Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine Update

journal homepage: www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine-update

Evaluation of risk factors for fall in elderly using Bayesian networks: A case study

Gulshan Sihag [a], Véronique Delcroix [*,a], Emmanuelle Grislin-Le Strugeon [a,b], Xavier Siebert [a,c], Sylvain Piechowiak [a], Cédric Gaxatte [d], François Puisieux [d]

[a] Univ. Polytechnique Hauts-de-France, LAMIH, CNRS, UMR 8201, Valenciennes F-59313, France
[b] INSA Hauts-de-France, Valenciennes F-59313, France
[c] Univ. de Mons, Faculté Polytechnique, Département de Mathématiques et Recherche Opérationnelle, Belgium
[d] Pôle de Gérontologie, Hôpital Universitaire de Lille, Lille Cedex 59037, France

ARTICLE INFO

Keywords:
Health data
Knowledge model
Fall prevention in elderly
Bayesian network
Classification
Reasoning with uncertainty
Partial observations

ABSTRACT

Background: Falls in the elderly are the number one cause of traumatic death in this population. Prevention of falls requires to evaluate which risk factors for fall are present for a person on the basis of available health information. Our objective is to predict the presence or the absence of 12 risk factors for fall in elderly people based on partial observations.

Methods: A data set of 1810 patients of the multidisciplinary falls consultation of Lille University Hospital covering fourteen years admissions were used to learn and evaluate a Bayesian network and four usual machine learning classifiers. Variable selection and data pre-processing were achieved on the basis of an ontology and interviews of the experts. The prediction of each target risk factor using the complete set of observations is first compared with the prediction based on a specific subset of variables, and second based on partial observation, from 10 to 90% of the variables.

Results: For 7 out of 12 target risk factors, the f1-score of classifiers using complete set of variables is slightly better than the specific subset of variables, with a difference of less than 3%. Bayesian Networks and other classiers perform equivalently in terms of accuracy and f1-score. The best prediction were obtained for the loss of autonomy and osteoporosis with a f1-score from 15 to 20% better than the baseline classifier when using the Bayesian network. At the opposite, for 3 risk factors, no classifier allows to improve the f1-score or the accuracy of more than 1% compared to the baseline classifier.

Conclusion: Our results show that the use of specific subsets of variables does not improve the prediction of risk factors, and that no classifier outperform the others. However Bayesian networks perform well and are interesting due to their explainability.

1. Introduction

Falls present a striking danger to health and safety in older people [1–3]. 3.8 million older people attend emergency departments each year with fall related injuries. The medical costs associated with fall-related injuries is approximately 25 billion euros per year. If effective prevention strategies are not established, the total cost of treating fall related injuries in European Union is expected to increase up to 45 billion euros per year by 2050 [4].

The use of Machine Learning algorithms to detect health related risks in patients is now common [5–7] and a large number of successful studies have addressed the problem of falls in elderly [2,3,8,9]. However, the evaluation of risk factors for fall remains a challenge since it requires time and expertise, and specific tests and devices may also be necessary. Moreover, the family physician who is one of the main actors of fall prevention generally does not have a lot of time, whereas fall prevention requires a pedagogical and repeated approach. As a consequence, the collection of information for a complete evaluation of risk factors is not feasible regularly and the risk factors for fall of a person should be assessed from an incomplete set of observations. In order to

* Corresponding author.
E-mail addresses: Gulshan.Sihag@uphf.fr (G. Sihag), Veronique.Delcroix@uphf.fr (V. Delcroix), Emmanuelle.Grislin@uphf.fr (E. Grislin-Le Strugeon), Xavier.SIEBERT@umons.ac.be (X. Siebert), Sylvain.Piechowiak@uphf.fr (S. Piechowiak), Cedric.GAXATTE@CHRU-LILLE.FR (C. Gaxatte).

https://doi.org/10.1016/j.cmpbup.2021.100035
Received 15 April 2021; Received in revised form 14 October 2021; Accepted 15 October 2021
Available online 16 October 2021

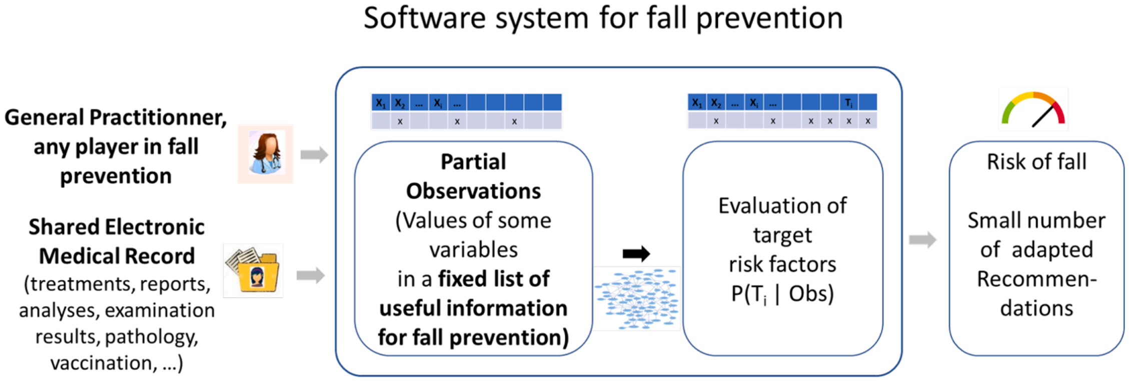

Software system for fall prevention
General Practitionner, any player in fall prevention
Shared Electronic Medical Record (treatments, reports, analyses, examination results, pathology, vaccination, ...)
Partial Observations (Values of some variables in a fixed list of useful information for fall prevention)
Evaluation of target risk factors P(T_i | Obs)
Risk of fall
Small number of adapted Recommen-dations

