Gaining Insight into Determinants of Physical Activity using Bayesian Network Learning

Simone C.M.W Tummers[1], Arjen Hommersom[2,3], Lilian Lechner[1], Catherine Bolman[1], and Roger Bemelmans[4]

[1] Faculty of Psychology, Open University of the Netherlands, Heerlen, The Netherlands
[2] Faculty of Science, Open University of the Netherlands, Heerlen, The Netherlands
[3] Department of Computer Science, Radboud University, Nijmegen, The Netherlands
[4] Zuyd University of Applied Sciences, Heerlen, The Netherlands

Abstract. Bayesian network modelling is applied to health psychology data in order to obtain more insight into the determinants of physical activity. This preliminary study discusses some challenges to apply general machine learning methods to this application domain, and Bayesian networks in particular. We investigate suitable methods for dealing with missing data, and determine which method obtains good results in terms of fitting the data. Furthermore, we present the learnt Bayesian network model for this e-health intervention case study, and conclusions are drawn about determinants of physical activity behaviour change and how the intervention affects physical activity behaviour and its determinants. We also evaluate the contributions of Bayesian network analysis compared to traditional statistical analyses in this field. Finally, possible extensions on the performed analyses are proposed.

Keywords: Machine Learning · Bayesian Network · E-health Intervention · Structure Learning · Physical Activity

1 Introduction

Nowadays there are various e-health intervention platforms that employ integrated behaviour change techniques in order to change health-related-behaviour of participants, for example increasing physical activity. These interventions apply theoretical psychological methods to influence behavioural determinants, which are factors determining a certain behaviour. These general techniques are translated to behaviour change strategies by tailoring the theoretical method to the target population and intervention setting [1]. To measure the effects of such interventions, various research studies have been performed, assessing physical activity with tools such as questionnaires and activity trackers. While there is now a good understanding of what the most important determinants for increasing physical activity are, little is known about how these determinants interact. Improved understanding of these relationships could be used to improve existing e-health interventions.

Supervised machine learning techniques are used to identify relationships underlying data with labeled input and output, and predict output results for a given input. These techniques could for example be used to model relations between diseases and symptoms and give expectations about the presence of various diseases given symptoms. Bayesian networks [8] represent probabilistic relationships between a set of variables, where relationships between the input variables can also be investigated. Such networks can make probabilistic predictions and provide a visual insight in relations among all variables of interest, thereby providing a potential useful tool to better understand determinants of physical activity.

In this article, a Bayesian network model is learned from data from a single intervention study, i.e., the *Active Plus intervention* [12], aiming at influencing physical activity behaviour among older adults. We discuss ways to learn from these complex data containing a significant amount of missing values. Based on these initial findings, results from previous analyses are compared to results from applying the Bayesian network model to the same data, to examine the added value of this technique compared to traditional ones. We show that learning a Bayesian network model for measurement data from the Active Plus project indeed reveals conditional dependence and independence relations that provide new insights and explanations for previously found results.

This paper is organised as follows. Section 2 provides technical background about methods and algorithms. Section 3 provides a description of the data and intervention study at hand, and how the data has been pre-processed. Furthermore, the analysis based on the Bayesian network model is explained including a description of the applied learning strategy, and a missing data analysis to select appropriate methods for handling the missing data. Then, in Section 4, results are given about the comparison of evaluated methods, and the comparison of the results from the Bayesian network model, determined using the best method, and the results from previous analyses. Finally, Section 5 concludes this paper and elaborates on possible extensions.

2 Preliminaries

This section gives an overview of the theoretical background relevant to perform the case study analyses, including a brief introduction of the modelling approach.

2.1 Bayesian network model

A Bayesian network [8] is a probabilistic graphical model represented as a directed acyclic graph $G = (V, E)$, where the set of nodes V represent random variables, and the set of arcs E represent probabilistic independencies among the variables. Associated with each node is a conditional probability distribution of that variable given its parents. The graphical structure implies conditional independence statements. Let $V = \{X_1, \ldots, X_n\}$ be an enumeration of the nodes in a Bayesian network such that each node appears after its children,

and let Π_i be the set of parents of a node X_i. The local Markov property in the Bayesian network states that X_i is conditionally independent of all variables in $\{X_1, X_2, \ldots, X_{i-1}\}$ given Π_i for all $i \in \{1, \ldots, n\}$. These local independences imply conditional independence statements over arbitrary sets of variables.

