EasyChair Preprint

№ 5173

Realistic Synthetic Health Condition Timelines: Generating the Patient History using Contextually Appropriate Disease Burden and Health Statistics

Scott McLachlan, Kudakwashe Dube, Clement Puybareau, Maxime Pitard, Derek Buchanan, Patience Chiketero, Thomas Gallagher, Bridget Daley, Graham Hitman and Norman Fenton

March 17, 2021

Realistic Synthetic Health Condition Timelines: Generating the Patient History using Contextually Appropriate Disease Burden and Health Statistics

Scott McLachlan
Risk and Information Management
EECS
Queen Mary University of London,UK
s.mclachlan@qmul.ac.uk

Kudakwashe Dube
School of Fundamental Sciences
Massey University,
Palmerston North, NZ
k.dube@massey.ac.nz

Clement Puybareau
ESEO
Institute of Science and Technology
Angers, France
clement.puybareau@reseau.eseo.fr

Maxime Pitard
ESEO
Institute of Science and Technology
Angers, France
maxime.pitard@reseau.eseo.fr

Derek Buchanan
School of Fundamental Sciences
Massey University
Palmerston North, NZ
derek.buchanan.nz@gmail.com

Patience Chiketero
Health and Wellbeing
Network Rail
Stratford, UK

Thomas Gallagher
Department of Information technology
Missoula College
University of Montana
Montana, USA
tom.gallagher@mso.umt.edu

Bridget J Daley
Centre for Genomics and Child Health
Blizard Institute
Queen Mary University of London
London, UK
b.j.daley@qmul.ac.uk

Graham A Hitman
Centre for Genomics and Child Health
Blizard Institute
Queen Mary University of London
London, UK
g.a.hitman@qmul.ac.uk

Norman E Fenton
Risk and Information Management
EECS
Queen Mary University of London
London, UK
n.fenton@qmul.ac.uk

***Abstract*—Synthetic patient populations and their electronic healthcare records (EHR) have been recognised to be valuable in many secondary uses including pandemic modelling while avoiding access to real health records, which breaches patient privacy. The problem of generating realistic synthetic EHR has remained an elusive challenge partly due to its knowledge-intensive and computationally expensive nature. Central to this challenge is the problem of generating the realistic health condition timelines (RS-HCT) for synthetic patients spanning from cradle to current age or to grave. This *position paper* is part of ongoing work, addresses this problem and presents an innovative approach to, and an algorithm for, generating the RS-HCT over the lifetimes of synthetic individuals within a given population without using real patient data. Statistics on disease burdens as well as clinical vocabulary, clinical expertise and population demographics across age groups are taken into consideration. This work is significant in that achieving the RS-HCT results in a skeletal realistic synthetic electronic healthcare record (RS-EHR) that would then be developed into a full RS-EHR using inexpensive methods that do not require access to the actual EHR for real patients.**

Keywords— synthetic data generation, synthetic health records, electronic health records

I. Introduction

Adoption of the *electronic healthcare record* (EHR) is now an essential part of patient care used in all healthcare settings. EHR are used by clinicians as a timeline, encompassing the flow of an individual's health and disease state from birth to current age. In 2014 *Realistic Synthetic Electronic Health Records* (RS-EHR) were proposed as a privacy preserving tool for enabling secondary health and health systems research without risk of needlessly exposing personal details of real patients [1-3]. Since then, a large number of works have proposed solutions for the synthetic health record problem, and several large projects have resulted. While the majority of RS-EHR *synthetic data generation* (SDG) solutions provide data relevant to a specific disease or intervention [4-7], many have lacked the chronological history to match a complete EHR. To address this, we propose the *realistic synthetic health condition timeline* (RS-HCT). The RS-HCT we propose would have two primary uses. *First*, as *skeletal* RS-EHR, which is the primary motivation for this paper; and *second*, in support of diagnostic, treatment and prognostic clinical decision-making. While RS-HCT could be mined or learned from real EHR, due to the need to preserve patient privacy any access to the *real* EHR should be limited to primary clinical uses. Hence, there is the need to develop approaches to generate RS-HCT without access to the real EHR. This position paper reports only the early part of ongoing work undertaken to develop privacy-preserving RS-HCT using publicly available aggregated health information for a representative population to which our synthetic person will 'belong'.

The RS-HCT must incorporate all health conditions common to the *health condition burden* for the population to which an individual belongs, and specific to the individual. This paper does not consider how a health condition is treated or managed. Rather, it concerns itself only with the generated

health conditions, their diagnosis, impact and treatment as data-points, and when they intersect with the synthetic individual's timeline. The RS-HCT *timeline* is segmented into *age ranges* common to the way health authorities report population-wide health statistics. The problem of this paper is finding approaches to populate an RS-HCT with health conditions appropriate to these age ranges and with consistency to the progression and lived experience of real patients. The health condition is seen in this paper only through the scope of the probability of it being suffered by a person from a given population within a particular age group. Thus, a person has a likelihood for experiencing a particular health condition, and this changes as the person passes through each age range subject to factors that may have existed at birth, or arisen due to prior disease, lifestyle and social status. The remainder of this paper is organised as follows: After discussion of the fundamental problems that must be overcome in order to generate the RS-HCT we review recent works that have focused on generating the longitudinal synthetic EHR. The Approach and Method section discusses each of the component problems before introducing the GenSeT method, RS-HCT knowledge model and health condition typology. The GenSeT method generation approach and algorithms are then presented, followed by discussion of the strengths and limitations of our proposed approach. We then summarise and conclude the paper.