Fig. 1. Fall prevention system.

tackle that problem, this article is a first step toward a fall prevention aiding system. We propose the use of Bayesian networks (BNs) since these probabilistic graphical models allow the updating of beliefs under uncertainty [10,11]. In addition, BNs are understandable and modifiable by the experts thanks to the graph. These models are inherently explainable and allow transparency and visibility while decision aiding. A large number of successful studies exists in the literature that present a BN modeling approach combined with expert knowledge and applied to real-world problems [12–17]. A systematic review of BNs in healthcare is presented in McLachlan et al. [6].

Thus the objective of this article[1] is to identify the risk factors for fall in the elderly people on the basis of partial set of observations about the person. Section 2 gives description about the problem and data used. Section 3 describes the data preprocessing steps and selection of variables. Methodology is presented in Section 4. Section 5 shows the results and discussion. Finally, we conclude the article in Section 6 with possible future directions.

2. Problem and data

In this section, we first describe the problem; second, we present the data collection.

2.1. The problem

Fig. 1 shows a simplified view of the general architecture of the fall prevention aiding system. Lots of actors could contribute to fall prevention (doctors, family members etc.). The objective of a fall prevention aiding system is to propose them a small number of adapted recommendations regarding the elderly they care for. These recommendations are selected on the basis of the most important risk factors for fall that are present for that elderly.

The evaluation of risk factors for fall is a multi-factorial problem, and some risk factors can not be evaluated by a simple and rapid question. Moreover, the potential actors of fall prevention usually do not have much time for these questions, which makes it necessary to store useful information in a personal database. Fortunately, the value of some risk factors and useful variables to evaluate them can be automatically extracted from shared electronic medical records. However, some kinds of information are rarely present in the medical file, and the amount of information available for each elderly person is very different. Thus, the issue addressed in this paper is the evaluation of presence or absence of some risk factors for fall on the basis of partial observation.

2.2. Subjects and data collection

The 1810 patients of the multidisciplinary falls consultation of Lille University Hospital between January 2005 and December 2018 are included in the study. The patients are admitted in that service for a complete day, during which they meet different medical personnel and each of them explores a set of factors such as history of falls, nutrition, physical activities, medical tests such as balance test etc. At each step, the data collected about the person are registered. Then a team of specialists gathers around the case file of the patient and discusses about the most appropriate recommendations on the basis of the observed risk factors of the person. At the end of the day, a small number of appropriate recommendations is selected and explained to the patient. The patients are invited to come back 6 months later in the service for a short consultation during which an assessment is done regarding the recommendations and the number of falls during the last 6 months.

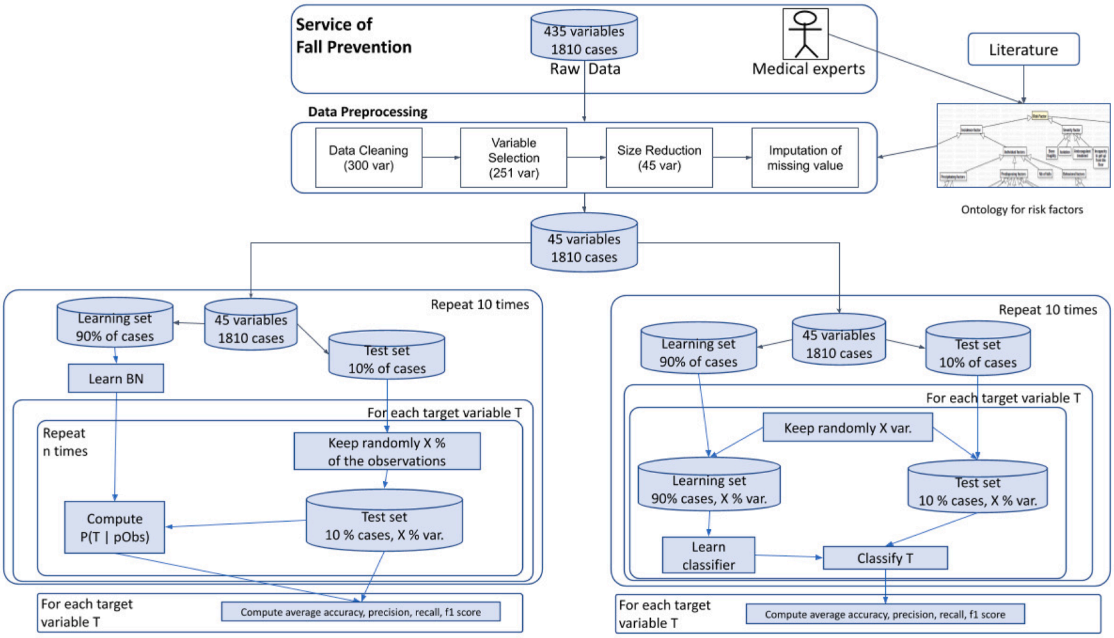
3. Data preprocessing and variable selection

