A Dynamic Bayesian Network model for simulation of disease progression in Amyotrophic Lateral Sclerosis patients

Alessandro Zandonà¹, Matilde Francescon¹, Maya Bronfeld²; Andrea Calvo³; Adriano Chiò³ and Barbara di Camillo¹

¹ Department of Information Engineering, University of Padova, Padova, Italy
² Prize4Life, Haifa, Israel
³ ‘Rita Levi Montalcini’ Department of Neuroscience, University of Turin, Turin, Italy

Corresponding Author:

Barbara Di Camillo¹
via Gradenigo 6B, 35131 Padova, Italy

Email address: barbara.dicamillo@unipd.it
A Dynamic Bayesian Network model for simulation of disease progression in Amyotrophic Lateral Sclerosis patients

Alessandro Zandonà¹, Matilde Francescon¹, Maya Bronfeld²; Andrea Calvo³; Adriano Chiò³ and Barbara di Camillo¹

¹ Department of Information Engineering, University of Padova, Padova, Italy
² Prize4Life, Haifa, Israel
³ ‘Rita Levi Montalcini’ Department of Neuroscience, University of Turin, Turin, Italy

Abstract

Background. Amyotrophic lateral sclerosis (ALS) is a progressive neurodegenerative disease primarily affecting upper and lower motor neurons in the brain and spinal cord. The heterogeneity in the course of ALS clinical progression and ultimately survival, coupled with the rarity of this disease, make predicting disease outcome at the level of the individual patient very challenging. Besides, stratification of ALS patients has been known for years as a question of great importance to clinical practice, research and drug development.

Methods. In this work, we present a Dynamic Bayesian Network (DBN) model of ALS progression to detect probabilistic relationships among variables included in the Pooled Resource Open-Access ALS Clinical Trials Database (PRO-ACT), which provides records of over 10,700 patients from different clinical trials, and with over 2,869,973 longitudinally collected data measurements.

Results. Our model unravels new dependencies among clinical variables in relation to ALS progression, such as the influence of basophil count and creatine kinase on patients’ clinical status and the respiratory functional state, respectively. Furthermore, it provided an indication of ALS temporal evolution, in terms of the most probable disease trajectories across time at the level of both patient population and individual patient.

Conclusions. The risk factors identified by our DBN model could allow patients’ stratification based on velocity of disease progression and a sensitivity analysis on this latter in response to changes in input variables, i.e. variables measured at diagnosis.

Introduction

Amyotrophic lateral sclerosis (ALS) is a progressive neurodegenerative disease primarily affecting upper and lower motor neurons in the brain and spinal cord. The incidence of ALS in Europe and in the United States is approximately 3 and 2 cases per 100,000 people, respectively; the highest prevalence rate of the disease is registered among males and in people above the age of fifty. The region of onset is typically within the upper limb, lower limb or bulbar musculature. Commonly, the first symptoms are a gait disturbance (foot drop, difficulty walking and lifting arms due to weakness), dysarthria followed by dysphagia or impaired fine movements in the upper extremities [2]. Once the first symptoms develop, ALS leads to a relentlessly loss of vital functions, weakness and worsening paralysis of voluntary muscles and eventually death. Mean life expectancy is three to five years, with respiratory failure being the most common cause of death [3, 4]. The most widely used rating instrument for monitoring the progression of disability in ALS patients is the ALS Functional Rating Scale (ALSFRS) and its revised version ALSFRS-R, which is based on 12 items rated on a 0-4 point scale evaluating
bulbar functions, fine and gross motor skills and respiratory functions. However, the heterogeneity in the course of ALS clinical progression and ultimately survival, coupled with the rarity of this disease, make predicting disease outcome at the level of the individual patient very challenging. This presents substantial barriers to the planning and interpretation of clinical trials of new treatments, leading to large, expensive, and potentially unbalanced trials. In addition, patients might respond to ALS medications differently: certain ALS therapies tested in clinical trials unsuccessfully were suggested to have beneficial effects on specific subpopulations of patients [2, 5]. Therefore, stratification of ALS patients has been known for years as a question of great importance to clinical practice, research and drug development [6,7].