The joint probability distribution over discrete variables follows from the conditional independence propositions and conditional probabilities:

$$\mathbb{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \mid X_1, \ldots, X_{i-1}) = \prod_{i=1}^{n} \mathbb{P}(X_i \mid \Pi_i),$$

where the first equation follows from the usual chain rule in probability theory and the second from the local Markov property. Note that the conditional probabilities $\mathbb{P}(X_i \mid \Pi_i)$ correspond to the arcs in the Bayesian network specification. In continuous Bayesian networks, usually a linear Gaussian distribution is assumed, where the joint density is factorised where each $X_i \mid \Pi_i \sim \mathcal{N}(\beta\Pi_i + \alpha, \sigma^2)$.

A temporal Bayesian network is an extension to the static counterpart in that it is a Bayesian network model over time, where the nodes represent the random variables occurring at particular time slices. The temporal Bayesian network model is subject to the condition that arcs directed to variables in previous time slices cannot occur. In case the temporal Bayesian network is time-homogeneous (or time-invariant), these models are also called dynamic Bayesian networks [6]. Since in this case study there are only a few time slices and differences between these slices are not constant, we do not assume time-invariance in the remainder of this paper.

2.2 Learning Bayesian networks

The following three common classes of algorithms are used to learn the structure of Bayesian networks from the data: constraint-based algorithms which employ conditional independence tests to learn the dependence structure of the data, score-based algorithms which use search algorithms to find a graph that maximises a goodness-of-fit scores as objective function, and hybrid algorithms which combine both approaches. Recent research has shown that constraint-based algorithms are often less accurate and seldom faster and hybrid algorithms are neither faster nor more accurate [11]. For this reason, we focus in the remainder of this paper on score-based structure learning algorithms, where local search methods are used to explore the space of directed acyclic graphs by single-arc addition, removal and reversal. In particular, we apply tabu search to the physical activity data in this case study as empirical evidence shows that this search method typically performs well for learning Bayesian networks [5, chapter 13.7].

There are several model selection criteria that are used in the search-based structure learning algorithms, where in this paper we have chosen the commonly-used Bayesian Information Criterion (BIC) [9]. To fit the parameters we have chosen a uniform prior distribution over the model parameters [4].

Algorithm 1 Structural EM algorithm, given $(M_0, \mathbf{o})$:

for $n = 0, 1, \ldots$ until convergence or predefined maximum number of iterations reached **do**
  Compute Θ^{M_n} using a parameter learning algorithm.
  Expectation-step:
  compute $\mathbf{h}^* = \arg\max_{\mathbf{h}} \mathbb{P}(\mathbf{h} \mid \mathbf{o}, M_n)$
  Maximization-step: apply structure learning to determine M_n using data $\mathbf{h}^* \cup \mathbf{o}$
  if $M_n = M_{n+1}$ or if stopping criterion is met **then**
    return M_n
  end if
end for

2.3 Handling missing data

Learning Bayesian networks with missing data is significantly harder as the log-likelihood does not admit a closed-form solution if values are missing. In this paper, we assume that data are missing at random, for which commonly used methods are listwise deletion, pair-wise deletion, single imputation, multiple imputation [7]. The deletion approaches omit (observed) values from analyses. In the listwise deletion approach on the one hand, all observations with missing values at any measurement are omitted completely. On the other hand, the pair-wise deletion method does not require complete data on all variables in the model, and mean and covariance estimations are here based on the full number of observations with complete data for each (pair of) variable(s). Imputation methods involve replacing missing values by estimates such as by the mean of observed values in the attribute, called mean imputation. Single imputation imputes a single value treating it as known, whereas multiple imputation replaces missing values by two or more values representing a distribution of possibilities. In multiple imputation, missing data are filled in an arbitrary number of times to generate different complete datasets to be analysed, and results are combined for inference. Finally, in Bayesian network learning, the Expectation Maximization (EM) algorithm [2] is often applied, which iteratively optimises parameters in order to find the maximum likelihood estimate, assuming the missing data is missing at random (MAR). The Structural EM algorithm (SEM) [3] combines this standard EM algorithm with structure search for model selection.