Data preprocessing has a significant impact on the performance of machine learning models because unreliable samples may lead to wrong outputs [19,20]. To perform a meaningful data preprocessing, either the domain expert should be integrated in the data analysis or the domain should be extensively studied before analysis [21]. In this study the understanding of the data is facilitated by the help of experts and an ontology about fall prevention [22] developed previously with the same service of fall prevention of Lilles Hospital. The initial data set includes more than 400 columns, including a lot of details and redundancy. We use the following 3 criteria to select relevant variables:

(1) **Data quality**: in order to have relevant and understandable observation, we remove 49 variables whose content is not useful (free text, very heterogeneous type of values), 61 old grayed out variables, 11 variables with very unbalanced classes, 14 variables with more than 30% of missing values (total 135 variables removed).
(2) **Focus on risk factors of fall**: we remove 15 variables associated with recommendations and 34 variables associated with the second appointment after 6 months (total 49 variables removed).
(3) **Model size limitation**: we remove 42 too specific variables, 93 variables associated with medicines, and 6 variables related to blood pressure, 2 variables about patient ID and year of consultation. Furthermore, variables having the same meaning or a very close meaning have been regrouped. This preprocessing leads to 45 variables (shown in Appendix A).

Moreover, we use k Nearest Neighbors (kNN) methods for imputation of missing values since it is very simple and easy to use and it can be applied irrespective to the data type. The number of neighbors has been

[1] This article is an extension of our work presented in Sihag et al. [18]



(a) Use a Bayesian network as a classifier

(b) Use an other classifier (LR, DT, RF, SVM)

Fig. 2. A schematic representation of the methodology of risk factors prediction using partial observations .

Table 1
Comparison of accuracy (acc) and f1-score (f1) using specific subset (sss) for each risk factor vs. complete data (45 var).

RFFs		BN		LR		DT		RF		SVM	
		sss	45 var	sss	45 var	sss	45 var	sss	45 var	sss	45 var
trMar	acc	86.24	86.8	86.88	87.06	80.65	80.36	86.65	86.57	**87.28**	87.24
	f1	0.92	0.92	0.92	0.92	0.88	0.88	0.92	0.92	0.93	0.93
peurTom	acc	78.7	**79.8**	78.98	79.04	73.94	72.93	77.16	79.02	78.82	78.99
	f1	0.87	0.88	0.87	0.87	0.83	0.82	0.86	0.88	0.87	0.88
dfOufaim	acc	68.69	69.49	70.06	69.5	60.04	60.19	69.32	**70.99**	70.15	70.59
	f1	0.8	0.8	0.79	0.79	0.69	0.69	0.79	0.8	0.8	0.8
trEq	acc	82.2	**82.51**	81.47	81.03	71.06	70.19	81.3	81.46	82.09	82.06
	f1	0.89	0.89	0.88	0.88	0.8	0.8	0.88	0.88	0.89	0.89
auTrNeur	acc	71.01	71.17	71.57	71.17	58.14	60.19	68.83	**71.71**	71.12	71.69
	f1	0.82	0.82	0.82	0.82	0.69	0.71	0.8	0.82	0.82	0.83
nbchu2	acc	57.15	59.19	61.57	61.34	55.26	55.55	58.52	61.61	**62.14**	61.98
	f1	0.69	0.71	0.7	0.7	0.61	0.62	0.67	0.71	0.72	0.71
ADLinf5	acc	78.5	79.16	**80.4**	80.22	70.9	79.71	79.68	79.91	80	79.9
	f1	0.53	0.55	0.54	0.54	0.45	0.44	0.49	0.47	0.48	0.49
demence	acc	66.48	67.22	68.15	**69.27**	58.18	58.64	65.88	68.6	67.86	68.8
	f1	0.55	0.54	0.59	0.61	0.51	0.51	0.56	0.58	0.58	0.58
newHypoT	acc	67.39	67.47	**67.72**	66.87	60.2	56.61	62.48	67.55	67.37	67.35
	f1	0.01	0.02	0.18	0.22	0.33	0.35	0.31	0.14	0.04	0.03
dep	acc	73.73	73.9	75.07	73.7	72.82	67.62	73.7	73.78	**75.11**	74.56
	f1	0.45	0.42	0.47	0.46	0.46	0.45	0.5	0.35	0.47	0.4
osteoconf	acc	81.65	82.26	83.24	**83.46**	76.41	76.2	80.97	82.3	82.72	82.52
	f1	0.49	0.52	0.46	0.48	0.38	0.4	0.39	0.3	0.4	0.32
parkOuSP	acc	83.4	83.48	**83.88**	83.69	74.94	72.91	81.7	83.3	83.44	83.47
	f1	0.06	0	0.12	0.19	0.27	0.25	0.16	0.02	0.01	0

set to five after evaluating different choices.

4. Methodology

In this section, we first introduce the construction of Bayesian networks and other classifiers; then we describe our target variables and the procedure of variable selection. Furthermore, we present the algorithm used to predict the presence or absence of risk factor for fall.