II. The problem of generating RS-HCT

This paper addresses the problem of providing a practical method for generating the RS-HCT as a skeletal structure for the lifelong RS-EHR without requiring the real EHR. Similar to that for *health condition* generating RS-EHR [2], the method under investigation uses declarative constraints (DC) for ensuring realistic properties in the generated HCT. Where this approach differs is that while methods like Synthea [3] and CoMSER [2] for generating RS-EHR consider DC as static components related only to the health condition being generated, the RS-HCT method provides *dynamic declarative constraints* (DDC) for the prior probabilities for generating each health condition instance. These are based *first* on defining an appropriate overall population to align the RS-HCT to, and *second* on resolving factors to provide the evidence for DDC to underpin RS-HCT generation that include: (i) synthetic patient ethnicity as a factor of parental ethnicities; (ii) inherited genetic, congenital and epigenetic conditions; (iii) demographic factors including their place of birth and the environment where the reside; and (iv) socio-economic factors both during childhood and, as the rest of the RS-HCT is generated, for later adult life. The aim of this paper is to present *Generating Synthetic health condition Timelines*, or GenSeT, which is a newly developed component that extends on the authors' prior work on RS-EHR [2]. To achieve this aim the paper presents: (1) the HCT knowledge model; (2) the method for using publicly available statistics along with *clinical practice guidelines* (CPG), caremaps incorporating clinical decisions [8, 9] and clinical expertise; (3) a typology for health conditions and their dependencies; and (4) application of the GenSeT method to a geographical and clinical area, thus demonstrating and validating the method and critically assessing the RS-HCTs that are generated.

III. Related works

Some works on synthetic health record generation provide complex technical detail regarding the generation method or solution architecture [5, 10, 11], while others focus on the health condition, symptoms, and an evaluation of the resulting synthetic data [6, 12]. A much smaller group present absent any detail for either the generation method or health condition [13]. Generally, two approaches exist for generating what are described in the literature as *fully synthetic* EHR. The *first* generates synthetic EHR from samples of real EHR [14, 15]. The *second* uses surrogates in place of real EHR that include some or all of: aggregated demographic, health incidence, treatment and outcome statistics [2, 3]. Leaving aside that clinical datasets are often found to be littered with missing entries which could affect the accuracy of any aggregated knowledge developed from them, it is argued that a key weakness of the first method is that it still has potential to pose a significant privacy risk to those patients whose real EHR are used during the sampling process [14]. While an inherent strength promoted of the second method is that it completely eschews access to real EHR at any point in the process [1, 2].

A recent focus in research has been approaches for generating the *digital twin* - a computer-based doppelganger for elements of, or entire, cities and nation states [16, 17]. Synthea™ [3], SPEW [18], SynC [19], CoMSER [2] and spatial microsimulation algorithms (SMA) [20] are all recent approaches that are receiving ongoing attention in the literature. Synthea's strengths included that it sought to synthesise the entire Massachusetts population, including both those who were patients and those who were not, and that it sought to create a framework for generation of multiple health conditions; so it had direct applicability to the problem of this paper. However, Synthea's weaknesses included that it generates each medical condition absent of important knowledge regarding the demographics and other health conditions suffered by the patient. This led to synthetic patients being generated with gender inappropriate medical or surgical interventions, and others having amputations for diabetic foot ulcers after they had already lost the leg onto which that foot had been attached [3]. SPEW, SynC and MSA are all approaches that apply weightings developed from census and population data to constrain generation. CoMSER, SPEW and SMA are built with common underlying methods such as Markov and Walker Alias models. Each work presents quite specific to the area being modelled but a key strength is that the underlying approach could be more generally applied. SynC provides a larger range of fields for generation and does demonstrate an ability to generate data other than simple demographics. Weaknesses include that these approaches were developed for generating generic populations with limited demographic fields and would need considerable redevelopment for use in generating a synthetic patient population with a much wider range of demographic, predisposition and socio-economic fields so as to be capable of supporting true RS-HCT and RS-EHR generation.

IV. Approach and Method

This section begins by exploring and resolving each of the component issues identified from the problem. It also introduces several components of the GenSeT method for generating the realistic synthetic HCT: the underlying knowledge model and a typology for health conditions.

A. Age ranges

Many health authorities and researchers provide prevalence and ethnic variance data in five-year age ranges[1] [21-23]. For this work, five-year intervals are used to segment the synthetic patient population during SDG. A higher degree of granularity could be applied, but for the purpose of this work, these arbitrary age intervals achieve a sufficiently realistic outcome. It is recognised that there are specific conditions that are particular to certain subsets within a chosen age interval that less granular data may overlook (e.g. neonatal jaundice is specific only to newborns). However, given the unique and innovative approach that we have developed, we believe such limitations are acceptable for what is a *proof of concept*. The choice of discrete numerical age intervals also avoids the debate regarding applicability of results that could arise from the use of descriptive life stages, for example: newborn, adolescent, young adult, middle age and older [24].

B. Population

The target population to be simulated should be one suitable to the purpose that the resulting RS-HCT and RS-EHR will be applied. That population could be global, national, state/county or even a local clinical catchment area. In the New Zealand (NZ) context shown in Figure 1, the local population could be: (i) an entire city like Auckland; or (ii) a District Health Board (DHB) catchment such as Waitemata DHB. For the United Kingdom (UK) the local population could encompass: (i) a county like Essex; (ii) a local council district like Tower Hamlets; (iii) or a National Health Service (NHS) Trust catchment area such as *Barts and the London NHS Trust*. Each local area context can present with minor differences to the wider area, or with significant differences that may dramatically alter the spread of ethnicity, age or disease prevalence.

Each random colour in Fig. 1 could represent a particular ethnicity, age range or primary medical condition. While general heterogeneity often exists amongst national and large city populations, there can be homogenous clusters within, and significant differences between, local level populations. While the three similarly sized local DHB all exist within the larger Auckland city area of NZ, each has unique population clusters that would significantly alter their aggregate statistics, dramatically changing their RS-HCT and RS-EHR requirements. These differences would also render knowledge developed solely from a national or even Auckland-based dataset less accurate if used in any *precision medicine* solution developed for or applied at the local level.

C. Health conditions

A *health condition* is a disease or injury experienced by the patient. In the RS-HCT we focus on the health condition at the point in time of its *onset*, as well as any ongoing *impact* it may have on the synthetic patient's future health conditions and life expectancy. This can include adjusting the probabilities of resulting or comorbid health conditions the synthetic patient may go on to develop.

