Lappenschaar, M., Hommersom, A., Lucas, P.J.F., Lagro, J., Visscher, S., Korevaar, J.C., Schellevis, F.G. Multilevel temporal Bayesian networks can model longitudinal change in multimorbidity. Journal of Clinical Epidemiology: 2013, 66(12), 1405-1416

http://www.jclinepi.com/article/S0895-4356(13)00265-5/abstract

Multilevel temporal Bayesian Networks can Model Longitudinal Change in Multimorbidity

Martijn Lappenschaar[a,*], Arjen Hommersom[a], Peter J.F. Lucas[a], Joep Lagro[b], Stefan Visscher[c], Joke C. Korevaar[c], François G. Schellevis[c,d]

[a] *Institute for Computing and Information Sciences, Radboud University Nijmegen, Nijmegen, The Netherlands*
[b] *Department of Geriatric Medicine, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands*
[c] *Netherlands Institute for Health Services Research (NIVEL), Utrecht, The Netherlands*
[d] *Department of General Practice and Elderly Care Medicine, EMGO Institute for Health and Care Research, VU University Medical Centre, Amsterdam, The Netherlands*

Abstract

Objective: Whereas the course of single diseases can be studied using traditional epidemiological techniques, these methods cannot capture the complex joint evolutionary course of multiple disorders. In this study, multilevel temporal Bayesian networks were adopted to study the course of multimorbidity in the expectation that this would yield new clinical insight. Study Design and Setting: Clinical data of patients were extracted from ninety general practice registries in the Netherlands. One and half million patient years were used for analysis. The simultaneous progression of six chronic cardiovascular conditions was investigated, correcting for both patient and practice-related variables. Results: Cumulative incidence rates of one or more new morbidities rapidly increase with the number of morbidities present at baseline, ranging up to 47% and 76% for three and five-year follow-up respectively. Hypertension and lipid disorders, as health risk factors, increase the cumulative incidence rates of both individual and multiple disorders. Moreover, in their presence, the *observed* cumulative incidence rates of combinations of cardiovascular disorders, i.e., multimorbidity, differ significantly from the *expected* rates. Conclusion: There are clear synergies between health risks and chronic diseases when multimorbidity within a patient progresses over time. The method used here supports a more comprehensive analysis of such synergies compared to what can be obtained by traditional statistics.

Keywords: multimorbidity, inter-practice variation, synergy, multilevel analysis, Bayesian networks, cardiovascular disease

*Corresponding author
Email address: mlappens@cs.ru.nl (Martijn Lappenschaar)

What is new?

Key findings:

- The urbanisation level of a general practice is associated with the cumulative incidence of chronic cardiovascular conditions, in particular those with a high prevalence, i.e., obesity, hypertension, dyslipidemia, diabetes mellitus, and ischemic heart disease.
- The overall multimorbidity rate of chronic cardiovascular (related) disorders rapidly increases when multimorbidity is already present at baseline.
- When multimorbidity progresses over time, certain disease combinations develop more quickly than what can be expected from individual disease progression. This synergistic effect happens particularly in the presence of hypertension and dyslipidemia.

What this adds to what is known:

- Multimobidity is not only about pairs of diseases, but also about how multiple diseases in patients interact and how this interaction changes over time. For the first time this proper perspective on multimorbidity is described and analysed using the new technique of multilevel temporal Bayesian networks. This new method not only supports finding multiple associations and how these change over time, but also which of these represent a direct association or a confounder, and how factors indirectly influence each other.

Implications:

- Whereas standard multilevel regression methods are very useful to explain a single disease with respect to a set of patient and practice related observable variables, multilevel Bayesian networks allow exploring the joint distribution of multiple diseases and their interactions, which is highly relevant in multimorbidity research.
- Clinical guidelines for patients with multimorbidity can be improved when the advice incorporates all the patient's specific characteristics. Since the network in the methodology used here can be personalised for a specific patient, it provides a valuable tool for the development of such tailored clinical guidelines.

Introduction

Recent epidemiological research indicates that up to two third of all patients older than 65 years in the western world have more than one chronic disorders at the same time. This is referred to as *comorbidity* or *multimorbidity*. Whereas comorbidity is usually defined in relationship to a specific index condition, as in the seminal definition of Feinstein [1], the term 'multimorbidity' has been introduced to refer to any co-occurrence of two or more medical, especially chronic, conditions within a person at the same time [2, 3]. Multimorbidity can be simply measured by computing various basic statistics: the number of chronic disorders per patient, corrected for age, gender, and socio-economical demographics [4–7], odds that describe the ratio between observed and expected prevalence rates for specific disease combinations [8–11], and disease clusters using principal component analysis [12–15].

Although systematic reviews [16–18] have given insight into the rates of cross-sectional co-occurrence, the progression over time of interactions between chronic cardiovascular diseases and related disorders is sparsely documented [19]. More insight into such interactions would help in personalizing the therapeutic management of patients with multimorbidity. As clinical knowledge is mostly organized around single diseases, that knowledge may not be fully applicable to patients with multimorbidity [20, 21]. The care of patients with multimorbidity can be improved by any method that tailors the advice to the patient's specific characteristics [22].

Recent explorations of patient data from primary care registries to quantify associations between chronic disorders, have shown these to be valuable for obtaining a broad picture of multimorbidity [21, 23, 24]. In this paper we will use such registries to assess three aspects of the joint progression of chronic cardiovascular multimorbidity: 1) its dependency on the practice's urbanity, 2) the synergistic effects between disorders when they evolve over time, and 3) the progression of the overall multimorbidity rate.

Patient data in primary care registries are often clustered by practices, introducing particular biases in the patient's diagnosis due to practice related effects. For example, the urbanity of the practice's area or the physician's experience. Multilevel regression analysis is the standard method of choice in these situations [25]. However, it does not allow analysing multiple disease outcomes simultaneously. Therefore, we adopted the method of *multilevel Bayesian networks* (MBNs) that does offer such support [26]. When used for the analysis of temporal data, the advantage of an MBN is that the disorders and their interaction are all treated as uncertain. The representation goes beyond showing how pairs of disorders are associated to each other. Furthermore, we can extend an MBN to analyse multiple outcomes at multiple time-points. The latter gives rise to *multilevel temporal Bayesian networks*, or MTBN for short.

In summary, we developed a multimorbidity model that yields a much better insight into interactions, progression over time, and the accumulation of chronic disorders than existing statistical

models are able to provide. Moreover, we show that posterior probabilities computed from the model at follow-ups can be tailored to any set of conditions present at baseline, which can provide valuable input for *personalised* clinical decision-making.