4.1. *Bayesian networks and other classifiers construction*

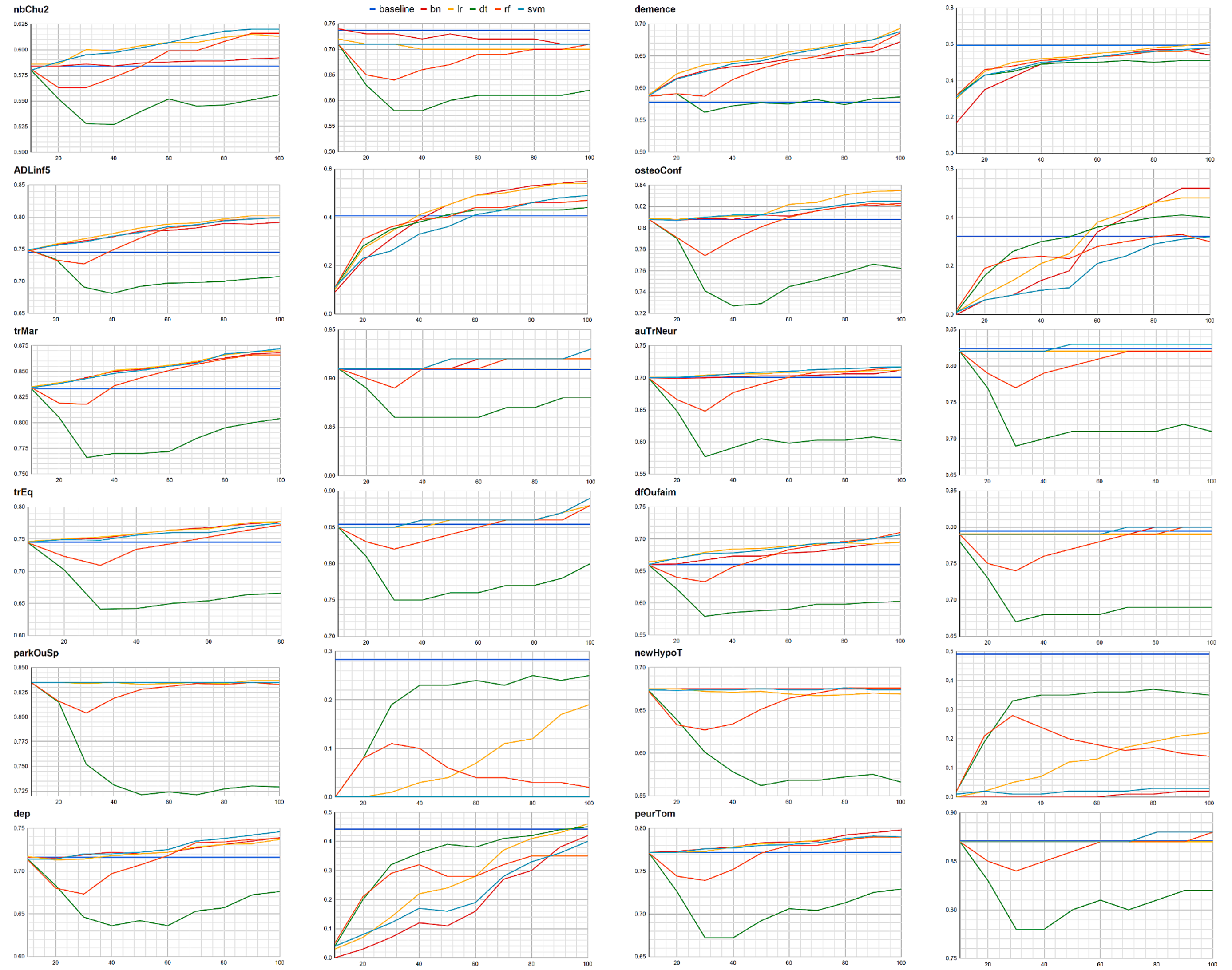
A Bayesian network (BN) is a graphical representation of a set of variables U = $\{X_1, X_2,..., X_n\}$ with a joint probability that can be factorized as follows:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|Parent(X_i))$$

where $Parent(X_i)$ is the set of variables that correspond to direct predecessors of X_i in the graph. It consists of a directed acyclic graph and a set of the local probability distributions, one for each node (variable) [11]. Both the structure of the graph and parameters (the conditional probabilities) can be learned from data or obtained by experts using the domain knowledge [23]. In this article we learn BN model from data.

Table 2
List of figures representing the result of accuracy and f1 score for a given risk factor respectively. Horizontal axis represents the % of available observations .

Accuracy	F1 score	Accuracy	F1 score



The methods to learn the structure of a BN model from data are categorized as follows: (1) *Constraint-based approaches* aim at building a graph structure to reflect the conditional independence relations in the data that match the empirical distribution, (2) *Score-based approaches* aim at maximizing the likelihood of the data given the model based on score functions, (3) *Hybrid approach* which is the combination of the two above [24]. We prefer the score based approach over a constraint based approach since they are often more accurate [25]. In that aim, we use hill climb algorithm [26] with Bayesian Information Criterion (BIC) scoring function because it provides a graph with not too many arcs, which is important since we have a moderate amount of data. Parameters are obtained by Bayesian estimator [27].

4.2. Target variables

Among the list of variables (in Appendix A), 12 target variables (shown in bold) are selected for prediction since information about these risk factors is frequently not available outside of specialized fall prevention services. The detection of these risk factors for fall contributes to evaluate the risk of fall and is essential to select the most useful recommendations. For some factors, it may also warn the physician that further investigation should be done, and in other cases, it may help to prevent future presence of that risk factor.