D. Disease burden

Disease burden is the accrued impact of living with illness or injury and premature mortality [25, 26]. Disease burden incorporates direct and incurred costs of treatment, medication, ongoing surveillance and lost life expectancy [25, 26]. Different models and measures have been used to calculate and describe disease burden including total *years of life lost* (YLL), *years lived with disability* (YLD), the *disability life-adjusted year* (DALY) that measures the difference between the current situation and an ideal situation [27], and a range of estimations and approaches for calculating the more frequently used *quality life-adjusted year* (QALY) [28]. RS-HCT consider disease burden in a more practical manner that focuses on current and ongoing impacts of the generated disease on the health state and life duration of the synthetic patient.

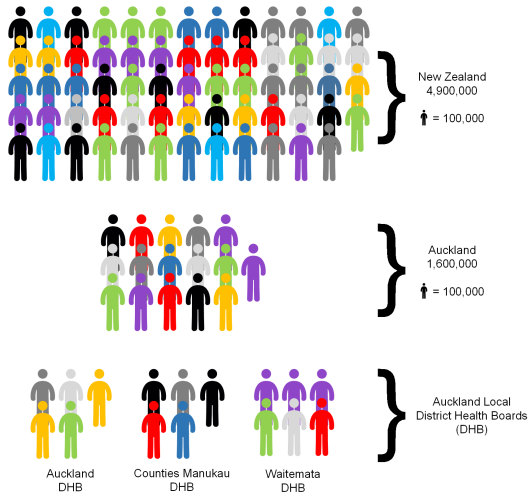

Fig. 1. Populations (New Zealand example, approx.)

E. The HCT knowledge model

This work focuses on the process of generating *realistic health condition onsets* for a patient from cradle to current age, and recognises that to be realistic they must be cognisant of the synthetic patient's: (a) genetic, demographic and socio-economic factors; and (b) their health status up to the current generation stage. To achieve this we must define an appropriate knowledge model from which and into which health conditions are generated, and a model that describes the source materials and timeline into which the health conditions can be chronologically prescribed. The HCT knowledge model is built of many static and dynamic fields as shown in Table 1. Static fields are those that, once populated, do not change throughout life, such as ethnicity and genetic predispositions. A static field can also include injuries and exposures that, once incurred, remain permanent, such as radiation and chemotherapy exposure, removal of a gland or organ, and amputation. Dynamic fields are those that may change during life, whether independently or resulting from interaction with other factors; and include conditions that, once treated or resolved, create no lasting burden for the individual. The overall knowledge model describes the synthetic patient as: (1) an accumulation of static and dynamic *demographic*, *predisposition* and *socioeconomic* factors at birth (generated as part of the gestation phase of the timeline shown in Fig. 2); (2) the value of *updated dynamic factors*

[1] We recognise that other age ranges are possible in publicly released health data. However, age ranges of 5 years were used in the examples cited herein and which were used in supporting our synthetic data generation processes.

TABLE I. TABLE I. FACTORS AND ELEMENTS OF THE HCT KNOWLEDGE MODEL

Factors	Elements	Description/Notes	Example	State
Demographic	Ethnicity	Including parents' ethnicities	Chinese, Canadian	Static
	Place of birth	City, Country and postcode of birth	Toronto, Canada, L3P 0A1	Static
	Place of residence	Current Suburb, City and postcode	Ilford, London, IG1 4RQ	Dynamic
	Birth setting	Whether birth was in hospital, home setting or other	Hospital	Static
	Gender	Gender at birth	Female	Static
	Sex	Sexual presentation	Female	Dynamic
	Sexual orientation	LGBTQ, Hetero	Hetero	Dynamic
Predisposition	Monogenic	Single gene disorders	PKD, cystic fibrosis, sickle cell anaemia	Static
	Congenital & Syndromic	Chromosomal or gene abnormalities with clustering of clinical signs and symptoms	Downs, Cleft Palate, Marfans Syndrome	Static
	Epigenetic	Epigenetic factors that predispose to disease	Fetal alcohol syndrome, epigenetic gene inactivation	Static
	Mitochondrial	Maternal inheritance	Diabetes and deafness, mitochondrial myopathy	S & D
	multifactorial	Genetic and environmental determinants	Diabetes, obesity, heart disease, arthritis	Static
Socio-Economic	Parental Income	Household income during childhood	<£20kpa, £40-60kpa	Dynamic
	Education	Highest level of education	University	Static
	Employment	Current employment status	Professional, Trades	Dynamic
	Income	Current income	<£20kpa, £40-60kpa	Dynamic
	Food	Quantity and quality	Abundant, Nutrient Poor	Dynamic
Disease	Instantial	Diseases that occur in isolation usually without residual effect	Seasonal cold,	Dynamic
	Residual	Diseases that have residual or recurring effect	Allergies, Malaria	Static
	Chronic	Diseases that are lifelong from diagnosis	Diabetes, RA	S & D
	Psychiatric	Psychiatric disorders	Depression, Schizoaffective	S & D
	Cancer	Malignant disease	Glioma, teratoma, melanoma	S & D
Injury	Resolving	Injuries that resolve independently or with treatment	Minor fracture, contusions	Dynamic
	Non-resolving	Injuries that are permanent or disfiguring	Amputation, Maxillofacial	Static
	Psychological	Psychological trauma	PTSD	S or D
Medication	Biologics	Disease modifying (DMARDs)	Long term Methotrexate, Rituximab	Dynamic
	Chemotherapy	Chemotherapeutic and radiation therapies	Bleomycin, Cisplatin, XRT/RTx	Static
	Overdose	Drugs which in overdose cause other or ongoing health effects	Panadol, chemotherapeutics	S & D
	Illicit/DDs	Illicit, addictive or dangerous drugs	Fentanyl, heroin, methamphetamine	S or D
	Hormone therapy	HRT, treatment to correct hormone deficiency	Sustanon, Premarin, Levothyroxine	S & D
	Chronic Treatment	Medications to treat chronic diseases and deficiencies	Diltiazem, Epogen, Warfarin,	S & D
	Vaccination	Allergies and other side effects from vaccines and excipients	ChAdOx1, Rotashield, PEG, Thiomersal	S & D