Methods

Data collection

The data used for analysis were obtained from the Netherlands Information Network of General Practice (LINH). All Dutch inhabitants are obligatory registered with a general practice, and the LINH registry contains information of routinely recorded data from about all patients of approximately 90 general practices. Longitudinal data of approximately one and half million patient years, covering the decade 2002-2011, from patients aged over 35 years, were used in our analysis. Patient data is available for the whole time frame, unless patient moved out of the practice or the practice itself opted out.

We used the definition of a 'chronic disorder' given by O'Halloran [27], which in turn was based on the international classification of primary care (ICPC) codes. Principally, our focus is on chronic cardiovascular diseases and related disorders, and in our model we included the following chronic disorders: obesity, hypertension, lipid disorder, diabetes mellitus, heart failure, stroke, ischemic heart disease, retinopathy, and nephropathy. The first three disorders are seen as health risks. Previous research has indicated that also some *non*-cardiovascular comorbidity is associated with cardiovascular disorders [21, 28]. Therefore, the diagnoses of other chronic non-cardiovascular disorders were modelled as well, but only as a single variable.

Lab results and medication were not always consistent with the diagnoses present in the LINH database. For example, insulin was sometimes prescribed to patient who were not diagnosed with diabetes mellitus according to the data. To compensate for such missing or incorrect information, we used lab results and medication to infer the diagnosis of obesity, hypertension, dyslipidemia and diabetes mellitus, in conjunction to the ICPC codes. This correction method has some limitations. For example, although statins are typically prescribed for lipid disorders, one cannot conclude that any patient who uses this drug has a lipid disorder [29]. Similarly, blood pressure lowering medication is sometimes prescribed for other cardiovascular disorders than hypertension.

Corrections were made by using the following rules adopted from the Dutch guideline on cardiovascular risk management [30], which are in line with the European guidelines on cardiovascular risk management in clinical practice [29]. For obesity: a body mass index over 30 kg/m^2; for hypertension: a high blood pressure (systolic $>$ 140 mm Hg or diastolic $>$ 90 mm Hg) within at least two recurring measurements; for dyslipidemia: an abnormal blood lipid profile (low density lipoprotein $>$ 3 mmol/l, high density lipoprotein $<$ 1 mmol/l, or triglycerides $>$ 2 mmol/l); and

for diabetes mellitus: a fasting glucose ≥ 6 mmol/l, or a prescription of either insulin or oral blood glucose lowering medication. See web materials for more details on data collection and the ICPC codes used for identification of the disorders used in the model.

Statistical analyses

As the patient data in the LINH dataset were obtained from several general practices, differences among those general practices may have a confounding effect on the probability distributions. Taking into account the hierarchical structure into statistical models demands for a multilevel approach. We used MBNs, i.e., multilevel Bayesian networks, in our analysis [26]. Bayesian networks provide a powerful framework for the representation of knowledge and reasoning under uncertainty [31], and they have had a significant impact on the modelling and the analysis of medical data [32]. The statistical relationships in such models can be learned from patient data. Recently, it was shown that when applying MBNs to a set of hierarchically structured disease variables, the outcome of multiple diseases can be very well predicted using a MBN [26]. With a receiver operating characteristic (ROC) curve it was demonstrated that a single MBN outperformed the use of multilevel regression models for each disease separately.

The use of a network-based approach to human disease, so-called *network medicine*, was recently acknowledged to be useful when researching complex disease pathways [33]. In an MBN, the disease variables are also represented as nodes in a network, but the associations have a direction and probabilistic associations are represented by arrows. When there is an arrow from node i to node j, then i is called a *parent* of j, and j is called a *child* of i. Although we cannot assume that these arrows represent true causality, often they do. Temporal arrows always point from the past to the future, and here a causal interpretation is even more natural. Each node is associated with a multinomial probability distribution for each configuration of the parent nodes. The interactions or moderating effects between parent nodes on their common child are therefore captured in these local probability distributions.

In the MBN used for this paper, we modelled the patient's status in terms of the predefined disorders at baseline using the first five years of the data in retrospect, with a registration minimum of three years. The population used here is a fixed cohort of patients that were alive at baseline. The disease status three and five years after the baseline was also included in the model, although patients might have been deceased or moved to another practice at that time. Building the complete MBN required two major steps:

1. Specifying the qualitative nature of the relationships in the network.
2. Specifying the local probability distributions of the disease variables.

For step 1 all the disorders were represented as binary nodes, i.e., the disorder is present yes or no. Age was discretised into four age groups. The subnetworks constituting the resulted three time

slices were then connected in such a way that each disorder variable had at least one directed arrow to the same variable in the next time slice. By doing this, the MBN is transformed into a MTBN, i.e., a multilevel, *temporal* Bayesian network. The urbanity of the practice was operationalized to control for several potential confounding factors, i.e., it was included as a higher level variable in the MTBN.

Although the relationships between the disease variables, i.e., the arrows in the network, can be specified by the user itself, we learned these relationships from the data. One reason for this, is to see whether unknown relationships could be revealed. Both the relationships between disease variables within a time slice and between different time slices were learned using a score-based searching method in the statistical R package *bnlearn* [34]. We assumed that the qualitative nature of associations between disease variables do not vary over time, i.e., the network structure remains the same for each time slice. To ensure a multilevel structure and specific medical knowledge, dependencies between disease variables can be secured or avoided through black- and whitelisting. For example, if an association between a demographic node C and a disease node D was found by the structure learning algorithm, the corresponding arrow in the network clearly should be $C \rightarrow D$. Therefore, all possible arrows $D^i \rightarrow C^j$ were blacklisted.

For step 2 we assumed that, although the network structure remains the same for each time slice, the parameters of the local probability distributions were allowed to change over time. The latter condition is known as the condition of non-stationarity. Once the complete structure was determined, the local probability distributions were estimated using one thousand bootstrapped samples from the dataset, i.e., we computed $P(D_t^i \mid parents(D_t^i))$ with $D_t^i = 1$ if disease i is present at time t and $D_t^i = 0$ otherwise. Note that, $parents(D_t^i)$ is determined by the network structure, it can contain disease variables of several types, e.g., age, a health risk factor, or another disease, from both the current time slice and the previous time slice. The variance induced by the urbanization level (a practice level variable), age, and gender, in the multilevel model was explained by using Markov Chain Monte Carlo simulation in WinBUGS [35]. For ease of use, the learned structure and parameters were put into the software package SamIam.