4.3. Variable selection for each target risk factor

In addition of using the complete set of variables to predict each target risk factor, we also use variable selection to remove non-informative or redundant predictors from the model. This approach prevents too large number of variables to slow the development and training of models. Many studies focus on the evaluation of the variables which are most affecting to a given risk factor (see for example about orthostatic hypotension [28], and about fear of falling [29]). In our study, we use the chi square method in order to identify a subset of variables associated with each of the target risk factors for fall. We consider a significance level of 0.05, which is the usual value, meaning that all the variables with a significance level of less than 0.05 are selected for a given target risk factor. We also compare the results with a significance level of 0.02 and it makes no change.

Table 3
List of variables and associated risk factors for fall (RFF) in the ontology.

Category of RFF	RFF	Variable	Short name	Prevalence
	age greater than 80	age greater than 80	*agegt80*	
	sex	sex	*sex*	
	body mass index	BMI	*BMI4*	
	number of falls	number of falls during the last six months	***nbChu2***	58%
precipitating factor	factors linked with medication	number of drugs	*nbMed3*	
	orthostatic hypotension	orthostatic hypotension	***newHypoT***	32.5%
	psychotropic drug	at least 1 psychotropic drug	*gt1psych*	
predisposing factor	balance impairment	balance impairment	***trEq***	74.5%
	gait impairment	gait impairment	***trMar***	83.3%
	sarcopenia	deficit of muscular strength or muscle weakness in the lower limbs	***dfOuFaiM***	66%
	loss of autonomy	activities of daily living (ADL) less than 5	***ADLinf5***	25.5%
	depression	depression	***dep***	28.4%
	neurological disorder	stroke or TIA	*AVCAIT*	
	neurological disorder	parkinson or parkinsonian syndrome	***parkOuSP***	16.5%
	neurological disorder	neurological disorder other than stroke, TIA, parkinson disease or dementia	***auTrNeur***	70.1%
	cognitive impairment and dementia	dementia (probable or confirmed or antecedent)	***demence***	42.2%
	locomotor system disorder	arthritis or rheumatoid arthritis	*arthPoly*	
	sensory disorder	vision disorder	*trVision*	
	sensory disorder	hearing disorder	*trAudit*	
behavioral factor	alcohol consumption	alcohol	*alc*	
	fear of falling	fear of falling	***peurTom***	77.2%
	use of assistive device	walking aids	*utiATM*	
severity factor	bone fragility	fracture during a fall or vertebral collapse	*fracturA*	
	bone fragility	confirmed osteoporosis	***osteoConf***	19.2%
	bone fragility	anti osteoporosis treatment	*newTrOst*	
	incapacity to get up off floor	was able to get up off floor on his own	*aSuSeRel*	
	incapacity to get up off floor	remained on the ground for more than one hour	*gt1hSol*	
	isolation	lives alone	*vitSeul*	

4.4. Algorithm to predict risk factors for fall

In order to estimate the risk factors based on available information, we build a BN model and compare the results with other classifiers, namely Logistic Regression (LR), Decision Tree (DT), Random Forest (RF) and Support Vector Machine (SVM). We first compare the results of the prediction using the complete set of variables and a specific subset obtained by variable selection as described above. Second, we achieve the prediction of the target risk factors based on partial observations, using a subset of randomly selected variables ranging from 10 up to 100% of the whole set of variables. Fig. 2 shows a schematic diagram about the methodology to predict the presence or absence of the risk factors. The left and right parts present the algorithm using BN, and the other classifiers (LR, DT, RF and SVM) respectively. The approach is roughly the same on both sides, except that the BN model (graph and parameters) is learned only once whereas the other classifiers have to be learned again for each target variable and each subset of variable. To evaluate the prediction model performance we use 10-fold cross validation. In each fold, 10% of cases are used as testing set and 90% of cases as training set. Then we compute the average over these 10 fold evaluation. The above procedure is repeated for several sets of observations with different size. We use accuracy, precision, recall and f1-score to evaluate the performance of the classifiers, and we compare these results with those obtained by a baseline classifier. For the comparison of accuracy (F1-score), we use a baseline classifier that always predicts the most frequent class (the positive class).

We use `pgmpy` [30] to learn the structure and parameters for our BN model and `Scikit-learn` for other classifiers. Based on several tests of hyperparameters, the best results are obtained for logistic regression with L2 regularization and *lbfgs* solver, for decision tree with *Gini* impurity, 100 trees, for random forest with *Gini* impurity, and for support vector machine classifier with *rbf* kernel.

5. Results and discussion

In this section, we first compare the quality of the prediction of target risk factors based on specific subsets of variables and the complete set of variables. Second, we present the evolution of the prediction quality as a function of the amount of available observations. Then we discuss about why BNs are a good choice to evaluate the risk factors for fall in real situations. Finally, we present the combined graph of the BNs learned from data.

5.1. Specific subset vs. complete set of variables

Table 1 presents the difference between the results when using complete set of variables (45 variables) and a specific subset of variables selected using hypothesis testing for each risk factors. We compare the accuracy and f1-score for each risk factor shown in the first and second rows, respectively. The maximum value of accuracy for a given risk factor is shown in bold.

Results in Table 1 shows that for 7 out of 12 target risk factors, the f1-score of classifiers using complete set of variables is slightly better than the specific subset of variables, with a difference of less than 3%. Since the difference is always very small, we can use any of the discussed scenario for prediction. We choose to use the complete set of variables because in our context where only partial information can be obtained, we have more chances to get the information about a given patient, since a piece of available information may partially compensate the absence of another element. We will discuss more about partial information below.