The variant of the structural EM algorithm that is used in this case study can be described as follows (see Algorithm 1 for an overview). Let $\mathbf{d}$ be a dataset over the set of random variables $\mathbf{V}$. Assume that $\mathbf{o}$ is part of the dataset that is actually observed, i.e., $\mathbf{o} \subseteq \mathbf{d}$. Furthermore, we denote the missing data by $\mathbf{h}$, i.e., $\mathbf{d} = \mathbf{o} \cup \mathbf{h}$, and $\mathbf{o} \cap \mathbf{h} = \varnothing$. The SEM algorithm aims to find a model from the space of Bayesian network models over $\mathbf{V}$, denoted by $\mathcal{M}$, such that each model $M \in \mathcal{M}$ is parametrised by a vector Θ^M defining a probability distribution $\mathbb{P}(\mathbf{V} : M, \Theta^M)$. To find a model in case of missing values, the complete data likelihood $\mathbb{P}(\mathbf{H}, \mathbf{O} \mid M)$ is estimated. The algorithm iteratively maximises the expected Bayesian network model score optimised by the score-based algo-

rithm. First the posterior parameter distributions, given the currently best model structure and observed data, are computed. In the expectation step, these distributions are used to compute the expected complete dataset, imputing missing values with their most probable values, also sometimes called *hard EM*. During the maximization, the currently best model structure is updated using a tabu structure learning algorithm, using the imputed data from the expectation step. Then parameter learning gives new distributions to be used as input for the next expectation step. To perform the first expectation, an initial network structure is given as input to the algorithm. In case a maximum number of iterations is reached or in case of convergence, the Bayesian network model is returned.

3 Description of the Data and Methodology

The experiments in this intervention case study aim to analyse performance of different methods to handle missing values and to learn the Bayesian network model for given intervention data in order to compare its results to previous analyses. This section describes the data, preprocessing phase, magnitude of the missing data problem and the approach to determine a suitable method in order to analyse the data by Bayesian network learning. The raw research data that has been collected during the Active Plus intervention was provided to the authors and is described in the first subsection.

3.1 Data acquisition and description

The raw research data has mostly been collected via questionnaires and consists of determinants, external factors, measurements of physical activity and intervention-related information at different time-slots, starting with a baseline measurement before the participant receives the intervention [14]. For example, the validated self-administrated Dutch Short Questionnaire to Assess Health Enhancing Physical Activity (SQUASH) is included in the questionnaires as subjective measurement of physical activity [15]. Figure 1 illustrates the intervention outline including moments of receiving intervention content and of measurement in time [12]. There is a distinction between control, intervention basic and intervention-plus groups, representing the intervention condition. This condition determines whether a participant receives an intervention or not and if environmental content is included in the intervention with additional information such as opportunities to be physically active in the own environment. Within these main groups, content is further personalised based on characteristics of participants, for example state of behaviour change (stage) measured at baseline or age. Since in the analyses in this article intervention content is proxied by a few main characteristics, this personalisation is beyond the focus of this article [12].

As depicted in Figure 1, data has been collected at 4 time-slots; at the baseline (before receiving the intervention, T0) and, to measure intervention effects, 3 (T1), 6 (T2) and 12 (T3) months after the baseline. About 1258 variables have

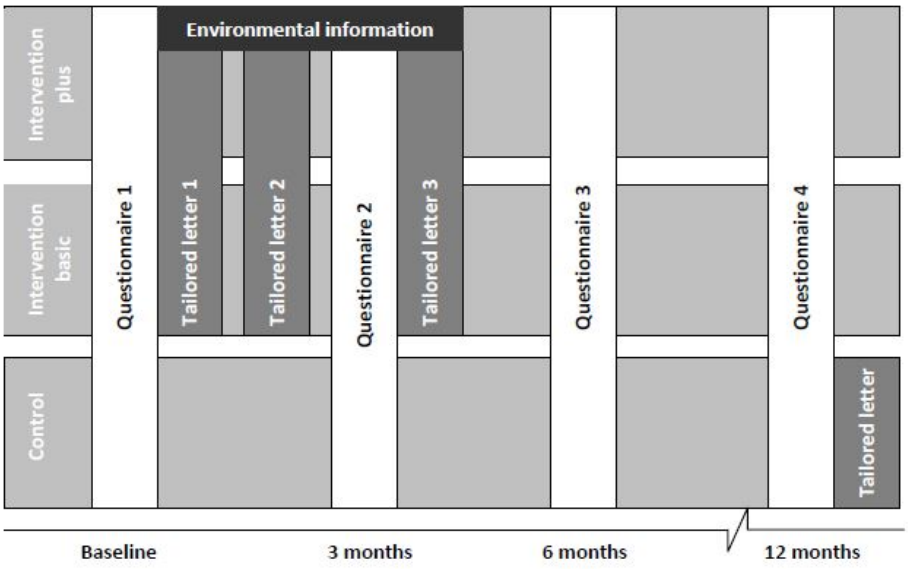

Fig. 1: Outline intervention program including moments of measurement [12].