The total number of disorders present simultaneously in patients, i.e., the multimorbidity rate, was calculated using nodes M_t that kept track of the number of disorders present in time slice t. The probability distribution of these nodes is deterministic, i.e.:

$$P(M_t = q \mid D_t^1, D_t^2, \cdots, D_t^n) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} D_t^i = q \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The value q then represents the number of simultaneously present disorders. See the web materials for more details on the implementation of MTBNs.

Once the local probability distributions were determined we were able to answer the questions mentioned in the introduction. The urbanity effects were derived by conditioning on a specific value of the corresponding node in the network. Secondly, for each time slice we determined whether cumulative incidence rates of disease combinations significantly deviated by increased occurrence from what might be expected from individual cumulative incidence rates, assuming statistical independence. Mathematically this can be expressed for two disorders i_1 and i_2 as:

$$P(D_t^{i_1}, D_t^{i_2} \mid R_t) \gg P(D_t^{i_1} \mid R_t)\ P(D_t^{i_2} \mid R_t) \tag{2}$$

with R_t the a set of health risks R_t^k present at time t. Since the size of the studied patient population favours reaching significance easily, we also examined the clinical importance of such deviations.

By conditioning on the multimorbidity rate M_t in a specific time slice the model allowed us to predict the multimorbidity rate in the next time slice, which is mathematically expressed by:

$$P(M_t = q \mid M_{t-1} = r) \tag{3}$$

We are particularly interested in the probabilities for $q > r$, because if these are large, the number of simultaneous disorders increases with time. These probabilities can be biased due to possible disease shifts, i.e., the patient acquires a new disorder but also loses one, keeping the multimorbidity rate equal. To evaluate this effect we calculated how much acquired disorders sustained in the next time slice, mathematically:

$$P(D_t^i = 1 \mid D_{t-1}^i = 1) \tag{4}$$

If these probabilities are close to one, the effects of disease shifts are considered to be minimal.

Results

The final MTBN consists three time slices modelling chronic cardiovascular disease progression. The associations in the MTBN are summarized in Table 1. The complete network structure representing all the parent-child relations, and thus the qualitative nature of the underlying multivariate distributions, is available on-line together with the corresponding code for parameter estimation. Evidently, age had a significant association with all other variables. However, gender did not had a significant association with CVD, except for ischemic heart disease.

Data of a total of 182,396 patients were used for analysis. The median and mean age of the patients at baseline were 53 and 55 years, respectively. At the end of the five year follow-up, 8.5% of these patients had dropped out of the registry. This happened because of death or patients

moved to a nursery home or practice not present in the registry. Their disease status until this event was included in our analysis.

For all health risks and chronic disorders that were incorporated into the model, cumulative incidence rates, with their standard errors, were estimated for each time slice. These were differentiated for age and urbanity, and to examine the model's validity further, we also made model predictions for diabetics and non-diabetics. Besides the individual rates, we also calculated rates of comorbidity patterns. Details on demographics and probability estimations are available on-line. The effect of urbanity on disease probabilities, corrected for age and gender is shown in Figure 1.

Table 2 shows the evaluation of synergies, as defined in Equation (2), in which cumulative incidence rates of the comorbidity patterns are compared with the rates of single disorders. As dyslipidemia and hypertension are the major predictors of cardiovascular morbidities, we computed the conditional probabilities in the absence and presence of these conditions from the MTBN. Some of the comorbidity patterns deviate significantly from the expected values. For example, at five-year follow-up the probability of ischemic heart disease and heart failure together is 5.4% when both dyslipidemia and hypertension are present. However, using Equation (2), the product of their individual rates is only 2.9%. The true incidence is thus almost twice as high, which indicates an interaction between the two disorders in relation to hypertension and dyslipidemia. This phenomenon can be found for several comorbidity patterns through each time slice. Since probabilities of disease combinations are relatively low, we consider absolute increments, rather than relative increments, where an increment of 0.5% is considered to be of clinical relevance.

Figure 2 shows the multimorbidity rates, as defined in Equation (3). For patients having one or more health risks, the probability of obtaining a new health risk is relatively low, in comparison to the presence of other cardiovascular disorders. In that case, the cumulative incidence rate rapidly increases with the number of conditions present in the previous time slice. For example, when having two disorders at baseline, the probability of obtaining one or more cardiovascular disorders after three year follow-up is approximately 19%+3%=22%, following the arrows from node two (at baseline) to node three and four plus (at three years follow-up) at the right-hand side of Figure 2. From the remaining 78%, those who attracted no new cardiovascular disorder within three years, another 32% gets one or more disorders at five years follow-up, making the total probability 47%.

Table 3 shows the persistence probabilities of individual disorders at follow-ups, as defined in Equation (4). For the majority of the disorders over 90% sustained in the next time slice. Obesity is the major exception on this.

Discussion

In this study, the new method of multilevel temporal Bayesian networks was used to precisely capture the *qualitative* and *quantitative* time course of chronic cardiovascular multimorbidity in

general practices. Bayesian network methods have not been used before in multimorbidity analysis. Although the discovered network dependences are sometimes similar to the associations described in medical literature, discovered by standard statistical means, the global picture of how chronic disorders and risk factors influence each other, as represented by a multilevel Bayesian network, is new. It gives an overview of *direct* and *indirect* associations and it quantifies transition rates, all in one representation. The results obtained are discussed in more detail below.

Evaluation of the network structure

The associations in summarized in Table 1 are compared with current knowledge reflected in recent clinical guidelines and the medical literature. A distinction is made between direct and indirect dependences, something not possible when using standard statistical methods. For example, represented in the MTBN there is an indirect association between dyslipidemia and heart failure via, i.e., conditional on, ischemic heart disease and hypertension. This is in line with clinical guidelines on heart failure, which states that there is no reason to prescribe a statin in the absence of an ischemic cause of heart failure [29].

Other disorders share a common parent in the network structure, e.g., both retinopathy and nephropathy share diabetes mellitus and hypertension as a common parent. The investigation of either of them in the presence of the other, is thus only of beneficial value if either diabetes mellitus or hypertension is also present.

The analysis shows that gender-induced associations are insignificant or small, and in case it is small, it is of little clinical significance. The exception on this is ischemic heart disease, where it is well established that gender is significant [36].

The comorbid associations between cardiovascular related disorders and other chronic disorders [28], are also recognized in our model; obesity, hypertension, dyslipidemia, and diabetes mellitus are associated with *non*-cardiovascular disorders. Moreover, the temporal associations show that diabetes mellitus is a direct predictor of such disorders in follow-ups.