5.2. Performance using complete set of variables

Table 2 represents the accuracy (left) and f1-score (right) for our targets calculated using complete set of variables for different percentages of available observations. The horizontal axis represents the percentage of randomly selected observations used to predict the target risk factor starting from 10% up to 100%.

BNs provide roughly the same quality of results for the different

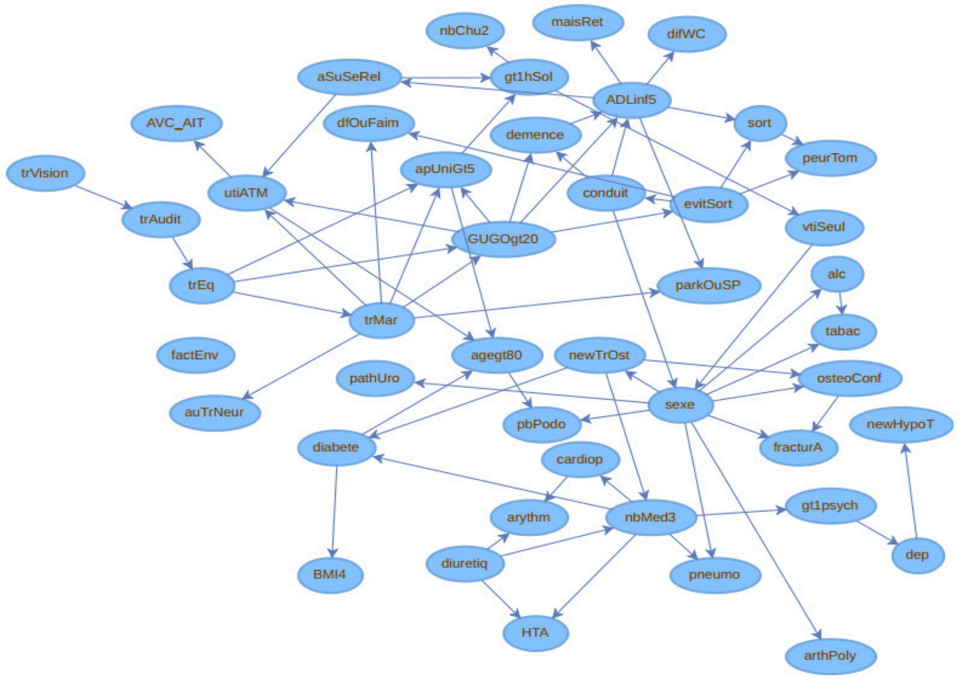


Fig. 3. A graph derived from one of the 10 BNs learnt during the 10 fold cross process: the arcs that belong to less than eight BNs have been removed, and the (single) arc that belong to eight (other) BNs has been added, after checking its direction.

target variables than the other classifiers with some variations according to the target risk factors for fall. None of the classifiers is clearly better than the others, even if LR provides sometimes slightly better results and DT most often slightly lower performance. For most of the variables, both accuracy and f1-score increase as the percentage of available observations increases. This illustrates clearly the influence of the quantity of observations on the ability to predict the presence of the risk factor. This increase is very clear for 4 out of the 6 variables whose prevalence is less than 50% (see Table 3): *ADLinf*5, *dep*, *demence*, *osteoConf*. For the variables *parkOuSP* and *newHypoT*, results are not good. For the variables whose prevalence is more than 50%, it seems that the behavior of most classifiers is to predict most often the presence of the risk factor, just as the baseline classifier.

The results in Table 2 show that Bayesian networks and other classiers perform equivalently in terms of accuracy and f1-score. We are able to evaluate most of the risk factors (even with a low improvement compared to the baseline classifier). The best prediction is obtained for the loss of autonomy (*ADLinf5*) and osteoporosis (*osteoConf*) with a f1-score from 15 to 20% better than the baseline classifier when using the Bayesian network. At the opposite, for 3 risk factors (orthostatic hypotension (*newHypoT*), Parkinson disease (*parkOuSP*) and other neurological disorders (*auTrNeur*), no classifier allows to improve the f1-score or the accuracy of more than 1% compared to the baseline classifier. From the point of view of our experts, our set of variables does not include enough details about the class of drugs that are known to be predictive of orthostatic hypotension. Likewise, about the variable *parkOuSP*, the expert states that it is not predictable from our set of variables.

5.3. Which classifier to use?

Among the 45 variables selected for this study, an arbitrary number of them can be observed, whether they are targets or not. Moreover, risk factors are not independent of each other, meaning that when one of them is observed, it should be used to improve the evaluation of the others, in addition with other observed features. That situation makes it very difficult to use of usual classifiers because a new model would have to be learned for each target variable, and for each possible subset of observed variables. BN models allow to overcome that problem, since the same model can be used to evaluate any variable of the model, regarding any subset of observations. In addition, BNs allow to combine general statistical knowledge and specific individual information, and to update belief on any node from incomplete observations. These features exactly answer to the problem of predicting risk factors in real life situations. Another advantage of BN is that the model can be built both from data and expert knowledge which is very interesting in the context of health. It is also very important to make the model interpretable and understandable by the final user (general practitioners) since it contributes to make the aiding system acceptable and augment the trust in results. So BN becomes a good choice to use because of the graphical representation that is easy to explain and understand.