been measured for a sub-population being a random sample of 1976 adults aged 50 and older. Measurements are at item-level of detail, where an item is a specific measurement, for example a question in the questionnaire. In preprocessing rules, it is described how concepts are calculated from item data in order to perform analyses at a higher level of abstraction.

3.2 Data preprocessing and concept design

The raw data is preprocessed, according to rules to integrate data from different studies and to aggregate, by calculating concepts from the raw data at item-level of detail, as mentioned in previous subsection. This subsection describes assumptions and decisions made during data preprocessing phase and rules to calculate the concepts included in analyses in this article.

In general, concepts are calculated by the mean or sum of items taking into account a maximum percentage of items allowed to be missing, except from a few concepts calculated using predefined formulas. In particular, the SQUASH-outcome measure, which is the number of minutes per week of moderate to intensive physical activity, is calculated in a standardised way [12]. In case more than 25 percent of the items are missing, the concept value is assumed to be missing. Besides these aggregation rules, preprocessing rules contain decisions about recalculation of raw data values to unipolar scale.

This article focuses on a selection of the data measured in the Active Plus intervention and, as already mentioned, analyses are performed at concept-level. The selection consists of data about the main determinants of physical activity

Concept	Number of items	T0	T1	T2	T3
Condition: intervention	1	X			
Condition: environment	1	X			
SQUASH outcome measure	-	X	X	X	X
Self-efficacy	10	X	X		
Attitude(-pros)	9	X	X		
Attitude(-cons)	7	X	X		
Intrinsic motivation	6	X	X		
Intention	3	X	X	X	X
Commitment	3	X	X	X	
Strategic planning	10	X	X	X	X
Action planning	6	X	X	X	
Coping planning	5	X	X	X	
Habit	12	X		X	X
Social modelling	1	X	X	X	
Social support	1	X	X		

Table 1: Overview of concept-level variables included in case study.

behaviour, including some social-related determinants, the main outcome measure from the SQUASH questionnaire and some variables indicating the intervention content the participant receives. As described, the intervention content that an individual participant has received is personalised and proxied in the analyses. The proxy of the intervention content is represented in the data by intervention condition variables, which thus play a central role in analyses. Table 1 gives an overview of these and all other concepts included in this articles analyses, indicating the number of item-level variables the concept variable aggregates and at which moments in time the concept is measured. Note that the number of items for the SQUASH outcome measure is not indicated since it is calculated by standard rules.

3.3 Missing data analysis

A significant part of this case study consists of the evaluation of ways to handle missing data values. This subsection illustrates the magnitude of the missing data problem in the case study and determines which methods are appropriate to be evaluated.

A total of 39 variables being concepts at certain moments in time are selected as subset for analyses. Table 2 demonstrates the number of missing values out of 1976 observations for each of the included concept-level variable. Since the time dimension is crucial to analyse intervention effects and, as can be seen in Table 2, more than a fourth of the values are missing for measurements after the baseline, applying pairwise deletion would result in an immense loss of information. Furthermore, the number of complete observations is for the selection of concepts 360 out of 1976 in total, meaning that applying list-wise deletion would neglect a large part of the dataset. Since deletion methods are not appropriate to deal with the missing data in this case study, we resort to the remaining methods for dealing with missing data, i.e., mean imputation and the SEM algorithm described in Section 2.3, are applied and results are compared.

Concept	Timeslot	Number of missing values (out of 1976)
Condition: intervention	T0	8
Condition: environment	T0	8
SQUASH outcome measure	T0	3
	T1	518
	T2	565
	T3	628
Self-efficacy	T0	229
	T1	638
Attitude(-pros)	T0	149
	T1	587
Attitude(-cons)	T0	167
	T1	597
Intrinsic motivation	T0	325
	T1	690
Intention	T0	141
	T1	571
	T2	654
	T3	748
Commitment	T0	31
	T1	531
	T2	573
Strategic planning	T0	156
	T1	601
	T2	652
	T3	661
Action planning	T0	182
	T1	604
	T2	686
Coping planning	T0	192
	T1	621
	T2	668
Habit	T0	136
	T2	633
	T3	662
Social modelling	T0	532
	T1	915
	T2	952
Social support	T0	68
	T1	561

Table 2: Overview of number of missing values in included concepts.