In this work, we present a Dynamic Bayesian Network (DBN) model of ALS progression to detect probabilistic relationships among variables included in the Pooled Resource Open-Access ALS Clinical Trials Database (PRO-ACT), which provides records of over 10,700 patients from different clinical trials, with over 2,869,973 longitudinally collected data measurements. A Bayesian Network [8, 9] is a mathematical model that represents the joint probability distribution of a set of random variables as a directed acyclic graph (DAG). A DBN extends a Bayesian Network to model dynamic processes, describing the dependencies among the variables along time [9].

Our model unravels new dependencies among clinical variables in relation to ALS progression, such as the influence of basophil count or bicarbonate on movement, communication and respiratory functional state. Furthermore, it provided an indication of ALS prognosis, in terms of the most probable disease trajectories across time at the level of both patient population and individual patient.

Materials & Methods

We considered data of ALS patients assessed over subsequent screening visits, included in the PRO-ACT database (https://nctu.partners.org/ProACT/). A broad spectrum of information was available for each patient including demographics, laboratory data (e.g., blood count, electrolytes), vital signs (e.g., blood pressure, pulse), ALS history, concomitant medication use, outcome measures (e.g., slow and forced vital capacity, survival) and ALSFRS, for a total of 186 variables. Time-dependent data collected over subsequent visits were defined as dynamic variables, while either time-independent covariates (e.g., race) or data collected at first visit only (e.g., age) were defined as static variables. Variables with missing values in more than 50% of patients were filtered out; 48 variables (7 static, 41 dynamic) remained after this step. Dynamic variables were longitudinally collected during clinical visits, thus both their status at a specific visit (time t) and at the previous one (t-1) were reported for each patient, in order to account for the relationship among two consecutive time points.

An additional variable (DELTA_step) was added to account for the time between two consecutive visits. Continuous variables were discretized either according to the 33rd and the 66th percentile or based on clinical threshold as given in PRO-ACT website (https://nctu.partners.org/ProACT/). For example, pulse was discretized in 3 levels: lower than 72, between 72 and 80, or higher than 80 beats per minute. The preprocessed dataset was then modelled as a DBN, a directed graph representing variables as nodes and conditional dependencies among them as directed edges. In order to learn the variable dependencies, the DBN was inferred through a two-steps procedure: i) learning the structure (i.e. the parents-children dependencies among nodes) and ii) learning the parameters (i.e. the probability that a variable assumes a specific value conditional to the value assumed by its parents) of the graph.

Variables were grouped into six layers (Table 1) where variables in layer j could be dependent only on variables in layers i ≤ j. In particular, variables onset_site and onset_delta (layer 2) can
be affected by gender and age as known in the literature [4, 5, 6]; the variables at visit \( t \) are affected by the static variables in first two layers, by variables at visit \((t-1)\) and by \( \text{DELTA\_step}, \) \( i.e., \) the delay between visit \((t-1)\) and visit \( t \) since this might vary for different patients. The variable status labels each patient either as dead during the PRO-ACT trials or still alive at the last visit; status can be affected by any other variable in the dataset.

Structure learning was performed using the Hill-Climbing (HC) algorithm [10], a search-and-score method that starts with the complete space of possible graph structures and repeats as long as the Bayesian-Dirichlet equivalent uniform scoring function is maximized. The set of parameters of the conditional probability distribution at each node were learnt using a Maximum a Posteriori estimation.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Variable</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gender, Age at ALS onset, Race, Riluzole</td>
<td>Static</td>
</tr>
<tr>
<td>2</td>
<td>Onset site, Onset delta (start of the trial - onset)</td>
<td>Static</td>
</tr>
<tr>
<td>3</td>
<td>Variables at time ( t-1 )</td>
<td>Dynamic</td>
</tr>
<tr>
<td>4</td>
<td>( \text{DELTA_step} )</td>
<td>Dynamic</td>
</tr>
<tr>
<td>5</td>
<td>Variables at time ( t )</td>
<td>Dynamic</td>
</tr>
<tr>
<td>6</td>
<td>Status</td>
<td>Static</td>
</tr>
</tbody>
</table>

Table 1. Layering and type of variables in the DBN.