5.4. Interpretation of Bayesian network's graph

Fig. 3 shows the most frequent links between the 45 variables of the 10 BNs learnt during the 10 fold-cross process. It only displays the arcs that appear in at least 8 graphs, without considering their direction. Since the BNs are learned from data, their graphs are not causal, and the same arc may belong to different BNs with opposite directions. For that reason, we focus on arcs without considering their direction in order to evaluate their meaning from a medical point of view. Following several interviews with two experts in fall prevention at Lille's Hospital, it appears that most of these links are explainable: each link is either between a cause and a consequence, or between associated factors that make sense in medical knowledge. For example, the number of drugs taken daily (*nbMed3*) is linked in the graph with variables that refer to medications (*newTrOst, diuretiq, gt1psych*) and with diseases that require medication (*cardiopathy, diabete, hypertension, pneumopathy*).

6. Conclusion and future works

We have presented a BN model for the evaluation of risk factors for fall on the basis of general statistical information (dataset), the knowledge about risk factors (ontology) and, partial observations for a given person.

The main lessons learned from this study are as follows:

1. It is better to use the complete set of variables to build the model instead of using the specific subset to predict the risk factors for fall because, as shown in this study, the results using complete set of variables are slightly better than the other. Also, in real life situations, the number of observations can be different for each patient. So for each new patient we have to learn a new model using the available information that can be very time consuming.
2. It is very important to have a deep understanding of the variables and their relationship for modeling so that the selection of variables can be done efficiently.
3. Despite the prediction of risk factors for fall remains a challenge, the prediction of the presence or the absence of risk factors for fall from incomplete information is possible; the size of the data set to learn the classifiers and the characteristics of the population used for the data set are very important.

This study has some limitations. First, we integrate the expert knowledge only at the level of the preprocessing of data, for feature selection, but not for the definition of the structure and the parameters of the BN. In future research, it would be interesting to take into account the knowledge of experts in order to provide a causal BN model. Second, we have a selection bias due to the fact that our data come from a population of people at high risk of falling. Thus, in future we will ensure that we have a larger database for our analysis. Finally, since the target risk factors for fall have unbalanced classes, we plan to test some methods to cope with that problem.

Recall that fall prevention requires to provide a small number of recommendations depending on the risk factors present for a person. Thus the evaluation of risk factors is the basis of fall prevention. But it appears that evaluation of risk factors is not an easy task to perform, neither by humans, (it may require time, expertise, specific tests and/or devices), nor automatically, as experimented in this study. Based on the results and discussion presented, we propose Bayesian networks as a possible solution to predict the presence or absence of the risk factors for a given person for prevention of fall using the information available about the person (whether complete or incomplete).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research has been supported by the Hauts-de-France region, the University of Mons, Belgium, the french Ministry of Higher Education and Research and the National Center for Scientific Research in France. Also, we are grateful to the service of fall prevention of Lille's hospital for their support.

Appendix A. Variable description

The list of 45 variables obtained from the steps described in Section 3 about data preprocessing and variable selection is shown in Table 3. The 12 target variables are shown in bold. The first variables are direct features of the person (age, sex and body mass index), together with a specific individual factor which is the number of falls during the last six months. The following 27 variables directly represent the main risk factors for fall identified in the ontology. The remaining 13 variables concern secondary risk factors for fall and associated variables, and are as follows: diabete (*diabete*), unipedal stance test greater than 5 s (*apUniGt5*), cardiac arrhythmia(*arythm*), cardiopathy(*cardiop*), drives her car (*conduit*), difficulty using the toilets (*difWC*), diuretic (*diuretiq*), avoids going out by fear of falling(*evitSort*), get up and go test greater than 20 s (*GUGOgt20*), high blood pressure (*HTA*), lives in a retirement home (*maisRet*), podiatric problem (*pbPodo*), pneumopathy (*pneumo*), goes out of his/her house (*sort*), tobacco (*tabac*). All the variables are binary (yes: 1, no: 0), except the variables *nbMed3* and *BMI4* (discretized in 3 or 4 intervals).

References

[1] A.S. Chiu, R.A. Jean, M. Fleming, K.Y. Pei, Recurrent falls among elderly patients and the impact of anticoagulation therapy, World J. Surg. 42 (12) (2018) 3932–3938.

[2] P.C. Dykes, D.L. Carroll, A. Hurley, S. Lipsitz, A. Benoit, F. Chang, S. Meltzer, R. Tsurikova, L. Zuyov, B. Middleton, Fall prevention in acute care hospitals: a randomized trial, JAMA 304 (17) (2010) 1912–1918.

[3] C.A. Pfortmueller, G. Lindner, A.K. Exadaktylos, Reducing fall risk in the elderly: risk factors and fall prevention, a systematic review, Minerva Med. 105 (4) (2014) 275–281.

[4] S. Turner, R. Kisser, W. Rogmans, Factsheet for falls among older adults in the EU-28, 2015, Available at https://eupha.org/repository/sections/ipsp/Factsheet-falls-in-older-adults-in-EU.pdf. Accessed: 2020-05-15.

[5] A. Kabeshova. Prédire la Chute de la Personne Âgée : Apports des Modèles Mathématiques non-Linéaires. Médecine Humaine et Pathologie, Université d'Angers, France, 2015. Ph.D. thesis.