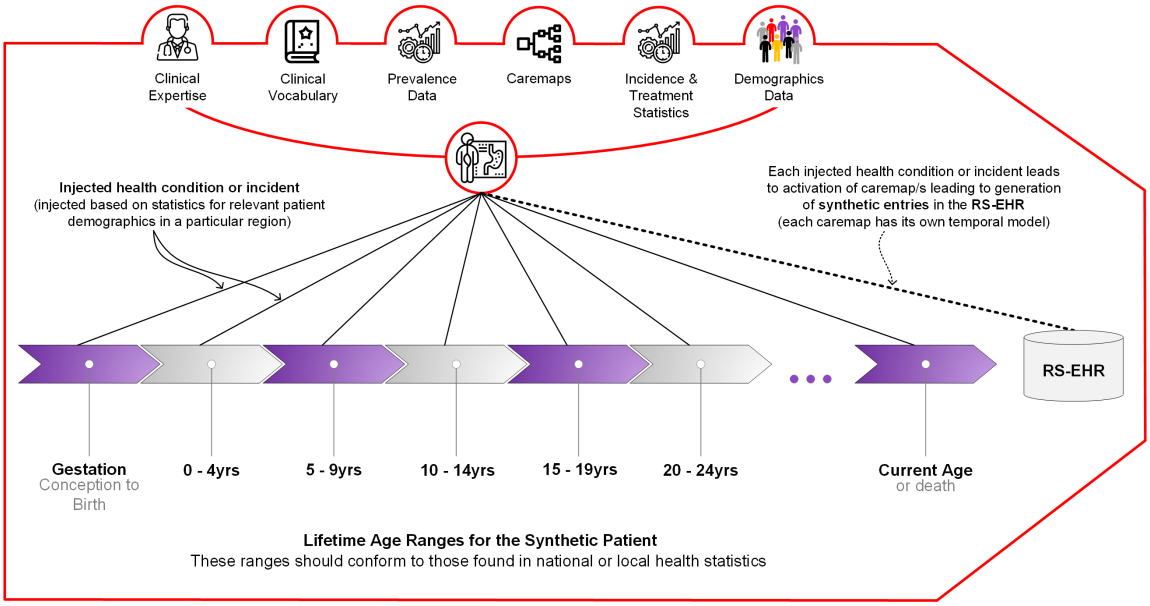

Fig. 2. The GenSeT Method knowledge model for generating the RS-HCT

re-generated at the beginning of each age range phase of the RS-HCT generation process; and (3) the diagnosed *health conditions* which are also generated during each age range phase of the RS-HCT generation process.

F. Generating the health condition

Generating a health condition on the RS-HCT requires that we identify the likelihood, or probability, of that health condition for the current synthetic patient. As shown in Fig. 2, these probabilities can be identified from *prevalence data*, which is the frequency of a given disease in the overall population, and *incidence statistics*, which are aggregate health statistics for a health condition from a national health department or clinical source, often presenting segregated by factors of interest, such as ethnicity and age range at diagnosis. Clinical expertise and clinical practice guidelines (CPG) are also used to identify risk factors for a given health condition, any dependent or co-morbid conditions that are likely to influence or be influenced by the health condition of interest, and when onset is more likely to occur.

This section also presents a typology for generation of health conditions. This typology is presented visually in Figures 3-6, and describes the generic dependencies or influences health conditions may have on each other. The typology describes health conditions that: (i) are independent; (ii) have a one-way dependency ; (iii) have multiple one-way dependences; or (iv) have cycle-bound dependencies.

Risk factor dependent health conditions: Many health conditions have known risk factors that predispose an individual to that health condition. Within a given population there will also be prevalence data published for a broad range of known health conditions. The prevalence represents the *population prior probability* for that health condition and is often expressed either as a rate within the population (e.g.: 1.4/1000) or as a percentage of the overall population (4%). If A is the health condition, the P(A) represents the population prior probability for that health condition.

If R is a risk factor for the health condition that is present for the current synthetic patient, the effect of R is to update P(A) for that health condition. The updated probability is described as the probability for having the health condition given the risk factor, and is written as P(A|R).

If x is a person (real or synthetic) with specific risk factors R1, R2, ...,R_n then we can think of P(A|R1, R2, ..., Rn) as the *personalised health condition* of x; that is to say, it is the probability that this person x has health condition A given the known risk factors for x.

Independent and dependant health conditions: The independent disease, as shown in Figure 3a, is one that could be described as being absent a patient-intrinsic cause or lasting consequence.

Some health conditions are known to be dependent risk factors for developing additional health conditions. These health condition relationships can be observed in several forms. Figure 3b shows a single relationship where health condition A influences the likelihood of health condition B. An example is where the presence of one autoimmune health condition is known to increase the likelihood of another, as in our later example where Type 1 Diabetes Mellitus (T1DM) influences the potential for Systemic Lupus Erythematosus (Lupus).

Fig 3a. Independent health condition

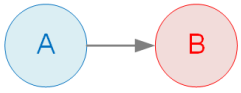

Fig. 3b. One-way health condition dependency

Fig. 4 shows multiple one-way relationships wherein: (i) health condition A influences the likelihood of health conditions B or C; (ii) the presence of health conditions A and C together influence the likelihood of health condition B; (iii) the presence of health condition C influences the likelihood of health condition D; and (iv) the presence of health conditions

B and C together influence the likelihood of health condition D.

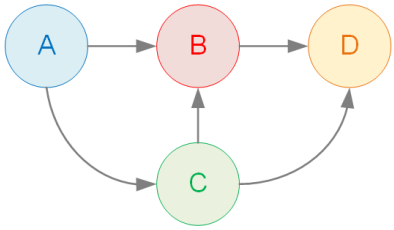

Fig. 4. Multiple one-way dependencies

Fig. 5 shows examples of cycle-bound relationships where: (i) a single health condition, once experienced, is recurrent; or (ii) two or more health conditions act to either amplify each other, or cause recurrent experience of the health condition.

