Temporal Data Simulation based on a real data set for fall prevention

Gulshan Sihag[1], **Véronique Delcroix**[1],
Emmanuelle Grislin-Le Strugeon[1,2], **Xavier Siebert**[3], **Sylvain Piechowiak**[1]
[1]Univ. Polytechnique Hauts-de-France, LAMIH, CNRS, UMR 8201, F-59313 Valenciennes, France
[2] INSA Hauts-de-France, F-59313 Valenciennes, France
[3] Univ. de Mons, Faculté polytechnique Département de Mathématique et recherche opérationnelle, Belgium
Gulshan.Sihag@uphf.fr, Veronique.Delcroix@uphf.fr, Emmanuelle.Grislin@uphf.fr, Xavier.Siebert@umons.ac.be,
Sylvain.Piechowiak@uphf.fr

Abstract

The motivation of this article is the need for a large temporal data set, which allows to reason about the changes regarding person's features over a long period of time. Faced with the difficulty of finding such a data set, we propose an algorithm to simulate such a data set, based on real static data provided by the service of fall prevention of Lille's hospital. We select five persistent variables, meaning that their value may change at most once, toward positive value for positive persistent variables. The algorithm is based on assumptions regarding the temporal evolution of each contextualized variable, as defined by a Bayesian network learned on the real static data set. The temporal data set simulated thanks to the proposed algorithm is evaluated by the comparison of the temporal distribution of each contextualized variable with the functions obtained by linear interpolation from the real data set.

Introduction

Data are the basis of a lot of works in artificial intelligence, often related with learning and reasoning. A temporal data set includes series of values over time for a set of variables, and a set of samples. With regard to data on people, describing for example their abilities, environment, behavior, etc, longitudinal studies allow to collect repeated observations over time of a phenomenon and/or a sample of individuals. However, such data collection is very costly and it often concerns either a short period of time, and / or a small number of persons, and / or a limited number of observations for a given variable.

In order to palliate that difficulty, we aim to simulate a complete temporal data set that includes the values of all variables at each time step for a long period (several decades) and for a large number of elderly.

When dealing with static data, it is frequent to simulate data from a Bayesian network (Gootjes-Dreesbach et al. 2020; de Vries et al. 2021). Temporal data set can also be simulated with dynamic Bayesian networks (DBN) (Murphy 2002). An example is given in (Marini et al. 2015) where a cohort is simulated for fifteen years, thanks to a DBN learned from a longitudinal study. In a DBN, variables are related to each other over two or more time slices. From a theoretical point of view, in order to consider sequences of arbitrary length, a solution is to consider that the probability distributions describing the temporal dependencies are time invariant. In that way, the relations defined between two time slices can be easily deployed (unrolled) for a particular number of steps. However, when we create a junction tree from an unrolled DBN, the cliques tend to be very large, often making exact inference tends to become intractable (Ducamp et al. 2020; Murphy 2002).

In that article, we propose another approach to simulate a temporal data set that allows us no to make the strong assumption that temporal evolution is time invariant. We present an algorithm to simulate a temporal data set on the basis of a real static data set. including information collected during the multidisciplinary consultation for fall prevention in Lille's Hospital. The real data set includes only one observation per person and per variable. We first present the context and the motivation to simulate a temporal data set, followed by the real static data set from Lille's hospital. Second, we explain the algorithm and related assumptions and definitions. We also provide some elements to evaluate the quality of the simulated data set. Finally, we give some perspective about how this data set could be used in the context of the prediction of risk factors for fall from a partial time stamped data set.

Context and motivation

This work takes place in the context of fall prevention. We present below our collaboration with Lille's hospital and our motivation to simulate a temporal data set from a real static data set provided by that service.

One third of people aged 65 and over living at home fall every year. This is the case for half of those over 85 years of age (Dargent-Molina and Bréart 1995). Falls account for 40% of all injury deaths (Rubenstein 2006). According to the World Health Organization (World Health Organization 2008), falls and consequent injuries are major public health problems that require frequent medical attention. Falls prevention is a challenge to population ageing, but it is one of the issues that have not been given a sufficient attention. Since falls result from a complex interaction of risk factors, an important step in fall prevention is to detect the presence of risk factors for falls.

At the hospital in Lille, patients are received in a day hospital for a multidisciplinary evaluation of the risk factors for falls. This leads to the selection of a small number of adapted recommendations. Most part of the time in this specialized consultation consists of data collection by different specialists, using specific equipments and tests. It provides a picture of the person's current state, behavior and environment, incorporating past events that can help to assess risk factors for falls.

However, outside the context of a specialized consultation on fall, such a complete data collection is not possible because of a lack of time, expertise and equipment. Though, there are many potential actors in the prevention of falls, and furthermore, it is possible to have almost instantaneously a partial set of information on a person from his or her personal medical record (electronic health record). These records regroup a collection of reports and information over the patient's life and are increasingly being used. It is therefore possible to extract quickly dated information about a person. But for some variables, the person's current condition may have changed making the information useless, even misleading.

In real life, when information is required immediately for a given person, such as for fall prevention by general practitioners, the available information can be seen as a partial time stamped observation set. Beyond this article, our overall objective is to allow an assessment of the risk factors for falls, based on a partial set of dated information (Delcroix, Grislin-Le Strugeon, and Puisieux 2021). For this purpose, we need knowledge about the dynamics of the considered variables, as well as a sufficient temporal data set in number of patients, covering a long period of time with a fine time step. For these reasons, we aim to simulate a realistic temporal data set about features of interest in fall prevention.