[6] S. McLachlan, K. Dube, G.A. Hitman, N.E. Fenton, E. Kyrimi, Bayesian networks in healthcare: distribution by medical condition, Artif. Intell. Med. 107 (2020) 101912.

[7] K.-J. Wang, J.-L. Chen, K.-M. Wang, Medical expenditure estimation by Bayesian network for lung cancer patients at different severity stages, Comput. Biol. Med. 106 (2019) 97–105.

[8] Y.S. Delahoz, M.A. Labrador, Survey on fall detection and fall prevention using wearable and external sensors, Sensors 14 (10) (2014) 19806–19842.

[9] A. Ungar, M. Rafanelli, I. Iacomelli, M.A. Brunetti, A. Ceccofiglio, F. Tesi, N. Marchionni, Fall prevention in the elderly, Clin. Cases Miner. Bone Metab. 10 (2) (2013) 91.

[10] D. Heckerman, D. Geiger, D.M. Chickering, Learning Bayesian networks: the combination of knowledge and statistical data, Mach. Learn. 20 (3) (1995) 197–243.

[11] T.D. Nielsen, F.V. Jensen, Bayesian Networks and Decision Graphs, Springer Science & Business Media, 2007.

[12] A.C. Constantinou, N. Fenton, W. Marsh, L. Radlinski, From complex questionnaire and interviewing data to intelligent Bayesian network models for medical decision support, Artif. Intell. Med. 67 (2016) 75–93.

[13] A.C. Constantinou, N. Fenton, M. Neil, Integrating expert knowledge with data in Bayesian networks: preserving data-driven expectations when the expert variables remain unobserved, Expert Syst. Appl. 56 (2016) 197–208.

[14] M.J. Flores, A.E. Nicholson, A. Brunskill, K.B. Korb, S. Mascaro, Incorporating expert knowledge when learning Bayesian network structure: a medical case study, Artif. Intell. Med. 53 (3) (2011) 181–204.

[15] W. Liao, Q. Ji, Learning Bayesian network parameters under incomplete data with domain knowledge, Pattern Recognit. 42 (11) (2009) 3046–3056.

[16] A.R. Masegosa, S. Moral, An interactive approach for Bayesian network learning using domain/expert knowledge, Int. J. Approx. Reason. 54 (8) (2013) 1168–1181.

[17] H.S. Sousa, F. Prieto-Castrillo, J.C. Matos, J.M. Branco, P.B. Lourenço, Combination of expert decision and learned based Bayesian networks for multi-scale mechanical analysis of timber elements, Expert Syst. Appl. 93 (2018) 156–168.

[18] G. Sihag, V. Delcroix, E. Grislin, X. Siebert, S. Piechowiak, F. Puisieux, Prediction of risk factors for fall using Bayesian networks with partial health information. Globecom AIdSH Workshop, IEEE, 2020, pp. 1–6.

[19] J.J. Davis, A.J. Clark, Data preprocessing for anomaly based network intrusion detection: a review, Comput. Secur. 30 (6–7) (2011) 353–375.

[20] S.B. Kotsiantis, D. Kanellopoulos, P.E. Pintelas, Data preprocessing for supervised leaning, Int. J. Comput. Sci. 1 (2) (2006) 111–117.

[21] A. Famili, W.-M. Shen, R. Weber, E. Simoudis, Data preprocessing and intelligent data analysis, Intell. Data Anal. 1 (1) (1997) 3–23.

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

[23] W. Buntine, A guide to the literature on learning probabilistic networks from data, IEEE Trans. Knowl. Data Eng. 8 (2) (1996) 195–210.

[24] S. Beretta, M. Castelli, I. Gonçalves, R. Henriques, D. Ramazzotti, Learning the structure of Bayesian networks: aquantitative assessment of the effect of different algorithmic schemes, Complexity 2018 (2018) 1591878:1–1591878:12.

[25] M. Scutari, C.E. Graafland, J.M. Gutiérrez, Who learns better Bayesian network structures: accuracy and speed of structure learning algorithms, Int. J. Approx. Reason. 115 (2019) 235–253.

[26] B. Xi, Z. Liu, M. Raghavachari, C.H. Xia, L. Zhang, A smart hill-climbing algorithm for application server configuration. Proceedings of the 13th International Conference on World Wide Web, 2004, pp. 287–296.

[27] J.K. Kruschke, Bayesian estimation supersedes the *t* test, J. Exp. Psychol. 142 (2) (2013) 573.

[28] C. Gaxatte, E. Faraj, O. Lathuillerie, J. Salleron, V. Deramecourt, V. Pardessus, M. H. Destailleur, E. Boulanger, F. Puisieux, Alcohol and psychotropic drugs: risk factors for orthostatic hypotension in elderly fallers, J. Hum. Hypertens. 31 (4) (2017) 299–304.

[29] C. Gaxatte, T. Nguyen, F. Chourabi, J. Salleron, V. Pardessus, I. Delabrière, A. Thevenon, F. Puisieux, Fear of falling as seen in the multidisciplinary falls consultation, Ann. Phys. Rehabil. Med. 54 (4) (2011) 248–258.

[30] A. Ankan, A. Panda, pgmpy: probabilistic graphical models using python. Proceedings of the 14th Python in Science Conference (SCIPY 2015), 2015, pp. 6–11.