Although the network structure indicates that obesity is a good predictor of hypertension and dyslipidemia, we observe that direct associations between obesity and conditions other than hypertension and dyslipidemia are missing. For example, one would have expected an arrow from obesity to diabetes mellitus, because it is well known that an elevated body mass index or waist circumferences is associated with diabetes mellitus [37]. Hence, the effect of obesity on other conditions is probably underestimated in the data explored, in particular due to low persistence at follow-ups in the database. The latter does not mean that most patients actually lose weight, but that registries do not properly keep track on this matter. However, if we leave obesity out of the model, it has little effect on the structure and the associated probabilities are minimally affected (data not shown).

Quantitative analysis

The prevalences of morbidity and comorbidity patterns at three-year follow-up are comparable to prevalences obtained from previous studies using LINH data, [11] and earlier results within the Netherlands [4]. The associations between age, diabetes and cardiovascular multimorbidity are quantitatively well recognized by the model. The multilevel approach allowed us to differentiate probabilities for practice related variables. Where other researchers showed an association between multimorbidity and socio-economic status [8, 21], we modelled the urbanity of the practice along with the disease variables. In case of obesity, hypertension, dyslipidemia, diabetes mellitus, and ischemic heart disease, the urbanization level had a significant effect on the prevalence. In these cases, the cumulative incidences of moderate and high urban areas were mostly above average, whereas these incidences of low and very high urban areas were mostly below average. Rural areas show on average incidences comparable to the overall average.

The effect of multiple cardiovascular risk factors was already outlined in [38], indicating that the five-year cardiovascular risk can go up to 44% when certain risk factors, e.g., hypertension, total cholesterol, smoking, high density lipoprotein, gender, diabetes, and high age, are present. Their results indicate, for example, that a smoking male patient over 60 years with diabetes, high total cholesterol and low high density lipoprotein, has a 5-year cardiovascular risk of 22% and 44%, for low and high systolic blood pressure respectively. We can do more or less the same exercise as described above: the 5-year cardiovascular risk for a male diabetic living in a rural area, aged between 65 and 80 years, and having a lipid disorder, is 28% for non-hypertensive patients and 55% for hypertensive patients.

However, the multiple-risk attributions in their approach could only be derived by adding the risk factors consecutively in a specific order. In our model there is no restriction on the number of disease variables used as predictor and the number of diseases variables being predicted. For example, the 5-year cardiovascular risk of *two or more* new diseases for is 10% and 23% for the patient described above. It also implies that we can condition on a specific cardiovascular disease already present. When conditioning on heart failure the effect on cardiovascular risk is the highest, e.g., a male hypertensive diabetic with dyslipidemia, living in a rural area, aged between 65 and 80 years, with heart failure already present at baseline, has a 5-year risk of 71% to obtain another cardiovascular disease, and 27% for two or more diseases.

Although any other cardiovascular risk score also represents a personalisation, the major difference with a MTBN is that in a MTBN not all disease variables need to be known. In fact a MTBN captures all predictions for *any* disease variable within the model for *any subset* of the remaining variables. This means that one can also reason the other way around: given the presence of certain diseases, one can make predictions about the presence of specific risk factors, e.g., hypertension or lipid disorders are more likely to be present in the presence of cardiovascular

diseases. In Figure 3 an example is shown of a personalisation, indicating that multimorbidity at baseline predicts future multimorbidity better than the demographics do. This is in line with the idea that the patient's *biological* age is of more importance in relationship to morbidity than *chronological* age [39], and the relation between frailty and the accumulation of deficits [40].

In summary, an MTBN can be used to make predictions for multiple diseases in many ways. To our knowledge this is new in multimorbidity research, and Table 2 only reflects a particular personalisation of cardiovascular risk. It reveals multiple interactions between chronic cardiovascular diseases and related disorders, which occur more frequently at follow-ups in comparison to baseline. We shall not discuss every interaction in detail, but it appears that the presence of hypertension or dyslipidemia are necessary preconditions for finding clinical significant interactions. For example, the combination of ischemic heart disease and heart failure is much higher as expected after three- and five-year follow-up. Alternatively, although cumulative incidences increase over time, they behave as expected, e.g., in the case of ischemic heart disease in combination with stroke. This fits with the fact that only an indirect association exists in the model.

Another new aspect of our model are the temporal associations. They cause the incidence of chronic disorders to rise quickly over time. In particular, the probability of acquiring at least one new chronic cardiovascular disorder increases with the number of chronic cardiovascular disorders already present, regardless of age (Figure 2). At three years follow-up this is respectively 9%, 20%, 22% and 47%, for zero to three disorders present at baseline. At five years follow-up this has increased to respectively 21%, 42%, 47% and 76%. In reality, these numbers can be even higher due to disease shifts. Since the probability of sustaining a chronic disorder is at least 90%, for the ones we used in our model, we believe this effect is minimal.

Cross-sectional research of other registries show that the prevalence of multimorbidity can be up to 90% [4–6, 21]. Moreover, 80% of the elderly patients with heart failure face at least *four* chronic comorbidities [28]. These numbers are comparable with the prevalences of cardiovascular comorbidity at follow-ups retrieved in our model. However, the prevalence of cardiovascular multimorbidity, in particular for diabetes mellitus, ischemic heart disease and stroke, is much lower at baseline in our model. This indicates the importance of the temporal dimension; estimates cannot be directly extrapolated to follow-ups, e.g., by using the prevalences of a higher age group.

Strength, limitations and implications

The major strength of the results is that the obtained MTBN allows analysing several aspects of multimorbidity in a single model. Our research encompassed an analysis of data obtained from public health registries, and because of the size of the data set used, significant results could be established. Although the data used here contained more patients and more disease variables as mostly present in controlled studies, it also contains more noise and typically involved more

preprocessing. Controlled studies are relatively small in size and often exclude patients with multimorbidity. There are some exceptions on this, e.g., recently a cohort study of nearly 15,000 elderly people had a focus on the epidemiology of chronic diseases using a variety of biomarkers and non-invasive measurements [41]. But the majority of its research has a single-disease focus. We recommend to apply Bayesian networks here as well to discover the coherence between the multiple biomarkers and disorders present in such studies.

Several aspects of the results could have been analysed by alternative methods. For instance, multi-state models could have been used to analyse the transition rates in the multimorbidity number, a separate multilevel regression model for each disease to investigate the urbanity effects, and a chi-squared test to see if the joint prevalence of two conditions is higher than would be expected. However, investigating more complex interactions, like in Table 2, would require logistic regression with added *interaction terms* [42]. Logistic regression demands building a separate model for each disorder; in this way, the insight into the qualitative nature of the interactions between the disorders would be lost.