3.4 Approach

This subsection discusses how a suitable method for handling missing data is determined in order to model the intervention data. To perform experiments,

Handling missing data	Mean log-likelihood	95% Confidence Interval
Mean imputation	-4779	[-4832;-4726]
SEM algorithm	-4127	[-4183;-4071]

Table 3: Results of cross-validation analysis for missing data methods.

the bnlearn package in R is used for Bayesian network learning [10]. Source code has been made publicly available[5].

In the comparison of the methods to handle missing data values evaluated in this article, we apply discrete dynamic Bayesian networks for preprocessed data that is discretised by manually creating intervals meaningful in the health psychology field. The models are learnt by the tabu search algorithm optimising the BIC score (see Section 2.2). In the intervention study at hand only system missing values occur, for example, in case a participant has not answered a specific question in the questionnaire or if the maximum amount of items allowed to be missing is exceeded. The methods evaluated both apply imputation where missing values are substituted by (maximum likelihood) estimators during the structure learning phase, namely mean imputation and the structural EM algorithm, introduced in Section 3.3. These two methods are compared by means of comparing the mean test-set log-likelihood using k-fold cross-validation (with $k = 10$).

Finally, a linear Gaussian temporal Bayesian network model for the Active Plus intervention data is constructed from the preprocessed selection of data by learning the network structure using SEM. It was chosen to learn a continuous network rather than a discrete one to prevent possible loss of information from the discretisation process. In order to evaluate significance of edges, a bootstrap analysis is applied. Edges that are identified in most bootstrap samples and in the original network are considered stable findings in the following.

4 Results

This section describes the performance comparison of the methods applied to handle missing values. Furthermore, the learnt Bayesian network to model the Active Plus data is presented and results are compared to previous analyses of relations between determinants in the study by Van Stralen et al. [13].

4.1 Comparison Bayesian network missing data strategy

Table 3 demonstrates the mean log-likelihood over the folds resulting from applying the implemented cross-validation algorithm to the selected methods for handling missing data.

The cross-validation analysis shows that the structural EM algorithm significantly outperforms mean imputation to handle missing data, since the mean

[5] `https://github.com/SCMWTUM/Active4life-datascience.git`

Model	Statistics	
Optimal Bayesian network	#nodes	39
	# arcs	188
	# undirected arcs	0
	Average markov blanket size	19.90
	Average neighbourhood size	9.64
	Average branching factor	4.82
Averaged Bayesian network	#nodes	39
	# arcs	170
	# undirected arcs	0
	Average markov blanket size	17.54
	Average neighbourhood size	8.72
	Average branching factor	4.36

Table 4: Statistics Bayesian network model versus averaged counterpart.

log-likelihoods over the folds significantly differ at 5% confidence level. In the next subsection, the learnt model is presented and results are compared to those from previous analyses.

4.2 Comparison of Bayesian network model to previous analyses

Figure 2 shows the union of the temporal Bayesian network model learnt by the tabu search algorithm, applying SEM and optimising BIC score, and the result of bootstrapping (which we call *averaged model*). A comparison of these models shows that only 149 edges appear in both models represented by black edges in Figure 2, 21 only in the averaged model represented by red edges, and 39 only in the optimal model learnt from the data represented by blue dashed edges. Table 4 gives the summary statistics of the temporal Bayesian network model learnt and its averaged counterpart and indicates that model complexity is decreased in the averaged model. This suggests that most edges are stable, but not in all cases. Quite some edges appear to be unstable, which is something that should be analysed further in future.