The simulated data will make possible to simulate a partial dated observation set, and to build and evaluate some models or algorithms to predict fall risk factors from a partial dated observation set. The prediction of the unobserved risk factors for fall contributes to fall prevention since adequate actions may reduce those risks and in turn reduce the risk of fall. In this article, the variables about loss of autonomy (*ADLinf5*) and dementia (*demence*) are two target risk factors for fall.

Below, we present a brief overview of some existing methods to predict fall risk, then, we present the static data set provided by Lille's hospital that we use to simulate temporal data.

Overview on some methods to predict fall risk

Several recent articles focus on the prediction of fall or fall risk based on models learned from large data sets (Marschollek et al. 2012; Marier et al. 2016; Homer et al. 2017; Ye et al. 2020) which comes from the population for which the data collection is facilitated (inpatients (Marschollek et al. 2012), nursing home residents (Marier et al. 2016), or people with a specific program of given health insurance, ensuring complete claims coverage (Homer et al. 2017). Despite this favorable data source, in (Marschollek et al. 2012), the authors mentioned the limitation to generalize their findings due to the significant amount of missing data for some sub-items. Our work is also motivated from this aspect with the final objective to propose a way to help in fall prevention for the whole elderly population, on the basis of their available information, even if it is very partial. Furthermore, existing Electronic Medical Record (EMR) systems do not provide an easy mechanism to synthesis and summarize information on *c*hanging risk variables collected in various portions of the EMR to support clinical decision making (Marier et al. 2016). This point brings us to our second consideration, which is determining how old data can be used to assess present risk. Finally, all of those articles are concerned with assessing fall risk, whereas we focus on the evaluation of risk factors for falls.

	Short name	**Description**	**Persistent variable**
G	GUGOgt20	true when the result of the Get Up and GO test is greater than 20 seconds	positive
C	conduit	true when the person still drives her car	negative
A	ADLinf5	true when the score of Activities of Daily Living (ADL) is less than 5	positive
D	demence	true when a dementia is probable or confirmed	positive
M	maisRet	true when the person lives in a retirement home	positive

Table 1: The persistent variables selected for that study

Lille's data set and variable selection

The real data set from the multidisciplinary consultation for fall prevention of Lille's Hospital includes personal data about 1810 persons, collected between 2005 and 2016. In that study, we keep only the 1752 cases with age between 65 and 95.

The original file contains more than 400 columns, among which we have first selected 65 variables for a previous study about the prediction of the main risk factors for fall (Sihag et al. 2020). We now present the five variables selected for this paper.

Variables selection from the real data set

We had several interviews with Pr. Puisieux, from the multidisciplinary consultation on fall prevention at Lille's hospital about the way the 65 variables previously selected evolve with time. As a result, we have identified a subset of variables whose temporal behavior is simple and that we name (positive) *persistent* variables. They are binary variables, whose value is most often false for young people and that can change at most once during the life of a person. Table 1 presents the 5 persistent variables that we select for the purpose of the current study.

The variables *demence* and *ADLinf5* are important predisposing risk factors for fall (Delcroix et al. 2019). The variable *ADLinf5* is an indicator of loss of autonomy. ADL measurements and scales can vary significantly (Mlinac and Feng 2016). The Katz Index of independence in ADLs (Katz et al. 1963) is one of the most commonly used tools to asses basic ADLs (bathing, dressing, toileting transferring, continence, and feeding). Both *demence* and *ADLinf5* are important to be predicted because information related with these risk factors can be difficult to collect, and because they are modifiable, meaning that specific actions can be conducted to reduce them.

The get up and go test (*GUGOgt20*) is related with gait disorder. When its score is greater than 20 seconds, it is considered as a risk factor for fall (Delcroix et al. 2019).

The four variables *GUGOgt20*, *maisRet*, *ADLinf5* and *demence* are positive persistent: when they become true for a given person, there is no chance that they become false again later. The variable *conduit* is negative persistent since it is generally true for adults and becomes false when the capacities of the elderly decrease, while people who did not drive as adults will not drive as elderly.

Definitions and assumptions

Before presenting our algorithm to simulate a temporal data set which represents information over time on a set of persons, we first introduce some notations, definitions, and the assumptions used for the temporal data simulation.

Notations

Here is a list of some notations:

- $\mathbf{X}$: main set of variables,
- $X, X_i \in \mathbf{X}$: some random variables,
- $Dom(X)$: domain of the variable X,
- $Dom(\mathbf{Y}) = Dom(Y_1) \times \ldots \times Dom(Y_m)$, where $\mathbf{Y} = \{Y_1, \ldots, Y_m\} \subset \mathbf{X}$
- $x \in Dom(X)$, $x_i \in Dom(X_i)$: a value of X or X_i [1],
- $\mathcal{T} = \{t_0, t_1, \ldots, t_p\}$ with $t_{i+1} = t_i + \Delta t$: period of time over which information is simulated and Δt is the length of a step,
- t, $t_k \in \mathcal{T}$: different times,
- $x^t \in Dom(X)$, $x_i^t \in Dom(X_i)$: values of the variables X and X_i at the time t for a given person.
- N: size of the population (number of samples),
- n: index of a specific person,
- $\mathcal{D}_{\mathcal{T}}$: complete temporal data set over the period $\mathcal{T}$,

We now present the definitions and related assumptions regarding the variables and their temporal evolution in a context defined by the Bayesian network.