With an MTBN we also avoid the redundancy that is obtained when using multiple separate regression models for each disease. For example, if we regress disease D on disease D' in one model and in another model D' on D, we obtain two parameters for the same association. This could produce certain ambiguity because the two models do not necessarily have to provide the same odds ratio for that pair. In summary, the MTBN used here allows analysing all the results presented in this paper without losing any of the epidemiological coherence between all the disease variables. To our knowledge, this is new in multimorbidity research, and there is no *single* alternative that analyses multimorbidity the way we did.

There are some aspects of registries that introduce a certain bias. Patients that did not visit their physician within the used time frame are not included in the data. Although we explored a decade of patient data within a time frame of ten years and patients of age 35 and above, making the proportion of missing patients likely to be very low, prevalences are probably slightly overestimated. On the other hand, in a public health registries data are missing and there are also incorrectly coded diagnoses, implying that prevalences might also be underestimated. Clinical guidelines often recommend specific additional investigations making it that certain disorders are discovered more likely than others. For example, retinopathy in a diabetic may only be discovered because of the recommendation mentioned in the guideline of visiting an ophthalmologist.

In our results, there is a considerable prevalence change between the baseline and follow-ups. Partly this is due to the fact that the population has aged five year. On the other hand, the absence of a diagnosis is interpreted as the absence of the corresponding disease, however, certain pathophysiology could already be present at baseline without knowing it. This delay in diagnosis also makes longitudinal associations between disorders less detectable.

Conclusions

Several attempts have been made in the literature to capture prevalences of multimorbidity. Lately, electronic databases of general practices are used more and more to quantify these numbers on a larger scale, but there is no clear method that fully describes how multiple disease evolve over time. Traditional statistical techniques are very useful to evaluate a single-disease framework by which most medical care, research, and education is configured. However, a multiple-disease orientation requires a more complementary strategy.

In that respect, the MTBN used here is a valuable step forwards in multimorbidity research. It combines the advantages of a temporal Bayesian network together with a multilevel analysis, and it was able to discover complex multimorbidity patterns of chronic diseases within health care data. First, the model was shown to be valid by comparing known disease interactions for diabetes mellitus with those present in the network structure. Second, several new disease interactions, qualitative and quantitative, for three and five years follow-up were discovered, showing that cumulative incidence rates are accelerated in the presence of multimorbidity. Especially, the presence of conditions such as hypertension, dyslipidemia, and diabetes mellitus accelerates cardiovascular risk significantly.

The paper only discusses the most significant results that can be obtained from an analysis of the MTBN model. The model itself can be used to extract many other relevant conclusions. We conclude that Bayesian network models make the analysis and visualization of the interactions between chronic disorders and their evolutionary course more comprehensive than traditional statistical techniques. They can be used to answer a variety of clinical and epidemiological questions without losing the context of these dependences out of sight. This is of great importance in the management of multimorbidity and the aim to adopt personalised clinical guidelines. The next step in multimorbidity research might be to address mortality rates, differentiated for cardiovascular and non-cardiovascular chronic diseases, in the same way as was done here for multimorbidity rates.

Acknowledgements

This paper is funded by The Netherlands Organisation for Health Research and Development. Data acquisition took place at the Netherlands Institute for Health Services Research, and the authors are grateful to all employees and general practitioners that participated in the data registration. The authors declare that there is no conflict of interest.

References

1 Feinstein A. The pretherapeutic classification of comorbidity in chronic disease. J Chronic Dis 1970;23:455–68.

2 van den Akker M, Buntinx F, Knottnerus J. Comorbidity or multimorbidity; what's in a name? a review of the literature. Eur J General Practice 1996;2:65–70.

3 van den Akker M, Buntinx F, Roos S, Knottnerus J. Problems in determining occurrence rates of multimorbidity. J Clin Epidemiol 2001;54:675–9.

4 van den Akker M, Buntinx F, Metsemakers J, Roos S, Knottnerus J. Multimorbidity in general practice: prevalence, incidence and determinants of co-occurring chronic and recurrent diseases. J Clin Epidemiol 1998;51:367–75.

5 Wolff J, Starfield B, Anderson G. Prevalence, expenditures, and complications of multiple chronic conditions in the elderly. Arch Intern Med 2002;162:2269–76.

6 Fortin M, Hudon C, Haggerty J, van den Akker M, Almirall J. Prevalence estimates of multimorbidity: a comparative study of two sources. BMC Health Services Research 2010;10:111–7.

7 Uijen A, van Lisdonk E. Multimorbidity in primary care: Prevalence and trend over the last 20 years. Eur J Gen Pract 2008;14(1):28–32.

8 Salisbury C, Johnson L, Purdy S, Valderas J, Montgomery A. Epidemiology and impact of multimorbidity in primary care. Brit J of Gen Prac 2011;61(582):e12–21.

9 Gunn J, Ayton D, Densley K, Pallant J, Chondros P, Herrman H, et al. The association between chronic illness, multimorbidity and depressive symptoms in an australian primary care cohort. Soc Psychiatry Psychiatr Epidemiol 2012;47:175–84.

10 Britt H, Harrison C, Miller G, Knox S. Prevalence and patterns of multimorbidity in australia. Med J Aust 2008;189(2):72–7.

11 van Oostrom S, Picavet H, van Gelder B, Lemmens L, Hoeymans N, Verheij R, et al. Multimorbidity and comorbidity in the dutch population - data from general practices. Nederlands Tijdschrift voor Geneeskunde 2011;155(A3193).

12 Holden L, Scuffham P, Hilton M, Muspratt A, Ng S, Whiteford H. Patterns of multimorbidity in working australians. Population Health Metrics 2011;9(1):115–20.

13 Schäfer I, von Leitner E, Schön G, koller D, Hansen H, Kolonko T, et al. Multimorbidity patterns in the elderly: A new approach of disease clustering identifies complex interrelations between chronic conditions. Plos One 2010;5(12):e15941–.

14 Marengoni A, Rizzuto D, Wang H, Winblad B, Fratiglioni L. Patterns of chronic multimorbidity in the elderly population. J Am Geriatr Soc 2009;57:225–30.

15 Kirchberger I, Meisinger C, Heier M, Zimmerman A, Thorand B, Autenrieth C, et al. Patterns of multimorbidity in the aged population - results from the kora-age study. PLoS One 2012;7(1):e30556–.

16 Caughey C, Vitry A, Gilbert A, Roughead E. Prevalence of comorbidity of chronic diseases in australia. BMC Public Health 2008;8:221–33.

17 Diederichs C, Berger K, Bartels D. The measurement of multiple chronic diseases - a systematic review on existing multimorbidity indices. J Gerontol A Biol Sci Med Sci 2011;66(3):301–11.