Compared to previous analyses, the Bayesian network model provides a more complete insight in the complexity of mechanisms influencing physical activity behaviour. Previously, mediation analyses have shown that factors such as social modelling, self-efficacy and intention are significant mediators of the intervention influencing physical activity behaviour. In Figure 3, a fragment of the stable part of the averaged model (Figure 2) is shown that includes these previously proven significant determinants, intervention effects, and effects on physical activity. It also includes coefficients, which represent the maximum likelihood estimators of parameters of the Gaussian conditional density distribution of variables given their parents. This part of the network suggests that intervention effect on physical activity levels is mainly mediated by influencing habit and intention, and the extension in which environmental components are added to the intervention does

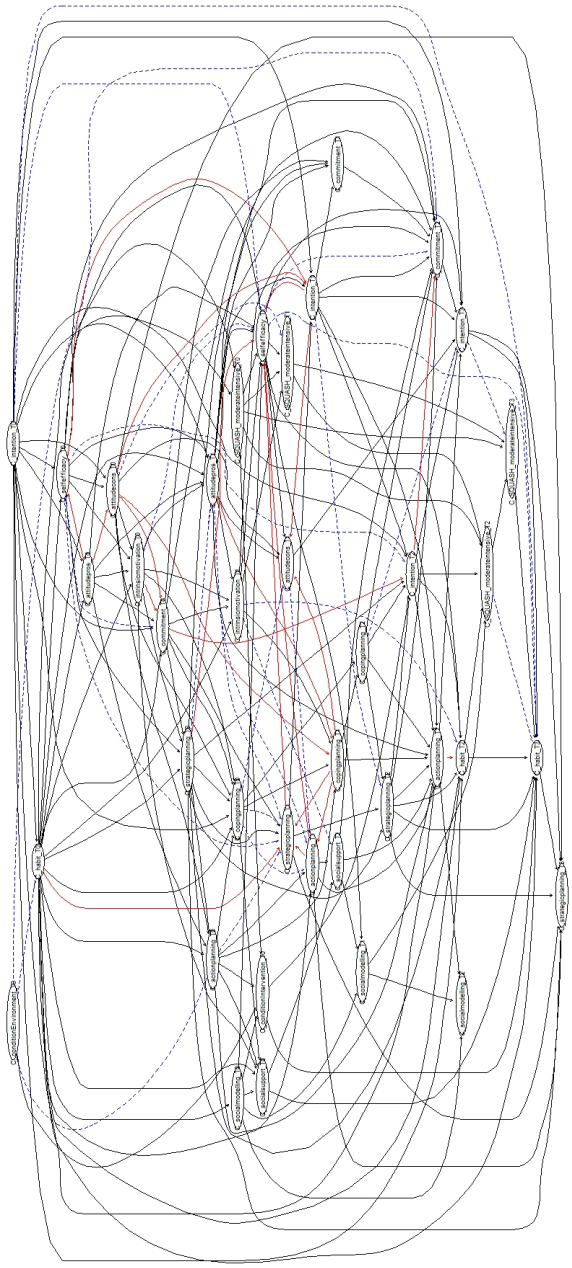

Fig. 2: Averaged model learnt from bootstrapping, which includes the false positives (in blue) and false negatives (in red) from the model learnt for the original dataset.

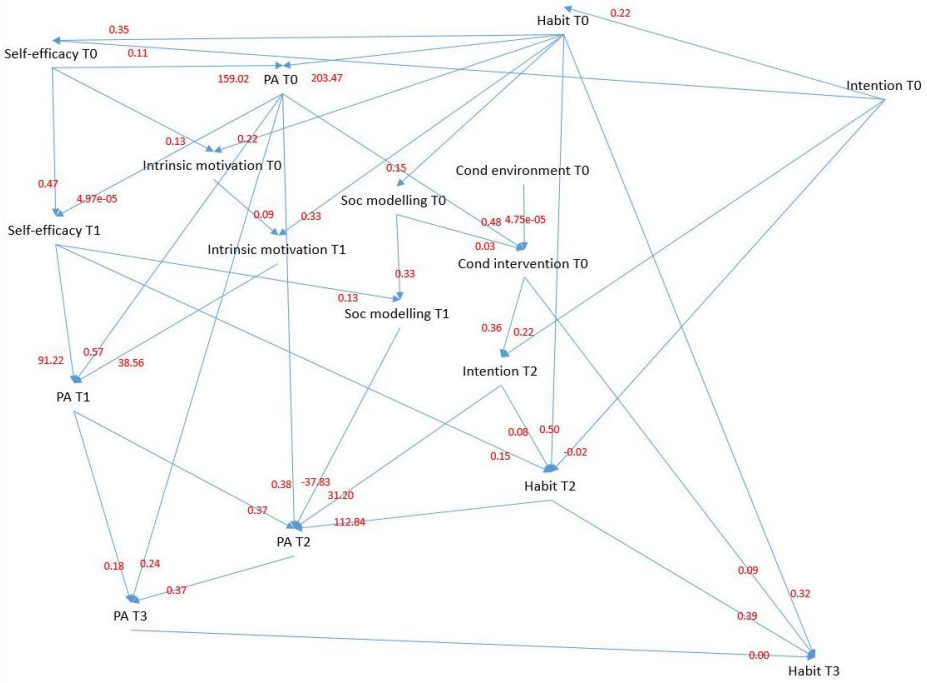

Fig. 3: Selected subgraph of the averaged model.