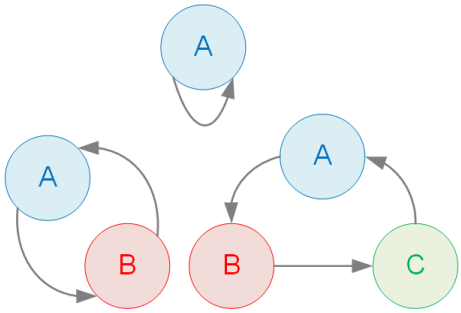

Fig. 5. Cycle-binding dependencies

Where health condition A influences health condition B, the effect is to update P(B) given health condition A. The updated probability is described as the probability for having health condition B given the presence of health condition A, and is written as P(B|A).

If person *x* (real or synthetic) has health condition A which influences health condition B, as well as specific risk factors R1, R2 and R3 that also influence health condition B, then we can think of P(B|A, R1, R2, R3) as the extended personalised health condition of *x*.

Figures 6 and 7 apply the health condition typology to two common health conditions: diabetes and malaria. Fig. 6 demonstrates that diabetes has a direct influence that predisposes the patient to nephropathy, cardio-vascular disease and retinopathy.

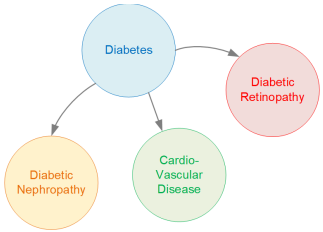

Fig. 6. Example of multiple one-way dependencies in diabetes

Even after treatment and a return to a seemingly healthy state, patients with malaria can experience relapse, or recrudescence. In this way, as shown in Fig. 7, malaria can be an example of a recurrent or cycle-bound health condition.

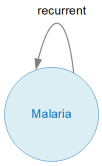

Fig. 7. Example of cycle-binding in recurrent malaria

V. Generating the Realistic Synthetic Health Condition Timeline

To simplify the process, the continuous health timeline is generated chronologically in discrete five-year-long age groups. The background prior for each disease that may be relevant to the current age group is identified and, where risk or other factors apply, updated. This process is demonstrated in Fig. 8. Demographic and predisposition factors have been generated during the Gestation period, and are used during the 0-4yrs period to update background priors for all diseases known to the system. In the example shown the synthetic patient has risk factors for Type 1 Diabetes Mellitus (T1DM)

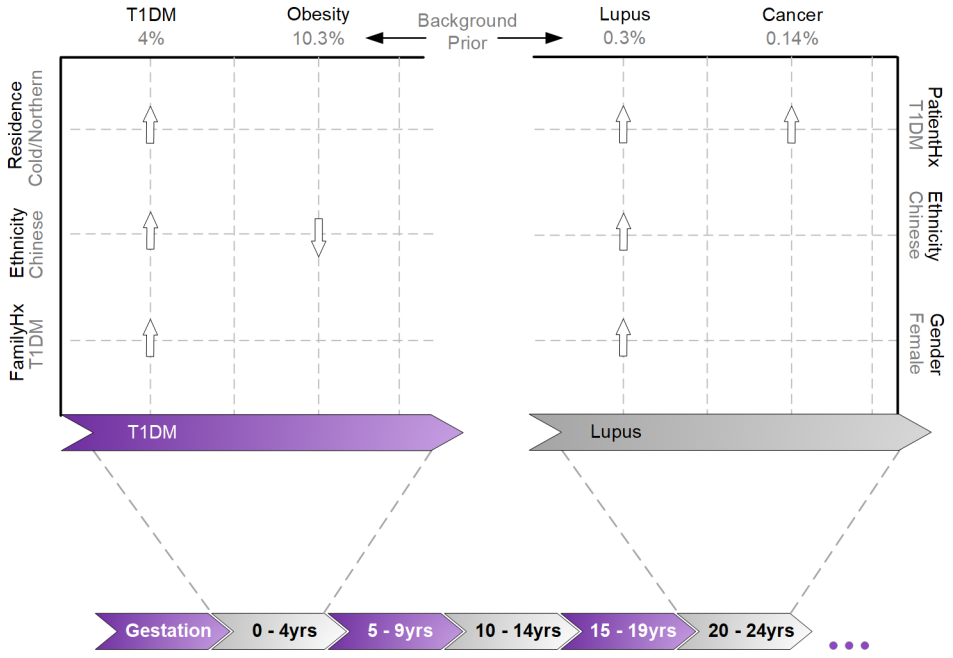

Fig. 8. How GenSeT generates health condition timelines

that include: (i) a family history (FamilyHx) of T1DM [29, 30]; (ii) being of an east Asian ethnicity [31, 32]; and (iii) living in a colder northern climate [33]. These risk factors individually and cumulatively increase the 4% background prior for T1DM, as shown by the upward pointing arrows. The system, based on these risk factors, has chosen to generate a diagnosis for T1DM. Contemporaneously the synthetic patient's ethnicity has acted to decrease the likelihood for obesity, as shown by the downward pointing arrow.

In the later 20-24yrs age range the accumulation of our synthetic patient's ethnicity [34], gender [35] and prior non-obese autoimmune T1DM diagnosis [36, 37] have all increased her likelihood for Lupus; which has been diagnosed. While T1DM increases her risk for certain cancers [38], a cancer diagnosis has not been generated during this age range cycle.

A. Generation algorithm

The GenSeT method consists of two phases. The first phase generates *synthetic patients* and populates their record with values for the *demographic*, *predisposition* and *socio-economic* factors described in Table 1. The second phase generates health conditions along the synthetic health condition timeline.

Generating synthetic patients: There are many ways to generate the synthetic patient. We have previously explained a process using a stepwise Walker's Alias algorithm for constrained realistic demographics generation in [2]. The GenSeT synthetic patient generator described in Listing 1 draws on national and local demographics data for such things as ethnicity and gender. It also draws on a combination of demographics data, prevalence data and incidence statistics to generate predisposition and socio-economic factors. Aside from preparing the locale-specific data and statistics, the other user-controlled feature is selecting the size of the synthetic population; or number of synthetic patients to generate. The accumulated fields generated during this phase populate the complete Gestation stage of the RS-HCT. The GenSeT synthetic patient generator delivers a database of synthetic patients ready to receive simulated health conditions on a health timeline generated by the GenSeT RS-HCT generator.