18 Marengoni A, Angleman S, Melis R, Mangialasche F, Karp A, Garmen A, et al. Aging with multimorbidity: A systematic review of the literature. Ageing Research Reviews 2011;10:430–9.

19 Glynn L, Buckley B, Reddan D, Newell J, Hinde J, Dinneen S, et al. Multimorbidity and risk among patients with established cardiovascular disease. British Journal of General Practice 2008;58:488–94.

20 Boyd C, Darer J, Boult C, Fried L, Boult L, Wu A. Clinical practice guidelines and quality of care for older patients, with multiple comorbid diseases: Implications for pay for performance. JAMA 2005;294(6):716–24.

21 Barnett K, Mercer S, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. Lancet 2012;380:37–43.

22 Guthrie B, Payne K, Alderson P, McMurdo M, Mercer S. Adapting clinical guidelines to take account of multimorbidity. BMJ 2012;345:e6341–.

23 Hippisley-Cox J, Hammersley V, Pringle M, Coupland C, Crown N, Wright L. Methodology for assessing the usefullness of general practice data for research in one research network. Health Informatics Journal 2004;10(2):91–109.

24 Wong A, Boshuizen H, Schellevis F, Kommer G, Polder J. Longitudinal administrative data can be used to examine multimorbidity, provided false discoveries are controlled for. J of Clin Epidemiol 2011;64:1109–17.

25 Hox J. Multilevel Analysis: techniques and applications. New York, USA: Routledge; 2010.

26 Lappenschaar M, Hommersom A, Lucas P, Lagro J, Visscher S. Multilevel Bayesian networks for the analysis of hierarchical health data. Artificial Intelligence in Medicine 2013;57:171–83.

27 O'Halloran J, Miller G, Britt H. Defining chronic conditions for primary care with icpc-2. Familiy Practice 2004;21:381–6.

28 van der Wel M, Jansen R, Bakx J, Bor H, OldeRikkert M, van Weel C. Non-cardiovascular co-morbidity in elderly patients with heart failure outnumbers cardiovascular co-morbidity. European Journal of Heart Failure 2007;9:709–15.

29 Perk J, de Backer G, Gohlke H, Graham I, Reiner Z, Verschuren WM, et al. European society of cardiology guidelines on cardiovascular disease prevention in clinical practice. European Heart Journal 2012;33:1635–701.

30 Wiersma T, Smulders Y, Stehouwer C, et al. . Multidisciplinary Guideline on Cardiovascular Risk Management. Houten, The Netherlands: Bohn Stafleu van Loghum; 2011.

31 Pearl J. Probabilistic Reasoning in Intelligent Systems. San Francisco, CA, USA: Morgan Kaufmann; 1988.

32 Lucas P, van der Gaag L, Abu-Hanna A. Bayesian networks in biomedicine and health-care. Artificial Intelligence in Medicine 2004;30(3):201–14.

33 Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet 2011;12(1):56–68.

34 Scutari M. Learning Bayesian networks with the bnlearn R package. Journal of Statistical Software 2010;35(3):1–22.

35 Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS User Manual; Version 1.4. Cambridge, UK: MRC Biostatistics Unit; 2001.

36 Vaccarino V, Badimon L, Corti R, de Wit C, Dorobantu M, Hall A, et al. Ischaemic heart disease in women: are there sex differences in pathophysiology and risk factors? position paper from the working group on coronary pathophysiology and microcirculation of the european society of cardiology. Cardiovascular Research 2011;90:9–17.

37 Ryden L, Standl E, Bartnik M, van den Berghe G, Betteridge J, de Boer M, et al. European society of cardiology guidelines on diabetes, pre-diabetes, and cardiovascular diseases: executive summary. European Heart Journal 2007;28:88–136.

38 Jackson R, Lawes C, Bennet D, Milne R, Rodgers A. Treatment with drugs to lower blood-pressure and blood cholesterol based on an individual's absolute cardiovascular risk. Lancet 2005;365:434–41.

39 Mitnitski A, Graham J, Mogilner A, Rockwood K. Frailty, fitness and late-life mortality in relation to chronological and biological age. BMC Geriatrics 2002;2:1–8.

40 Rockwood K, Mitnitski A. Frailty in relation to the accumulation of deficits. Journal of Gerontology: Medical Sciences 2007;62A(7):722–7.

41 Hofman A, van Duijn C, Franco O, Ikram M, Janssen H, Klaver C, et al. The rotterdam study: 2012 objectives and design update. Eur J Epidemiol 2012;26:657–86.

42 Jaccard J, Turrisi R. Interaction effects in multiple regression. Sage University Papers Series on Quantitative Applications in the Social Sciences 2003;:07–072Thousand Oaks, CA: Sage.

43 Nielen M, Schellevis F, Verheij R. Inter-practice variation in diagnosing hypertension and diabetes mellitus: a cross-sectional study in general practice. BMC Family Practice 2009;10:1–6.

44 Rong J, Yu C, Yang P, Chen J. Diabetic retinopathy: A predictor of coronary artery disease. Diabetes and Vascular Disease Research 2012;online first:1–8.

45 McMurray J, Adamapoulos S, Anker S, Auricchio A, Böhm M, Dickstein K, et al. European society of cardiology guidelines for the diagnosis and treatment of acute and chronic heart failure. European Heart Journal 2012;33:1787–847.

46 Weiner D, Tighiouart H, Amin M, Stark P, MacLeod B, Griffith J, et al. Chronic kidney disease as a risk factor for cardiovascular disease and all-cause mortality: A pooled analysis of community-based studies. J Am Soc Nephrol 2004;15:1307–15.

47 Magri C, Calleja N, Buhagiar C, Fava S, Vassallo J. Factors associated with diabetic nephropathy in subjects with proliferative retinopathy. Int Urul Nephrol 2012;44:197–206.