not significantly influence physical activity nor its determinants. Furthermore, there is a distinction between determinants of physical activity in the short (T1) and in the long (T2 and T3) run. In the short run, effects on physical activity are mainly determined by self-efficacy and intrinsic-motivation, which mediates effects of habit and self-efficacy. In the long-run, social modelling, intention and habit are important, where habit has the strongest correlation with long run physical activity levels.

Looking at intervention effect analysis, comparing these results to previous results by [13], significant influence on social modelling and self-efficacy in the short run is not demonstrated in the network. Looking at mediator effects on physical activity, [13] has not found that intrinsic motivation is relevant in the short run, whereas in the Bayesian network this determinant does have a significant direct influence on physical activity level in the short run. Also, where in previous analyses results show significant influence of the environmental extension on physical activity and determinant levels, this relationship is not found in the Bayesian network model. All in all, the added value of applying the Bayesian network model compared to traditional analyses is that the model provides new in-depth information relevant for understanding the working mechanisms of the intervention. For example, intrinsic motivation might mediate effects of other determinants found in [13], such as attitude-cons, on physical activity in the

short run, which might explain that the Bayesian network model leads to the difference compared to classic mediator analyses that has been found.

To conclude this section, differences found between previous results and results from the Bayesian network model are explored and possible explanations are provided. First of all, previously-found important mediators of intervention effects on physical activity such as social modelling, self-efficacy and intention are confirmed by the network model. In the network, habit is also a significant mediator. [13] did not include this determinant in analyses, so no comparison can be made with respect to habit being a significant mediator of intervention effects on physical activity. An important difference is that [13] found differences between effects in groups of participants having received environmental content and those who did not receive this extension. In the Bayesian network, those differences are not found. However, taking into account uncertain edges, there are some interesting correlations of the environmental extension with, for example, commitment at T2. Further analyses could explore these relations in order to explain the differences. There are also differences found with respect to intervention effects on determinants and mediation effects on physical activity. [13] has not found intrinsic motivation being a significant mediator, whereas the network model shows that the effect of self-efficacy on physical activity is both direct and mediated by intrinsic motivation. The network explores the mechanism in which self-efficacy influences physical activity, so that intrinsic motivation emerges as mediator. In the network, the intervention does not have a direct effect on social modelling nor on self-efficacy. This can be explained by looking at the whole model, where for example the intervention influences intention, which is correlated with action planning that is again correlated with social modelling. In this way, some determinants previously-found to be influenced by the intervention directly, are indicated in the network to be influenced via other determinants. Hence, the network provides a more in-depth view in the dependencies and the structure in which determinants and physical activity are influenced by the intervention.

5 Discussion and Conclusions

In this article, the Bayesian network modelling technique has been applied to an e-health intervention case study to potentially better understand relations between determinants of physical activity, since this technique has not been applied often in this field and traditional analyses are not sufficient to reveal the dependence structure between determinants. The magnitude of the major challenge of missing values in performing machine learning in real-world studies in general is examined for this case study and is shown to be of such an order that conventional methods to handle missing values cannot be used. The performance of different methods to handle missing data in Bayesian network modelling (i.e. mean imputation and the structural EM algorithm), considered to be appropriate in this case study, has been evaluated. Although the comparison between the mean imputation and structural EM method is not very novel from a machine

learning point of view, it has been carried out to evaluate their performances in this specific context. Also, since this modelling technique has not yet often been applied in this research field, its added value compared to more classic analyses in health psychology is evaluated by learning a Bayesian network for the case study and comparing its results to those of previous analyses on the same data.