Generating synthetic health timelines: The GenSeT RS-HCT generator described in Listing 2 extends the synthetic patient through each age range using the accumulated patient data generated in the Gestation stage, along with prevalence data, incidence and treatment statistics. The clinical vocabulary and clinical expertise are used to temper the statistical data with knowledge of which conditions are independent and dependent, and the strength of relationship between dependent conditions.

Since the number of age ranges and health conditions may be considered to be constants, it follows that the computational complexity of this algorithm is of order O(n), where n is the number of synthetic patients required to be generated together with their RS-HCT.

For each health condition, the prior probability is identified from national or local prevalence data and incidence statistics. Drawing on clinical expert knowledge, it is then updated to account for the influence of specific risk factors and any dependent health conditions the synthetic patient has. The final step is for the system to use the updated probability in a decision process that will render a Boolean decision for whether the patient is or is not diagnosed with the health condition. There are a number of SDG decision processes that may be used such as Walker's Alias, Markov Chains, Generative adversarial networks and probabilistic Bayesian networks.

VI. Strengths, Limitations and Future Work

It is important to emphasise here that this is a position paper with the limitation that it reports work in progress that is currently incomplete but worthy of reporting due to the current attention [39-41] and funding [42, 43, 41] being made available for research into development of realistic *digital twin* solutions for environments, civic systems and populations [44].

Many published synthetic EHR solutions draw use health condition prevalence as a fixed rate, effectively salting their synthetic dataset of patients with that diagnosis at the given frequency. A key strength for GenSeT is use of the prevalence value as a prior probability that is dynamically updated based on risk factors and dependent health conditions that are already known for this patient. This process means GenSeT

Listing 1: *GenSeT Synthetic Patient Generation Algorithm: RS-HCT Generation Algorithm for the GenSeT Method*

```
Inputs:
    1.  prevData - Prevalence data;
    2.  stats - Incidence and treatment statistics;
    3.  popDemo - Population Demographics;
    4.  patientPop - Numeric value for the required patient population size

Output: synPatientDB - Synthetic patients database

Pre-condition: synPatientDB is initially empty;

Post-condition: synPatientDB is populated and has size patientPop
```

```
 1  GenSeT.genSynthPatients()
 2  Begin
 3     repeat:
 4      a.  patientDemographics ← GenerateDemographicFactors(popDemo);
 5      b.  predispositions ← GeneratePredispositionFactors(popDemo, prevData, stats);
 6      c.  socioEcoFactors ← GenerateCurrentSocioEcoFactors(popDemo);
 7      d.  synthPatient ← GenerateSynthPatient(patientDemographics, predispositions,
 8          socioEcoFactors);
 9  addPatient(synthPatient, synPatientDB);
10     Until synthPatientDB.size() == patientPop;
11  End;
```

Listing 2: *GenSeT Synthetic Health Condition Timeline Algorithm: RS-HCT Generation Algorithm for the GenSeT Method*

```
Inputs:
    1. clinVoc - Clinical vocabulary;
    2. prevData - Prevalence data;
    3. stats - Incidence and treatment statistics;
    4. synPatientDB - Synthetic Patients with Demographics;
    5. clinExp - Clinical expertise
    6. stdClinicalAgeRanges - Age ranges common to stats;

Output:
    1. synPatientDB.rsHCT - Synthetic patients database with Health Condition Timelines

Pre-condition:
    1. synPatientDB.rsHCT is initially empty for each patient,
    2. stdAgeRanges contains age ranges sorted by temporal order;

Post-condition: synPatientDB.rsHCT is populated for each patient
```

```
 1  GenSeT.rsHCT()
 2  # generates synthetic HCT for each range in order
 3
 4  Begin
 5  For each ageRange in stdClinicalAgeRanges do
 6      For each synthPatient in synPatientDB Do
 7          For each healthCondition in PrevData Do
 8              i.   priorProb ← GetPriorProbability(prevData)
 9              ii.  priorProb ← riskFactorUpdate(priorProb, clinVoc, clinExp, prevData, stats,
10                   synthPatient)
11              iii. priorProb ← dependentHealthConditionsUpdate(priorProb, clinVoc, clinExp,
12                   prevData, stats, synthPatient)
13              iv.  hasDiagnosis ← isDiagnosed(priorProb)
14              v.   If (hasDiagnosis)then
15                     a.  date ← determineDateOfDiagnosis(ageRange)
16                     b.  Update(synPatientDB.rsHCT, healthCondition, date)
17          Repeat (nextHealthCondition)
18      Repeat (nextPatient)
19  Repeat (nextAgeRange)
20  End;
```

ensures conditions are generated for synthetic patients that are more likely to experience them. Another strength is use of phased generation using age ranges as this enables ongoing updating of the prior probability based on the ongoing health experience of the synthetic patient. This means dependent health conditions are more likely to be generated only for those patients with the related primary health condition, for example: that *diabetic foot ulcers* are more likely to be generated for: (a) diabetic patients; who (b) still have the limb on which is the ulcerated foot is attached.

Limitations exist that require further work. It is necessary to develop an approach for computing the boundary for how many conditions any one patient may be capable of bearing and therefore, at what point it is most appropriate to impute death. An approach is also needed for evaluating the synthetic HCT to validate whether the accumulated conditions experienced by each patient remain realistic. This approach must be capable of evaluating clinical knowledge and identifying three types of situations that make the new condition: (1) more likely; (2) extremely unlikely; and (3) impossible. For example, identifying that: (1) someone with an autoimmune disease is more likely to experience additional autoimmune diseases or renal and hepatic failure; (2) after bi-lateral tubal ligation a woman is unlikely to be found pregnant (but it has happened and therefore is not impossible); and (3) a woman who has undergone a full hysterectomy could not be diagnosed with ovarian cancer or poly-cystic ovarian syndrome.