Variables, observations and temporal data set

In that study, we consider only binary variables. For any variable X, $Dom(X) = \{0, 1\}$, where 1 is called the positive value. We also consider only hard observation (see (Mrad et al. 2015) about uncertain observations and (Delcroix, Grislin-Le Strugeon, and Puisieux 2021) about their use in fall prevention). Let's precise definitions and notations regarding dated information.

Definition 1 ***Time stamped observation***
A time stamped observation o on a variable X for a given person n is a tuple $o = (X, x, t, n)$ where $x \in Dom(X)$ is the value of X observed at t.

An time stamped observation $o = (X, x, t, n)$ of a binary variable X is said to be *positive* when the observed value is positive ($x = 1$).

Definition 2 ***Complete temporal data set***
A time stamped data set $\mathcal{D}$ on the set of variables $\boldsymbol{X}$ and a set of persons indexed by $[1..N]$ is said to be complete over a period $\mathcal{T}$ when the set $\mathcal{D} = \{(X, x, t, n), X \in \boldsymbol{X}, t \in \mathcal{T}, n \in [1..N]\}$ includes exactly one value for each element $(X, t, n) \in \boldsymbol{X} \times \mathcal{T} \times [1..N]$.

When we consider a specific ordered subset of variables $\mathbf{X_J} = (\ldots, X_j, \ldots)$ and one of its a possible setting $\mathbf{v} = (\ldots, v_j, \ldots)$, we write that $(\mathbf{X_J}, \mathbf{v}, t, n) \in \mathcal{D}$ to denote that for each variable $X_j \in \mathbf{X_J}$ and its value v_j, the element (X_j, v_j, t, n) belongs to the temporal data set $\mathcal{D}$.

Bayesian network

In order to set the way each variable evolves with time, it appears useful to take into account the value of some other variables. Indeed, for a given variable, different schema of temporal changes can be defined depending on the values of some other variables.

In that aim, we use a Bayesian network to define the dependence between variables (Naïm et al. 2011). We denote $pa(X)$ the parents of the variable X in the graph of the Bayesian network.

In that work, we assume that the Bayesian network graph does not change with time. As a consequence, each variable is associated with a *context* defined by the values of its parents in the graph.

We denote $(X, \mathbf{v})$ a variable X in a context $\mathbf{v}$, where $\mathbf{v}$ is one of the possible combination of values of the parents of X in the graph of a Bayesian network $\mathcal{B}$: $\mathbf{v} \in Dom(pa(X))$. We name such a couple a *contextualized variable*, and say that $\mathbf{v}$ is one of the context of X in $\mathcal{B}$.

In order to simplify the notation, when X has no parent in the graph of the Bayesian network, ($pa(X) = \emptyset$), the couple $(X, \mathbf{v})$ represents the variable X.

In the following, we present a schema of temporal evolution for each contextualized variable $(X, \mathbf{v})$.

Persistent variable

We got a better understanding of the way the variables change with time thanks to the interviews with Professor Puisieux. It appears that variables can be classified in several classes regarding the characteristics of their change over

[1] We do not use a specific notation to distinguish the different values of X in $Dom(X)$

time. Except constant variables that never changes, such as the sex, we define the concept of persistent variable as the simplest class regarding temporal evolution.

Definition 3 ***Positive (resp. negative) persistent variable***
A binary variable X with $Dom(X) = \{0, 1\}$ is said to be positive persistent in a temporal data set $\mathcal{D}$ when its value never changes after the value becomes 1 for a given person indexed by n :

$$\forall t, t' \in \mathcal{T},\ \textit{with}\ t' > t, (X, 1, t, n) \in \mathcal{D} \Rightarrow (X, 1, t', n) \in \mathcal{D}$$

Respectively, the value of a negative persistent variable never changes after it becomes zero.

As a consequence, when we consider a population composed of a group of persons, the proportion of persons with X being positive increases with the age of the persons. Thus, when a variable X is positive persistent, the function $f(age) = P(X = 1 \mid age)$ is an increasing function, where $P(X = 1 \mid age)$ denotes the probability for a variable X to be positive among the given age group[2].

In this work, we consider only persistent variables. Figure 1 shows the graph of the Bayesian network for the five variables that we consider in this article. To get it, we first learned a Bayesian network from the real data set, then we removed the arc $conduit \rightarrow demence$ so that every node has at most two parents. Indeed, the number of combinations of the values of the parents is higher with three parents, making possible that some cases have no representing sample in the data set.

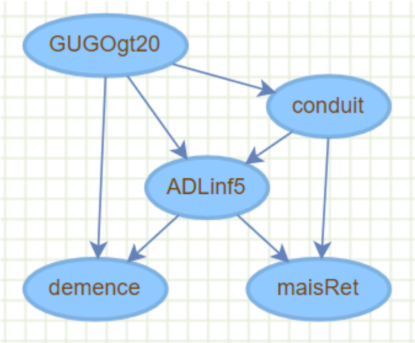

Figure 1: Graph of the Bayesian network.

Since our real data set includes the age of the persons, we uses that information to extract temporal behavior of the variables : we assume that the distributions of each variable in function of the age on the whole population allows us to derive the evolution of the variables for a given person regarding her age. In that aim, we plot the distribution of each variable regarding the age of the persons from the real data set (see Figure 2).