Analysis of missing data in the case study dataset demonstrates that the magnitude of this problem causes methods to handle missing data based on deletion to be inappropriate, since this would result in a significant loss of information for this type of data. Two suitable methods, i.e., mean imputation and the structural EM algorithm, have been compared and we show that applying the structural EM algorithm leads to the best results in terms of fitting the data when learning a Bayesian network model for intervention data. The model learnt for the case study data applying this algorithm to handle missing values, suggests that the intervention does influence physical activity behaviour, that some concepts do not play a direct role influencing this behaviour or are not directly influenced significantly by the intervention and, most importantly, that there is some structure of how determinants explain this behaviour. Furthermore, there is some room for improvement with respect to increase confidence in some relationships in the model. Focusing on significant edges in a submodel, some differences regarding significant direct correlations are found compared to previous analyses. In brief, it can be concluded that applying Bayesian networks to e-health intervention study data provides more insight in the complexity of how interventions cause behavioural change (physical activity) and therewith are a useful technique to better understand dependence mechanisms of determinants of behaviour change.

In future work, analyses in this article could be extended for example by evaluating other imputation methods to be implemented in the structural EM algorithm, such as a distribution over values instead of imputing the value with highest probability (*soft EM*). From a technical perspective, we will also consider exploring constraint-based structure learning algorithms, other score-based algorithms, alternative parameter learning algorithms or alternative model selection criteria. From the application perspective, future research could further elaborate on the structure, in which determinants are related to each other and physical activity, and on the differences found in the Bayesian network model compared to previous (regression) analyses. Also, it would be interesting to perform analyses in more detail by using item variables in order to clarify the correlations between concepts found in the learnt network model presented in this paper. Finally, a combined model could be designed for an integrated dataset including measurements from several different e-health intervention studies, on different sub-populations, in order to examine if the general model yields different or additional results compared to the submodels for a smaller amount of data from single studies. However, even with data from a single study, this paper shows that exploring the differences between results from previous analyses and from the Bayesian network model, the network provides a more complete and in-depth insight in dependency structures. More specifically, the network reveals

relations between variables where a variable influences another via a third one. In previous analyses, only some of the hypothetical mediator effects are explored by regression analyses. Hence, our results provide new opportunities to analyse and confirm our findings using traditional statistical methods.

Acknowledgements This work is part of the research programme Active4Life with project number 546003005, which is financed by ZonMw.

References

1. Brug, J., van Assema, P., Lechner, L.: Gezondheidsvoorlichting en gedragsverandering Een planmatige aanpak. Koninklijke Van Gorcum, Assen, 9th edn. (2017)
2. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological) **39**(1), 1–22 (1977)
3. Friedman, N.: The Bayesian structural EM algorithm. In: Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence. pp. 129–138 (1998)
4. Ji, Z., Xia, Q., Meng, G.: A review of parameter learning methods in Bayesian network. Advanced Intelligent Computing Theories and Applications **9227**, 9–12 (2015)
5. Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. MIT press (2009)
6. Murphy, K.: Dynamic Bayesian Networks: Representation, Inference and Learning. Ph.D. thesis, UC Berkeley (2002)
7. Nakai, M., Ke, W.: Review of the methods for handling missing data in longitudinal data analysis. International Journal of Mathematical Analysis **5**(1), 1–13 (2011)
8. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo, CA (1988)
9. Schwarz, G., et al.: Estimating the dimension of a model. The annals of statistics **6**(2), 461–464 (1978)
10. Scutari, M.: Package 'bnlearn'. Bayesian network structure learning, parameter learning and inference, R package version 4.4 **1** (2019)
11. Scutari, M., Graafland, C.E., Gutiérrez, J.M.: Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms. International Journal of Approximate Reasoning **115**, 235–253 (December 2019)
12. van Stralen, M.M., Kok, G., de Vries, H., Mudde, A.N., Bolman, C., Lechner, L.: The active plus protocol: systematic development of two tailored physical activity interventions for older adults. BMC Public Health **8** (2008)
13. van Stralen, M.M., de Vries, H., Bolman, C., Mudde, A.N., Lechner, L.: Exploring the efficacy and moderators of two computer-tailored physical activity interventions for older adults: a randomized controlled trial. Annals of Behavioral Medicine **39**(2), 139–150 (2010)
14. van Stralen, M.M., de Vries, H., Mudde, A.N., Bolman, C., Lechner, L.: Determinants of initiation and maintenance of physical activity among older adults: a literature review. Health Psychology Review **3**, 147–207 (2009)
15. Wendel-Vos, G., Schuit, A., Saris, W., Kromhout, D.: Reproducibility and relative validity of the short questionnaire to assess health-enhancing physical activity. J Clin Epidemiol **56**, 1163–1169 (2003)