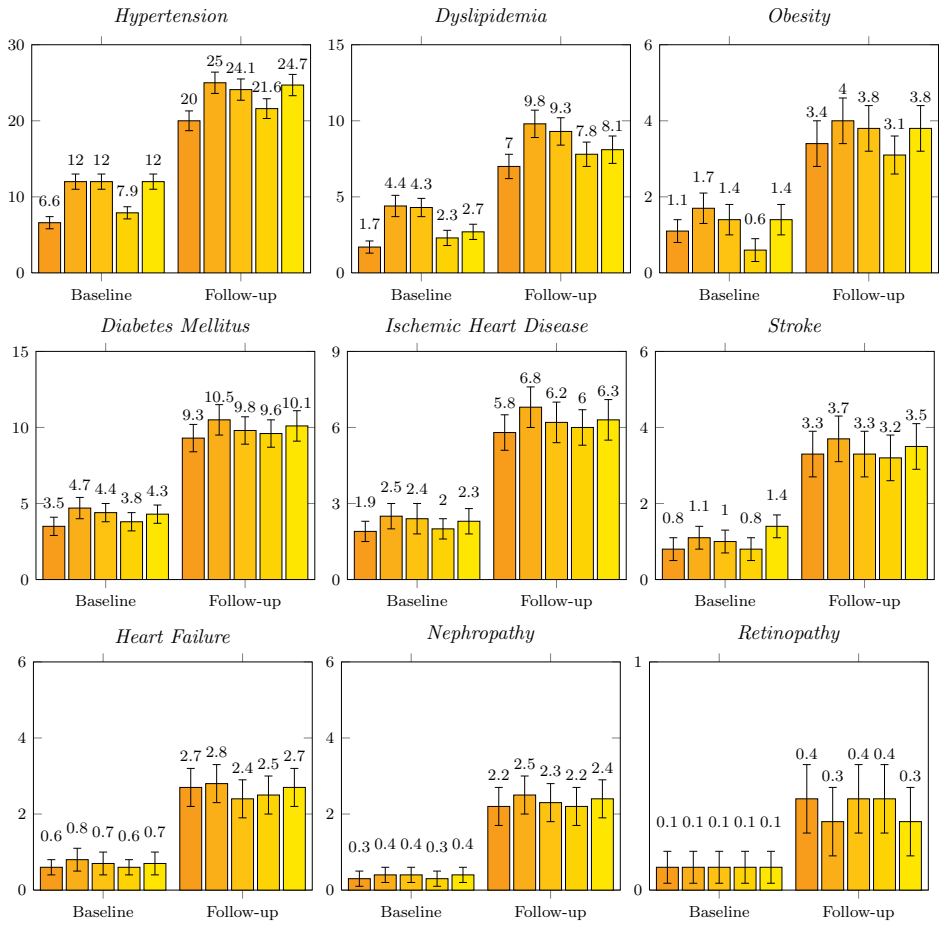

Figure 1: The effect of the ubanization level, varying from very high urban areas (> 2500 addresses per km^2, orange) to rural areas (< 500 addresses per km^2, yellow), on cumulative disease incidence at base-line and five year follow-up. Numbers are corrected for age and gender and provided with a 95% confidence interval.

Chronic Disease	Associations known from the literature	Associations learned from the data	
		Direct*	Indirect**
diabetes mellitus	age, dyslipidemia, hypertension, ischemic heart disease, heart failure, nephropathy, retinopathy obesity, stroke [37]†, practice [43]	age, dyslipidemia, hypertension, ischemic heart disease, heart failure, nephropathy, retinopathy,	practice, obesity, stroke
ischemic heart disease	age, gender, obesity, dyslipidemia, hypertension, diabetes mellitus, heart failure, stroke [29]†, retinopathy [44], practice [23]	age, gender, dyslipidemia, hypertension, diabetes mellitus, heart failure	practice, obesity, stroke, *nephropathy*, retinopathy
heart failure	age, obesity, dyslipidemia, hypertension, diabetes mellitus, ischemic heart disease, nephropathy, stroke [45]†	age, hypertension, diabetes mellitus, ischemic heart disease, stroke, nephropathy	*practice*, obesity, dyslipidemia, *retinopathy*
stroke	age, obesity, dyslipidemia, hypertension, diabetes mellitus, ischemic heart disease, heart failure [29]†, practice [23]	age, dyslipidemia, hypertension, heart failure	practice, obesity, diabetes mellitus, ischemic heart disease, *nephropathy*, *retinopathy*
nephropathy	age, gender, hypertension, diabetes mellitus, ischemic heart disease, heart failure, stroke [46], retinopathy [47]	age, hypertension, diabetes mellitus, heart failure	*practice*, *obesity*, dyslipidemia, ischemic heart disease, retinopathy
retinopathy	age, dyslipidemia, hypertension, diabetes mellitus [37]†, ischemic heart disease [44], nephropathy [47]	age, hypertension, diabetes mellitus	*practice*, *obesity*, dyslipidemia, ischemic heart disease, *heart failure*, nephropathy

Table 1: Associations between cardiovascular diseases, known from the literature (clinical guidelines are marked with a †), and learned from the data. * Direct associations correspond with a direct arrow between diseases in the MTBN. ** Indirect associations correspond with diseases that have another disease between them or share a common child or parent in the MTBN. Italic written associations were not directly clear from the used literature.

	BaseLine				3 years follow-up				5 years follow-up			
Risk Factors	None	DL	HT	DL+HT	None	DL	HT	DL+HT	None	DL	HT	DL+HT
Comorbidity												
DM+IHD	0.2	2.6	1.8	6.2	0.7	5.8	4.4	11.2	1.0	6.8	5.4	14.0
DM+HF	<.1	0.5	0.7	1.3	0.4	1.6	2.1	3.9	0.5	2.5	3.0	5.0
DM+NP	<.1	0.3	0.4	0.9	0.3	1.1	2.0	3.8	0.5	1.7	3.1	5.0
DM+ST	<.1	0.7	0.7	2.6	0.2	1.9	2.1	5.3	0.4	2.4	2.9	6.4
DM+RP	<.1	0.1	0.1	0.2	0.1	0.2	0.2	0.3	0.1	0.3	0.3	0.4
IHD+ST	<.1	0.6	0.4	1.7	0.2	1.6	1.3	3.8	0.3	2.3	2.0	4.9
IHD+NP	<.1	<.1	0.2	0.5	0.1	0.6	1.0	2.0	0.2	1.1	1.7	3.4
IHD+HF	<.1	0.8	0.9	1.8	0.4	2.1	2.2	3.9	0.6	2.8	3.2	5.4
ST+HF	<.1	0.1	0.3	0.4	0.2	0.5	0.9	1.4	0.3	0.9	1.4	2.2
NP+HF	<.1	<.1	0.1	0.2	0.1	0.4	0.9	1.1	0.3	0.6	1.5	2.0

Table 2: Probability of having comorbid combinations of chronic cardiovascular diseases at baseline, and after three and five years follow-up, under condition of the presence or absence of health risks. Abbreviations: HT=hypertension; DL=dyslipidemia; DM=diabetes mellitus; IHD=ischemic heart disease; HF=heart failure; ST=stroke; NP=nephropathy; RP=retinopathy. Results are shown in percentages. The yellow part of the circle represents the expected value based on individual rates, the surplus is coloured in orange or red (see also Equation (2) of the paper). Red circles represent cumulative incidence rates which deviate significantly ($p < 0.001$) from the expected values and have a clinical importance as well (absolute increase $> 0.5\%$). They indicate the clinical significant interactions.