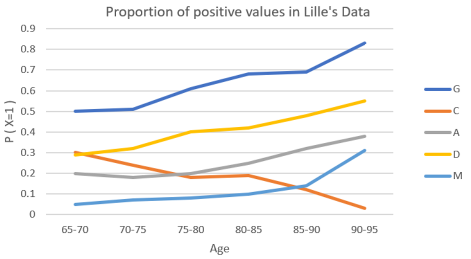

Figure 2: Proportion of positive values for each variable according to the age in Lille's real data set.

In order to simulate temporal series for each of these variables, we consider a linear interpolation for each curve. This approximation, combined with the feature of persistent variables, is used to compute the probability of a variable to become positive at a defined time step, when it was negative at the previous time step. It is important to note that more complex interpolation functions could be used in the algorithm that we propose.

In addition, we also want that the simulated data set reflects the dependencies between variables, such as described by a Bayesian network. In that aim, we make further assumption described below.

Parent-persistent contextualized variable

Our goal is now to combine information given by dependencies between variables in the Bayesian network and information from the distribution of positive value in function of the age, in order to simulate a temporal data set. In that aim, we introduce a concept related with the evolution of a variable in the context of its parents' values, which extends the concept of persistent variable.

Consider a Bayesian network $\mathcal{B}$ on a set of variables including a binary variable X, and let $\mathbf{v} \in Dom(pa(X))$ be a context of X in $\mathcal{B}$.

Definition 4 ***Positive (resp. negative) parent-persistent contextualized variable***
A contextualized variable $(X, \mathbf{v})$ is said to be positive parent-persistent in a temporal data set $\mathcal{D}$ when its value in the context $\mathbf{v}$ never changes after the value becomes 1 for a given person indexed by n:

$$\begin{aligned} &\forall t, t' \in \mathcal{T},\ \textit{with}\ t' > t, \\ &\textit{if for each couple}(Y, y)\ \textit{with}\ Y \in pa(X) \\ &\textit{and}\ y\ \textit{is the value of}\ Y\ \textit{in}\ \mathbf{v}, \\ &(Y, y, t, n) \in \mathcal{D}\ \textit{and}\ (Y, y, t', n) \in \mathcal{D},\ \textit{then} \\ &(X, 1, t, n) \in \mathcal{D} \Rightarrow (X, 1, t', n) \in \mathcal{D} \end{aligned} \tag{1}$$

Respectively, we consider also negative parent-persistent contextualized variable.

Remark: If a variable X is positive persistent in a data set $\mathcal{D}$, then for any context $\mathbf{v}$, the contextualized variable $(X, \mathbf{v})$ is positive parent-persistent in $\mathcal{D}$.

In this work, we thus assume that all contextualized variables are parent-persistent. For convenience, we also speak

[2] In our simulated temporal data set, the number of persons is constant whatever the age group. On the contrary, in the real static data set, the distribution regarding the age is not constant, making important to consider the conditional probability and not the joint probability.

about contextualized variable when the set of parents is empty.

Figure 3 shows the distribution of each contextualized variable in function of the age, computed from our real static data set. These plots are based on intervals of five years for the age. For each contextualized variable $(X, \mathbf{v})$, we plot the proportion

$$\frac{\#samples(X = 1, Pa(X) = \mathbf{v}, Age = a_l)}{\#samples(Pa(X) = \mathbf{v}, Age = a_l)}$$

where $Dom(Age) = a_1, \dots a_l, \dots$, computed from the real static data set.

More data is needed to obtain smoother curves.

Linear assumption

Figure 3 shows the distribution of contextualized variables regarding the age of patients. These curves are based on a discretization of Lille's data set with intervals of 5 years, which is a compromise between information quality and statistical quality.

In order to generate temporal data with any desired temporal granularity, while remaining faithful to the real data set, we replace these curves by interpolated functions. We choose linear interpolation:

- the functions $f(age) = P(X = x \mid age)$ are linear functions, for all $X \in \mathbf{X}$,
- the functions $f(age) = P(X = x \mid \mathbf{X_J} = \mathbf{v}, age)$ are linear functions, for all $\mathbf{v} \in Dom(\mathbf{X_J})$ where $\mathbf{X_J} = pa(X)$.

From the distributions plotted in Figure 3, we have defined a linear function associated with each contextualized variable.

About Survival Analysis

The functions shown in Figures 2 and 3 are very similar with survival functions and hazard rate conditional (Kleinbaum and Klein 2010). We show the evolution of the risk whereas survival functions usually show the chances that a person survives. In our case, the event of interest is the change of value of a risk factor, from absent (0) to present (1). Some methods to estimate Survival function are based on the assumption that data follows some distribution (such as exponential, gamma, weibull, log-normal etc.) and then we calculate its parameters. Other methods such as 'Kaplan-Meier' estimator do not have any prior assumptions. However, estimating survival function from data supposes that data include information about the response for each subject. In that kind of data, the subject is always "alive" when the study period starts, and the event of interest may or not occurs before the end of that period. When the event does not occurs, the survival time is labelled as 'Censored'.

