2003) 14–21.

[14] I. Buchan, J. Ainsworth, E. Carruthers, P. Couch, M. O'Flaherty, D. Smith, R. Williams, and S. Capewell, IMPACT: A generalisable system for simulating public health interventions, *Stud Health Technol Inform*, **160** (2010), 486–490.