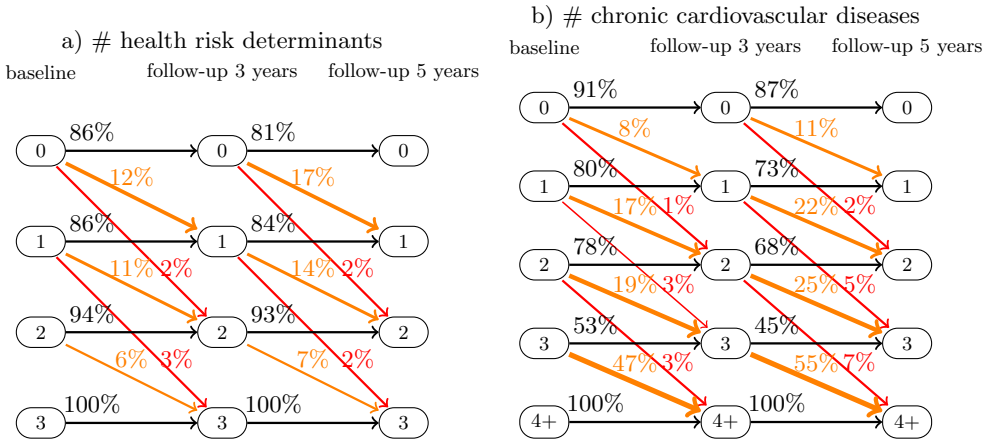

Figure 2: Transition probabilities of a) health risks, and b) chronic cardiovascular diseases (see also Equation (3) of the paper). Health risks are obesity, dyslipidemia, and hypertension. Chronic cardiovascular diseases are diabetes mellitus, ischemic heart disease, heart failure, stroke, retinopathy, and nephropathy. The left (black) percentages and lines represent patients who do not acquire a new determinant or chronic disease within the next time slice, as where the middle (orange) and right (red) percentages and lines represent patients acquiring respectively one or two new health risks or chronic diseases within the next time slice.

Chronic Disorder	$t_0 \rightarrow t_1$	$t_1 \rightarrow t_2$
Obesity	40	33
Dyslipidemia	78	90
Hypertension	95	97
Diabetes Mellitus	95	91
Ischemic Heart Disease	94	98
Heart Failure	95	99
Stroke	90	91
Nephropathy	98	99
Retinopathy	99	99
Other	66	90

Table 3: Persistence of individual chronic diseases, i.e., the probability (in percentages) of a disorder being present at follow-ups under the condition that this disorder was present in the previous time slice (see also Equation (4) of the paper). t_0=baseline, t_1=follow-up 3 years, and t_2=follow-up 5 years.

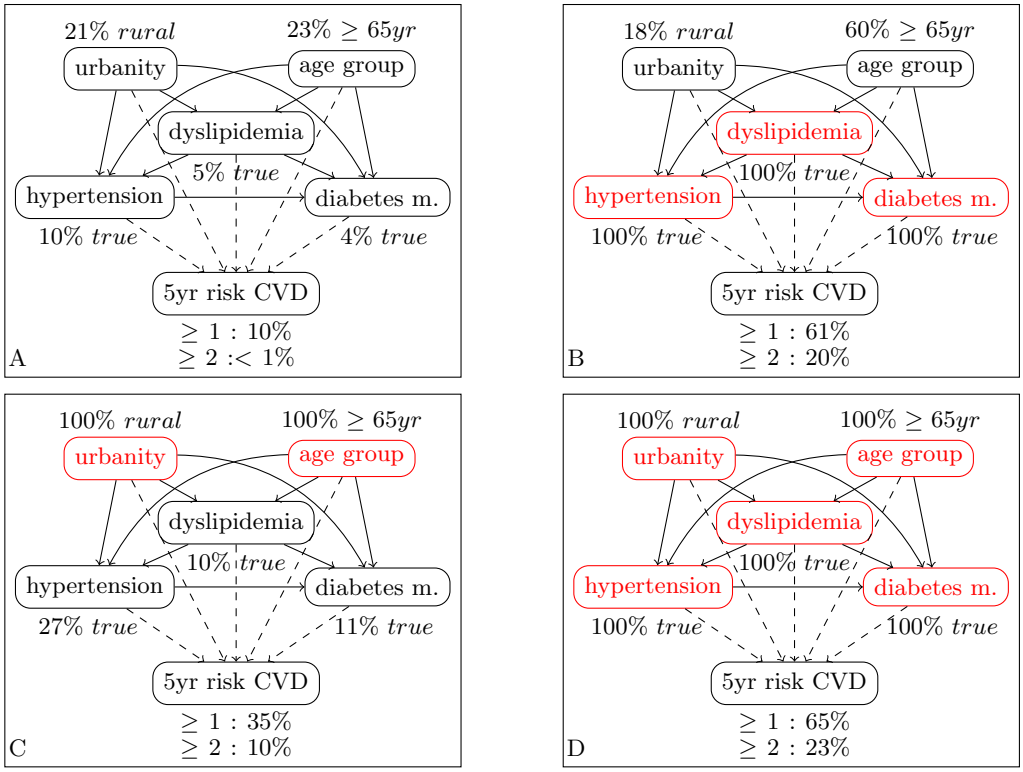

Figure 3: personalisation of a subset of the disease model. Straight lines represent the atemporal associations at baseline, and dashed lines represent the temporal associations between baseline conditions and the five year cardiovascular risk. Subfigure (A) shows the 5-year risk on a cardiovascular (related) disease (ischemic heart disease, heart failure, stroke, retinopathy, or nephropathy) of the general dutch population aged over 35 years. This probability is 10%, and within this population we have the following prior probabilities of risk factors: 21% lives in a rural area, 23% is older than 65 years, 10% has hypertension, 5% has dyslipidemia, and 4% has diabetes mellitus. The subfigures (B), (C) and (D) show how probabilities change when specific information of the patient is incorporated. (B) represents patients with hypertension, dyslipidemia and diabetes at baseline, but without knowledge about the patient's demographics. The 5-year risk has increased to 61%, and it is estimated that 18% lives in a rural area and 60% is over 65 years. (C) represents patients older than 65 years, living in a rural area, but without knowledge about the patient's disease status. The 5-year risk on CVD has increased to 45% and it is estimated that cumulative incidences of risk factors are doubled or more. (D) is the combination of (B) and (C), showing that the 5-year risk has gained only a little with respect to (B).