In our case, our data from Lille's Hospital are very different since it corresponds to a single moment of observation for each subject, and we do not know when the risk occurs. At the moment of the observation, the risk is present for some person and absent for others. Because we do not

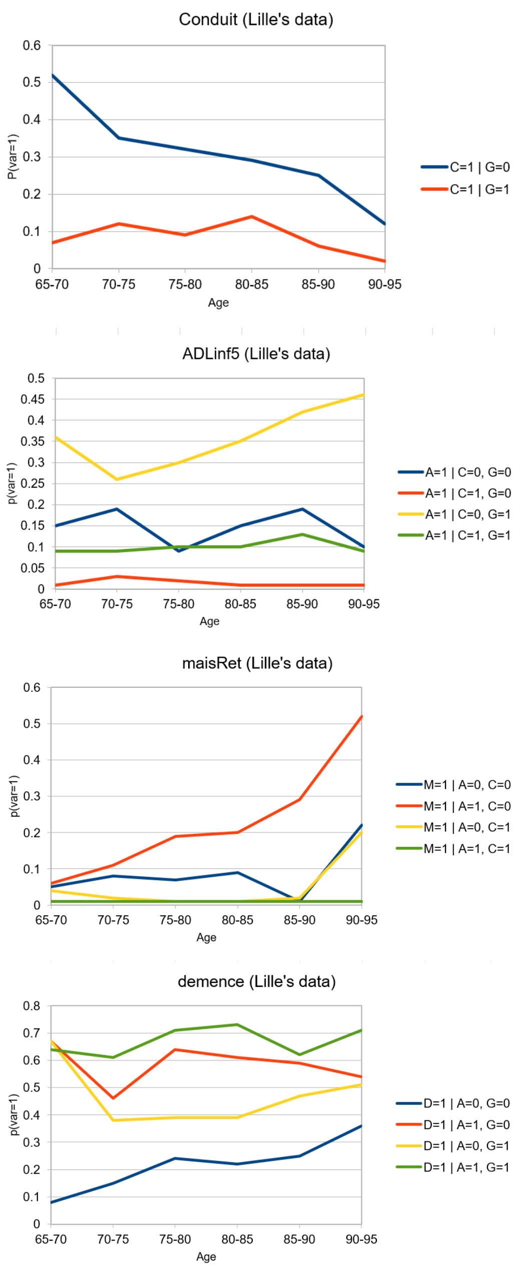

Figure 3: Proportion of positive values of each contextualized variable in function of the age.

have information about the time when the event of interest occurs, we take benefit from the fact that the observed population involves persons of different ages, and we assume that the proportion of persons with a risk factor at a given age may give us a way to estimate the survival function.

Assumptions regarding the period of time over which data are simulated

In order to simplify the simulation of the data set, we assume that the period $\mathcal{T}$ starts at time t_0 with all persons being 65 years old. This assumption can be easily removed later by shifting each data row randomly in time. This would allow to get a data set in which people of any age over 65 are considered at time t_0.

Second, we simulate data for all persons during the complete period, meaning that we do not consider the death of people. When we want to remove that second assumption, the age of death could be simulated on the basis of general knowledge about the distribution of the age of death.

In addition, let's precise that we consider $\Delta t = 6$ months.

Algorithm to simulate temporal data set from a static data set

The objective is to generate a temporal data set. The real data set includes the age of the person.

We now describe an algorithm to generate a temporal data set on a set of variables $\mathbf{X}$, on the basis of a real data set that includes a set of observations collected only once for a given person. We follow two main ideas: (1) respect the proportion of positive values for each variable given the age of the patient, (2) respect the dependence between the variables as described by a Bayesian network $\mathcal{B}$ learned on $\mathbf{X}$.

Let $\mathbf{X} = \{X_1, \ldots X_i, \ldots\}$ such that the order of the variables in $\mathbf{X}$ is compatible with the partial order defined by the graph of the Bayesian network $\mathcal{B}$.

The algorithm `simulateTDS` generates a value for each variable X_i at each time on a given period, based on the values of the parents of X_i in the graph of the Bayesian network at the same time and on the value of X_i at the previous time.

As explained above, we extract the temporal behavior of each contextualized variable regarding the age of the person.

Two basic functions are used by the algorithm to simulate the temporal data set: `parents(X)` and `linearF(X,`$\mathbf{v}$`,a)`. These functions are based on a Bayesian network $\mathcal{B}$ and the set of linear functions associated with each contextualized variable $(X, \mathbf{v})$, where X is a variable associated with a node of $\mathcal{B}$, and $\mathbf{v}$ is a vector of values of the parents of X in $\mathcal{B}$. The function `parents(X)` returns the list of variables that are parents of the variable X in the graph of the Bayesian network. The function `linearF(X,`$\mathbf{v}$`,a)` returns the value of the linear function associated with the contextualized variable $(X, \mathbf{v})$ for the age a. In addition, the function `generate(p)` returns 0 or 1 with probability distribution $(1-p, p)$ where the parameter p is a value in $[0, 1]$.

In order to simplify the presentation of the algorithm, we assume that we have only positive persistent variables. Indeed, a negative persistent variable can be replaced by a positive persistent variable by exchanging values 0 and 1. The temporal data are generated with a regular time step for a given number of iterations, and a given number of samples. The `simulateTDS` algorithm generates each new value and fills gradually a 3D table whose dimensions correspond to the samples, the variables and the time (lines 1–3).

The values are generated following a partial order of the variables so that the values of the parents of a given variable can be used to generate the value of this variable, and following the temporal order, so that the previous value of a variable can be used to generate its next value.