VII. Summary and Conclusions

The realistic synthetic health condition timeline (RS-HCT) forms a strong basis for generating a more comprehensive and realistic synthetic electronic healthcare record (RS-EHR) for a synthetic patient population. Some works in the literature have recognised this by the inclusion of a step for mining the HCT from the real EHR within their algorithm for generating the synthetic EHR. Other works place no emphasis on generating the RS-HCT despite that it forms the skeletal form of the desired RS-EHR. This position paper has presented an approach and method in our early efforts in developing a framework and software tool for generating the RS-HCT for the patient segment of a population *without access to the EHR of real patients*, which potentially could breach patient privacy. Through incorporation of publicly available datasets and clinical expert knowledge and applied to common age ranges, this paper has presented a novel strategy for generating the RS-HCT. In the method presented here, synthetic patients with a broad range of demographic and predisposition data are also generated as a precondition for generating the RS-HCT for those patients. The uniqueness of the approach and method presented here lies in: (i) segmenting the HCT based on age groups; and (ii) running patients through the common age group-based temporal segments; while (iii) applying health condition prevalence statistics and health expertise; together with (iv) health condition dependency considerations in the form of dynamic forward adjustment of prior probabilities as patients move along the timeline. We contend the RS-HCT will have as much of an impact on developing health-related *digital twin*

solutions as the RS-EHR has been seen to have during the last seven years.

Contribution Statement

SM prepared the first draft with assistance from KD. NF reviewed the probabilistic, typology and constraint approach. GAH, BJD, DB and PC provided clinical content and review. CP and MP proposed the health condition typology and prepared beta code for *proof of concept* testing.

Acknowledgment

SM, BJD, GAH, and NF acknowledge support from the EPSRC under project EP/P009964/1: PAMBAYESIAN: Patient Managed decision-support using Bayes Networks. MP and CP acknowledge the support of Massey University and ESEO for providing the opportunity to attend Massey and undertake this work with the HiKER Group.

Competing Interests

No author identified a competing interest relevant to this research.

References

[1] Dube, K. and T. Gallagher. (2014). Approach and method for generating Realistic Synthetic Electronic Healthcare Records for secondary use. Paper presented at the International Symposium on Foundations of Health Informatics Engineering and Systems, Berlin, Heidelberg.

[2] McLachlan, S., K. Dube and T. Gallagher. (2016). Using CareMaps and health statistics for generating the realistic synthetic Electronic Healthcare Record. Paper presented at the International Conference on Healthcare Informatics (ICHI'16), Chicago, USA.

[3] Walonoski, J., M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, and S. McLachlan. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. Journal of the American Medical Informatics Association, 25(3), 230-238.

[4] Avino, L., M. Ruffini and R. Gavalda. (2018). Generating synthetic but plausible healthcare record datasets. ArXiv Preprint, arXiv:1807.01514.

[5] Buczak, A., S. Babin and L. Moniz. (2010). Data-driven approach for creating synthetic electronic medical records. BMC medical informatics and decision making, 10(1), 1-28.

[6] Dash, S., R. Dutta, I. Guyon, A. Pavao, A. Yale and K. Bennett. (2019). Synthetic event time series health data generation, arXiv:1911.06411.

[7] Piacentino, E., A. Guarner and C. Angulo. (2021). Generating Synthetic ECGs using GANs for Anonymising Healthcare Data. Electronics, 10(389).

[8] McLachlan, S., E. Kyrimi, B. Daley, K. Dube, M. Marsden, S. Finer, G. Hitman and N. Fenton. (2020). Incorporating Clinical Decisions into Standardised Caremaps. Paper presented at the IEEE International Conference on Health Informatics (ICHI), DOI: 10.1109/ICHI48887.2020.9374381.

[9] McLachlan, S., E. Kyrimi, K. Dube and N. Fenton. (2019). Clinical caremap development: How can caremaps standardise care when they are not standardised? Paper presented at the 12th International Joint Conference on Biomedical Systems and Technologies (BIOSTEC 2019), volume 5: HEALTHINF, Prague, Czech Republic.

[10] Gaba, S., Y. Havinga, T. van der Weide, J. Visser, E. Hoijtink, H. Brons, J. Kijne, P. Dijksta, and F. Walraven. (2020). Portavita Benchmark: A Dataset Generator for Healthcare. Retrieved from: online. https://portavita.com/sites/default/files/whitepapers/AXLE_Healthcare_Dataset_Generator.pdf

[11] Yale, A., S. Dash, R. Dutta, I. Guyon, A. Pavao and K. Bennett. (2020). Generation and evaluation of privacy preserving synthetic health data. Neurocomputing, 416, 244-255.

[12] Walonoski, J., S. Klaus, E. Granger, D. Hall, A. Gregorowicz, G. Neyarapally, A. Watson and J. Eastman. (2020). Synthea™ Novel coronavirus (COVID-19) model and synthetic data set. Intelligence-based medicine, 1(100007).

[13] Liu, Y. and Y. Theng. (2020). Development of Synthetic Health Records to Support Urban Planning for Healthy Aging. Innovation in Aging, 4, 12.

[14] El Emam, K., L. Mosquera and J. Bass. (2020). Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation. Journal of Medical Internet Research, 22(11), e23139.

[15] Elliot, M. (2014). Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team. Retrieved from: https://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/2015-02%20-Report%20on%20disclosure%20risk%20analysis%20of%20synthpop%20synthetic%20versions%20of%20LCF_%20final.pdf

[16] El Maria, Q., T. Taleb and J. Song. (2020). Roads Infrastructure Digital Twin: A Step Toward Smarter Cities Realization. IEEE Access ACCESS.2017.2657006.

[17] Srinivasan, R., B. Manohar and R. Issa. (2020). Real-Time Demand Response Using Digital Twin. In (Ed.), Cyber-Physical Systems in the Built Environment. Cham.: Springer.