The DBN was inferred on 8635 patients (the training set) and used to simulate the temporal evolution of 1289 patients (the validation set) by sampling, at each discrete time point, the state of a patient conditioned on his/her state in the previous time point in accordance to the conditional probability tables encoding variable dependencies. We started our simulation at time \( t = 0 \) (the first visit) and let the simulated patients evolve for 100 months with a time-step of 1 month. For each real patient, we generated 100 different temporal evolutions so to have a distribution of variable values in time, for a total of \( 1289 \times 100 \) simulated patients. In addition, to provide an estimation of a level of confidence on the edges between variables, 100 bootstrap samples of the original dataset were considered and the resulting DAGs were converted into the corresponding partially directed acyclic graph (PDAG) representing the corresponding I-equivalent classes and aggregated in a Weighted Partially Directed Acyclic Graph (WPDAG). WPDAGs encode the confidence on the presence of each edge as the fraction of bootstrap samples with that edge.

Data pre-processing and network learning were implemented in R, combining a set of in-house scripts with \textit{bnstruct} [11], a recently developed R package providing structure and parameter learning of Bayesian networks even in presence of missing values, which is a common situation in clinical setting, and automating the possibility of using bootstrap to construct WPDAGs.

Results

Figure 1A reports the distribution of the temporal evolution of variable Q10 (respiratory functional state) in the validation set and in the DBN prediction (from training set), showing a good correspondence (Mean Squared Error: 1.1e-01) between simulated and real distributions. Similar results were obtained for the other dynamic variables. As shown in figure 1B, our model can be used also to simulate, for a single patient, the time-dependent probability of a variable being in a certain state. Figure 1C shows the WPDAG learnt from 100 bootstrap samples of the original dataset; only edges with the maximum confidence, hence observed for all 100 DAGs, are represented. Notably, our model identified some expected dependencies between variables, such as the influence of gender and age on the onset site. Indeed, it is known that the spinal onset is more common in men than in women and that bulbar phenotype typically affects women older than 65 years [12]. In addition, some new dependencies among variables were
highlighted. For instance, our analysis revealed the dependency between patients’ clinical status (status encoded by ALSFRS Q1-Q10) and the basophil count, suggesting an ALS-induced inflammatory state [13]. A relationship between the level of creatine kinase - correlated to creatinine [14] - and both respiratory insufficiency and the difficulty in eating was also suggested by our results.

Figure 1. A) Distribution of the temporal evolution of variable Q10 (respiratory functional state) in the validation set (blue line) and in the prediction (red line), based on probabilities modelled by DBN. B) Stacked histogram of the probable Q10 values (color-coded) predicted along 100 months (x-axis) for a single patient. 100 different temporal evolutions were generated so to have a distribution of variable values in time. C) Subset of the WPDAG obtained on 100 bootstrap samples. Only edges with the maximum confidence (present in all of 100 networks) are shown. Dynamic variables at time $t$ (green nodes) depend on their state at time $t-1$ (edges not shown), and can depend on other dynamic variables both at time $t$ and at time $t-1$ (blue nodes) as well as on static variables (orange nodes). Q1,…. Q10: ALSFRS scale items.
onset_delta: Time between start of the trial and ALS onset (days<-771, -771<days<-443, days>-443)

onset_site: Site of disease onset (bulbar, limb)

CK: Creatine Kinase (male: <38 u/L, 38-174 u/L, >174 u/L; female: < 96 u/L, 96-140 u/L, >140 u/L)

FVC: Forced Vital Capacity (<60%, 60-80%, >80%)

DELTA_step: Time between two consecutive visits (days<28, 28<days<54, days>54)

bp_diastolic: diastolic blood pressure (<78 mmHg, 78-85 mmHg, >85 mmHg)

Conclusions

ALS clinical heterogeneity and progression hinder the diagnosis, the prognosis and the treatment of the disease. Our study took advantage of a wide spectrum of clinical, longitudinal ALS patient information and built a Dynamic Bayesian Networks (DBN) unravelling the probabilistic dependencies among variables over time and allowing the prediction of ALS temporal evolution in time, starting from a single patient, screen visit data. Consequently, our approach could provide an indication of ALS prognosis: this would be the basis of a prevention strategy and a personalized therapy, because the treatment could be guided and re-targeted based on the most probable disease progression. In future study, we will use our model to identify risk factors that allow patients' stratification based on velocity of disease progression and on how sensitive this latter is to changes in input variables, i.e. variables measured at diagnosis.

Funding

This work was funded by the bilateral Italian-Israel project CompALS (Computational analysis of the clinical manifestations and predictive modelling of ALS), supported by the Italian Ministry of Foreign Affairs and International Cooperation and the Ministry of Science, Technology and Space of the State of Israel.

References