The operations to generate a value for a given person (or sample), a given variable and a given time are detailed in the `simulateOne` algorithm. At first, the context of the variable is identified by extracting the value of its parents from the data already simulated (lines 1–2). In order to generate a value of a variable at the first time step (age = 65), one generates randomly 0 or 1 with a uniform probability corresponding to the value of 65 for the linear function associated with the contextualized variable (lines 3–5). When a previous value has already been generated for a given variable, the value to be generated depends on it: When the previous value is 1, the new value has to remain 1, by definition of a positive persistent variable (lines 6–7). When the previous value is 0, one generates randomly 0 or 1 with a uniform probability corresponding to the increase of the linear function associated with the contextualized variable during one step of time, and reduced to the negative cases (lines 9–12). Remark that this step is based on the interpolated functions, but does not required these functions to be linear.

```
Algorithm 1: simulateTDS($\mathbf{X}$, $K$, $\Delta t$, $N$)
   Input: $\mathbf{X}$ ▷ an ordered set of variables
   Input: $K$ ▷ number of temporal iterations
   Input: $\Delta t$ ▷ length of the time step
   Input: $N$ ▷ number of samples
   Output: $D$ ▷ a 3-dimension table containing the simulated
           temporal data set on $\mathbf{X}$ over the period $\mathcal{T}$. The
           cell $D[n, i, t]$ contains the simulated value of
           sample $n$ for the variable $X_i$, at time $t$.
1  $D \leftarrow 0$                 ▷ initialize the 3D array to zero
2  foreach $k \in [1..K]$ do           ▷ generate data at time $t_k$
3      foreach person $n \in [1..N]$ do   ▷ generate $N$ samples
4          foreach variable $X_i \in \mathbf{X}$ (in topological order) do
               ▷ generate value of $X_i$ at $t_k$
5              $D[n, i, k] \leftarrow \mathrm{simulateOne(n, i, t_k, \Delta t, D)}$
6  return $D$
```

Evaluation of the simulated temporal data set

Using this algorithm and the real static data set of Lille, we have simulated a temporal data set of 2000 cases, over a period of 30 years, with a time step of 6 months.

In order to evaluate the quality of the simulated temporal data set, we plot the proportion of positive values for each variable in the simulated data set (Figure 4). In comparison to Figure 2, the result clearly shows that the linear assumption is faithfully reproduced in the simulated data set, when

```
Algorithm 2: simulateOne($n, i, t_k, \Delta t, D$)
Input: $n$ ▷ sample index
Input: $i$ ▷ variable index
Input: $t_k$ ▷ time to be simulated
Input: $\Delta t$ ▷ length of the time step ($t_k - t_{k-1}$)
Input-Output : $D$ ▷ a 3-dimension table containing the
               already simulated temporal data set on $\mathbf{X}$
               over the period $\mathcal{T}$
Data: a Bayesian network $\mathcal{B}$
Data: Linear functions associated with each contextualized
      variable $(X, \mathbf{v})$, regarding $\mathcal{B}$
$\mathbf{X_J} \leftarrow$ parents($\mathrm{X_i}$)
$\mathbf{v} \leftarrow$ values of $\mathbf{X_J}$ generated at $t_k$        ▷ a context of $X_i$
if $k = 0$ then                                        ▷ first time step $t_0$
    $p \leftarrow \mathrm{linearF}(\mathrm{X_i}, \mathbf{v}, 65)$ $D[n, i, 0] \leftarrow \mathrm{generate(p)}$
else                     ▷ generate data at time $t_k$ from value at $t_{k-1}$
    if $D[n, i, k-1] = 1$ then                        ▷ previous value
        $D[n, i, k] \leftarrow 1$                  ▷ positive persistent variable
    else  ▷ compute the probability to become 1 among the
          negative cases
        $c \leftarrow \mathrm{linearF}(\mathrm{X_i}, \mathbf{v}, \mathrm{t_k}) - \mathrm{linearF}(\mathrm{X_i}, \mathbf{v}, \mathrm{t_{k-1}})$
        $p \leftarrow c/(1 - \mathrm{linearF}(\mathrm{X_i}, \mathbf{v}, \mathrm{t_{k-1}}))$
        $D[n, i, k] \leftarrow \mathrm{generate(p)}$
```

considering each variable separately.

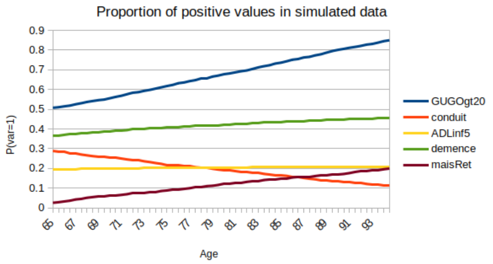

Figure 4: Proportion of positive values of each variable in simulated data.

In addition, the algorithm of data simulation is also based on information provided by the Bayesian network learned on the real static data set, by taking into account the way each variable changes in a given context. In order to evaluate that second point, we show in Figure 5 the proportion of positive values for each variable in a given context along with time in the simulated data set, and we compare with the linear functions computed for each contextualized variable defined from the static data set (based on Figure 3).

The comparison of the linear functions and the plot from simulated data shows that in most cases, the proportion of positive values in the simulated data is faithful with the linear functions.