[18] Gallagher, S., L. Richardson, S. Ventura and W. Eddy. (2017). SPEW: Synthetic Populations and Ecosystems of the World. Carnegie Mellon University. ArXiV Preprint. Retrieved from https://arxiv.org/pdf/1701.02383.pdf

[19] Wan, C., Z. Li, A. Guo and Y. Zhao. (2019). SynC: A Unified Framework for Generating Synthetic Populationwith Gaussian Copula. University of Toronto. ArXiv PrePrint. Retrieved from https://arxiv.org/pdf/1904.07998.pdf

[20] Harland, K., A. Heppenstall, D. Smith and M. Birkin. (2012). Creating Realistic Synthetic Populations at Varying Spatial Scales: AComparative Critique of Population Synthesis Techniques. Journal of Artificial Societies and Social Simulation, 15(1).

[21] DiabetesUK. (2010). Diabetes in the UK: Key Statistics. Retrieved from: https://www.diabetes.org.uk/resources-s3/2017-11/diabetes_in_the_uk_2010.pdf

[22] NZMoH. (2019). Report on Maternity Web Tool. Retrieved from: https://minhealthnz.shinyapps.io/Maternity_report_webtool/

[23] UKONS. (2019). Suicides in the UK: 2018 Registrations. Retrieved from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/suicidesintheunitedkingdom/2018registrations

[24] Geifman, N., R. Cohen and E. Rubin. (2013). Redefining meaningful age groups in the context of disease. Age, 35(6), 2357-2366. doi:10.1007/s11357-013-9510-6

[25] Ekwueme, D., P. Strebel, S. Hadler, M. Meltzer, J. Allen and J. Livengood. (2000). Economic evaluation of use of diphtheria, tetanus, and acellular pertussis vaccine or diphtheria, tetanus, and whole-cell pertussis vaccine in the United States, 1997. Archives of pediatrics & adolescent medicine, 154(8), 797-803.

[26] Fadiga, M., C. Jost and J. Ihedioha. (2013). Financial costs of disease burden, morbidity and mortality from priority livestock diseases in Nigeria. Retrieved from: Nigeria. https://cgspace.cgiar.org/rest/rest/bitstreams/44bed655-da59-4ce4-a547-931c3ba89aae/retrieve

[27] WHO. (undated). The Global Burden of Disease concept. Retrieved from: https://www.who.int/quantifying_ehimpacts/publications/en/9241546204chap3.pdf

[28] Smith, M., M. Drummond and D. Brixner. (2009). Moving the QALY forward: rationale for change. Value in Health, 12.

[29] Bonifacio, E., M. Hummel, M. Walter, S. Schmid and A. Zeigler. (2004). IDDM1 and multiple family history of type 1 diabetes combine to identify neonates at high risk for type 1 diabetes. Diabetes Care, 27(11), 2695-2700.

[30] Peng, H. and W. Hagopian. (2006). Environmental factors in the development of Type 1 diabetes. Reviews in endocrine and metabolic disorders, 7(3), 149-162.

[31] Mayer-Davis, E., R. Bell, D. Dabelea, R. D'Agostino, G. Imperatore, J. Lawrence, L. Liu and S. Marcovina. (2009). The many faces of diabetes in American youth: type 1 and type 2 diabetes in five race and ethnic populations: the SEARCH for Diabetes in Youth Study. Diabetes Care, 32, S99-S101.

[32] Spanakis, E. and S. Golden. (2013). Race/Ethnic Difference in Diabetes and Diabetic Complications. Current Diabetes Reports, 13(6).

[33] Levy-Marchal, C., C. Patterson and A. Green. (1995). Variation by age group and seasonality at diagnosis of childhood IDDM in Europe. Diabetologia, 38, 823-830.

[34] Lau, C., G. Yin and M. Mok. (2006). Ethnic and geographical differences in systemic lupus erythematosus: an overview. Lupus, 15(11), 715-719.

[35] Bruce, I., M. Urowitz, D. Gladman, D. Ibanez and G. Steiner. (2003). Risk factors for coronary heart disease in women with systemic lupus erythematosus: the Toronto Risk Factor Study. Arthritis and Rheumatism, 48(11), 3159-3167.

[36] Al Ahmed, O., V. Sivaraman, M. Moore-Clingenpeel, A. S., S. Bout-Tabaku and CARRA. (2020). Autoimmune thyroid diseases, autoimmune hepatitis, celiac disease and type 1 diabetes mellitus in pediatric systemic lupus erythematosus: Results from the CARRA Legacy Registry. Lupus, 29(14), 1926-1936.

[37] Esteban, L., T. Tsoutsman, M. Jordan, D. Roach, L. Poulton, A. Brooks, O. Naidenko, S. Sidobre, D. Godfrey, and A. Baxter. (2003). Genetic control of NKT cell numbers maps to major diabetes and lupus loci. The Journal of Immunology, 171(6), 2873-2878.

[38] Shu, X., J. Ji, X. Li, K. Sundquist and K. Hemminki. (2010). Cancer risk among patients hospitalized for Type 1 diabetes mellitus: a population ‐ based cohort study in Sweden. Diabetic Medicine, 27(7), 791-797.

[39] Dembski, F., U. Wossner, M. Letzgus, M. Ruddat and C. Yamu. (2020). Urban Digital Twins for Smart Cities and Citizens: The Case Study of Herrenberg, Germany. Sustainability, 12(2307).

[40] FrontierSI. (2021). Call for Digital Twin Proposals. Retrieved from https://frontiersi.com.au/digital-twin/

[41] SCJ. Moscow "Digital Twin" project received the international ISOCARP award. Smart City Journal. Retrieved from https://www.thesmartcityjournal.com/en/cities/moscow-digital-twin-project-received-the-international-isocarp-award

[42] Ketzler, B. (2020). Digital Twin cities receives prestigious Epic Mega Grant. Retrieved from https://dtcc.chalmers.se/2020/06/30/digital-twin-cities-receives-prestigious-epic-mega-grant/

[43] RAEng. (2021). UK IC Postdoctoral Research Fellowships. Retrieved from https://www.raeng.org.uk/grants-prizes/grants/support-for-research/ic-postdoctoral

[44] CSIRO. (2021). Digital twins at CSIRO's Data61: From objects and systems to precincts and cities. Retrieved from https://data61.csiro.au/en/Our-Research/Our-Work/Future-Cities/NSW-Digital-Twin/NSW-Digital-Twin