Perspective and conclusion

This article proposes a first attempt to simulate temporal data using a Bayesian network in the aim to complete a real static data set to be applied in the context of fall prevention for

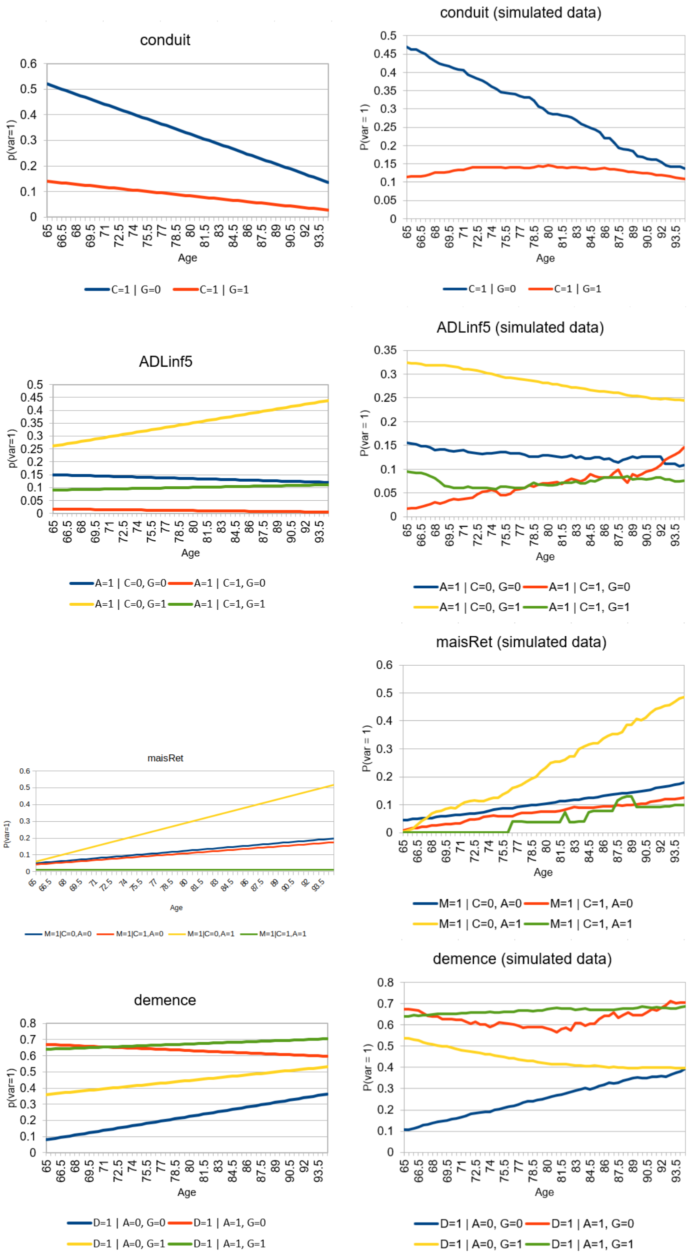

Figure 5: Linear functions associated with each contextualized variable (left) and Proportion of positive value of each contextualized variable in simulated data (right).

elderly people.

We combine several assumptions and expert knowledge in order to provide a temporal data set that is faithful with the real data set. In that aim, we select a small number of variables of the real data set regarding their schema of temporal evolution. We focus on a subset of persistent variables whose value may evolve only once in the life of a person, e.g., from zero to one for the positive persistent variables. This concept emerged during discussions with experts about temporal changes of a large set of variables of interest for fall prevention.

The persistent feature of the selected variables is visible on the plot of the distributions of positive values as functions of the age. We assume that these distributions can be used as a basis of the evolution of the associated variables for a given person. We also consider a set of possible contexts for each variable, as defined by a Bayesian network learned on the static real data set. Finally, we use linear interpolation to get a simple model of the proportion of positive values for each contextualized variable.

On this basis, we propose an algorithm to simulate a temporal data set. The results are evaluated through the comparison of the temporal distributions in the temporal data set generated thanks to the algorithm and the linear functions computed from the real data set.

We are aware that the data are generated on the basis of two strong assumptions: the persistence of the concerned variables and the linear approximation of their distribution according to the age. However, the data generation algorithms and their first results consist in a first step toward the necessary filling of the data gaps in our health application context.

As a perspective, we now intend to exploit this data set in the context of fall prevention. More precisely, the objective is to predict some risk factors for fall based on a partial set of time stamped observations. In this problem, the challenge is first to take benefit from old observations, meaning that the value of some variable observed in the past may have changed, and second to reason with a number of observations that can be arbitrary small.

About perspectives on data simulation, it could be interesting to compare our data simulation with the one obtained by a dynamic Bayesian network when linear functions are used to approximate the dynamic of contextualized variables, since it makes changes time invariant. Other perspectives concern the use of non linear functions for interpolation, the inclusion of variables with other temporal schema, such as semi-persistent variable, and variables with larger domain.

Acknowledgments

This research is supported by the Hauts-de-France region, the University of Mons, Belgium, the Ministry of Higher Education and Research and the National Center for Scientific Research. Also, we are grateful to Pr. F. Puisieux and Dr. C. Gaxatte and the service of fall prevention of Lille's hospital for their support.

References

Dargent-Molina, P.; and Bréart, G. 1995. Epidemiology of falls and fall-related injuries in the aged. *Revue d'Épidémiologie et de Santé Publique* 43(1): 72–83.

de Vries, J.; Kraak, M. H.; Skeffington, R. A.; Wade, A. J.; and Verdonschot, P. F. 2021. A Bayesian network to simulate macroinvertebrate responses to multiple stressors in lowland streams. *Water Research* 194: 116952. ISSN 0043-1354.

Delcroix, V.; Essghaier, F.; Oliveira, K.; Pudlo, P.; Gaxatte, C.; and Puisieux, F. 2019. Towards a fall prevention system design by using ontology. *en lien avec les Journées francophones d'Ingénierie des Connaissances, Plate-Forme PFIA* .

Delcroix, V.; Grislin-Le Strugeon, E.; and Puisieux, F. 2021. A knowledge based system for the management of a time stamped uncertain observation set with application on preserving mobility. *International Journal of Approximate Reasoning* 134: 53–71.

Ducamp, G.; Bonnard, P.; Nouy, A.; and Wuillemin, P. 2020. An Efficient Low-Rank Tensors Representation for Algorithms in Complex Probabilistic Graphical Models. In Jaeger, M.; and Nielsen, T. D., eds., *International Conference on Probabilistic Graphical Models, PGM 2020, 23-25 September 2020, Aalborg, Denmark*, volume 138 of *Proceedings of Machine Learning Research*, 173–184. PMLR.

Gootjes-Dreesbach, L.; Sood, M.; Sahay, A.; Hofmann-Apitius, M.; and Fröhlich, H. 2020. Variational Autoencoder Modular Bayesian Networks for Simulation of Heterogeneous Clinical Study Data. *Frontiers in Big Data* 3: 16. ISSN 2624-909X. doi:10.3389/fdata.2020.00016. URL https://www.frontiersin.org/article/10.3389/fdata.2020.00016.

Homer, M. L.; Palmer, N. P.; Fox, K. P.; Armstrong, J.; and Mandl, K. D. 2017. Predicting Falls in People Aged 65 Years and Older from Insurance Claims. *The American Journal of Medicine* 130(6): 744.e17–744.e23.

Katz, S.; Ford, A. B.; Moskowitz, R. W.; Jackson, B. A.; and Jaffe, M. W. 1963. Studies of illness in the aged. The index of ADL: a standardized measure of biological and physical function. *journal of American Medical Association* 185: 914–9.

Kleinbaum, D. G.; and Klein, M. 2010. *Survival analysis*, volume 3. Springer.

Marier, A.; Olsho, L.; Rhodes, W.; and Spector, W. 2016. Improving prediction of fall risk among nursing home residents using electronic medical records. *Journal of the American Medical Informatics Association* 23(2): 276–82.

Marini, S.; Trifoglio, E.; Barbarini, N.; Sambo, F.; Di Camillo, B.; Malovini, A.; Manfrini, M.; Cobelli, C.; and Bellazzi, R. 2015. A Dynamic Bayesian Network model for long-term simulation of clinical complications in type 1 diabetes. *Journal of Biomedical Informatics* 57: 369–376. ISSN 1532-0464. doi:https://doi.org/10.1016/j.jbi.2015.08.021. URL https://www.sciencedirect.com/science/article/pii/S1532046415001896.

Marschollek, M.; Gövercin, M.; Rust, S.; Gietzelt, M.; Schulze, M.; Wolf, K.-H.; and Steinhagen-Thiessen, E. 2012. Mining geriatric assessment data for in-patient fall prediction models and high-risk subgroups. *BMC Medical Informatics and Decision Making* 12(19): 1–6.

Mlinac, M. E.; and Feng, M. C. 2016. Assessment of Activities of Daily Living, Self-Care, and Independence. *Archives of Clinical Neuropsychology* 31(6): 506–516.

Mrad, A. B.; Delcroix, V.; Piechowiak, S.; Leicester, P.; and Abid, M. 2015. An explication of uncertain evidence in Bayesian networks: likelihood evidence and probabilistic evidence - Uncertain evidence in Bayesian networks. *Appl. Intell.* 43(4): 802–824.

Murphy, K. P. 2002. *Dynamic bayesian networks: Representation, inference and learning*. Ph.D. thesis, University of California, Berkeley.

Naïm, P.; Wuillemin, P.; Leray, P.; Pourret, O.; and Becker, A. 2011. *Réseaux bayésiens*. Algorithmes. Eyrolles. ISBN 9782212047233. URL https://books.google.fr/books?id=7d_Jq2ehb0oC.

Rubenstein, L. Z. 2006. Falls in older people: epidemiology, risk factors and strategies for prevention. *Age and ageing* 35(suppl_2): ii37–ii41.

Sihag, G.; Delcroix, V.; Grislin, E.; Siebert, X.; Piechowiak, S.; and Puisieux, F. 2020. Prediction of Risk Factors for Fall using Bayesian Networks with Partial Health Information. In *Globecom AIdSH Workshop*, 1–6. IEEE.

World Health Organization. 2008. WHO global report on falls prevention in older age. https://apps.who.int/iris/handle/10665/43811.

Ye, C.; Li, J.; Hao, S.; Liu, M.; Jin, H.; Zheng, L.; Xia, M.; Jin, B.; Zhu, C.; Alfreds, S. T.; Stearns, F.; Kanov, L.; Sylvester, K. G.; Widen, E.; McElhinney, D.; and Ling, X. B. 2020. Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm. *International Journal of Medical Informatics* 137: 104105.