Journal of Biomedical Informatics 58 (2015) 133–144

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Prognostic models based on patient snapshots and time windows: Predicting disease progression to assisted ventilation in Amyotrophic Lateral Sclerosis

André V. Carreiro [a,*], Pedro M.T. Amaral [a], Susana Pinto [b], Pedro Tomás [a], Mamede de Carvalho [b], Sara C. Madeira [a,*]

[a] INESC-ID Lisbon and Instituto Superior Técnico, Universidade de Lisboa, Portugal
[b] Translational Clinical Physiology Unit, Institute of Molecular Medicine, Faculty of Medicine, Universidade de Lisboa, Portugal

ARTICLE INFO

Article history:
Received 19 February 2015
Revised 18 September 2015
Accepted 23 September 2015
Available online 8 October 2015

Keywords:
Prognostic model
Disease progression
Amyotrophic Lateral Sclerosis
Time windows
Patient snapshots

ABSTRACT

Amyotrophic Lateral Sclerosis (ALS) is a devastating disease and the most common neurodegenerative disorder of young adults. ALS patients present a rapidly progressive motor weakness. This usually leads to death in a few years by respiratory failure. The correct prediction of respiratory insufficiency is thus key for patient management. In this context, we propose an innovative approach for prognostic prediction based on patient snapshots and time windows. We first cluster temporally-related tests to obtain snapshots of the patient's condition at a given time (patient snapshots). Then we use the snapshots to predict the probability of an ALS patient to require assisted ventilation after *k* days from the time of clinical evaluation (time window). This probability is based on the patient's current condition, evaluated using clinical features, including functional impairment assessments and a complete set of respiratory tests. The prognostic models include three temporal windows allowing to perform short, medium and long term prognosis regarding progression to assisted ventilation. Experimental results show an area under the receiver operating characteristics curve (AUC) in the test set of approximately 79% for time windows of 90, 180 and 365 days. Creating patient snapshots using hierarchical clustering with constraints outperforms the state of the art, and the proposed prognostic model becomes the first non population-based approach for prognostic prediction in ALS. The results are promising and should enhance the current clinical practice, largely supported by non-standardized tests and clinicians' experience.

1. Introduction

Amyotrophic Lateral Sclerosis (ALS) is a neurodegenerative disease characterized by a rapidly progressive muscular weakness. It causes denervation of axial, bulbar and respiratory muscles. This leads to a progressive functional impairment (in general without major cognitive decline [1]), and ultimately death [2]. It has no cure and its causes are yet to be discovered. Maintaining the patients' quality of life is thus of major relevance.

Respiratory complications account for the majority of deaths in ALS. Most patients succumb from hypoventilation with hypoxemia and hypercapnia, often associated with respiratory infection [3]. Predicting the onset of hypoventilation is therefore of major importance, to anticipate timely interventions such as the start of non-invasive ventilation (NIV). NIV was shown to be effective in improving both the quality of life and the survival of ALS patients, in particular in patients without major bulbar muscles weakness [4,5]. There is a number of non-evidence based guidelines to support clinicians in their decision to start NIV in ALS patients. These take into account the clinical observation and the results of respiratory tests, and are based on consensus agreement [4]. However, no criteria is available to indicate the probability of respiratory failure within a defined time interval. In fact, in clinical practice, the decision to start NIV is highly dependent on the clinician's experience, together with NIV acceptance by the patient and caregivers. In this scenario, being able to predict the probability of a particular patient to progress to respiratory insufficiency within a certain period of time (such as before the next visit), would be of great clinical value. This information would have critical implications, regarding prognosis, health-costs and quality of life [6].

Unlike in other diseases such as cancer, cardiovascular diseases, Alzheimer's disease and Parkinson's disease [7,8], the state of the art in ALS relies on population-based approaches such as Kaplan–Meier survival tables and multivariable Cox proportional hazard

* Corresponding author.
E-mail addresses: andre.carreiro@tecnico.ulisboa.pt (A.V. Carreiro), sara.madeira@tecnico.ulisboa.pt (S.C. Madeira).

http://dx.doi.org/10.1016/j.jbi.2015.09.021
1532-0464/

regression models [9]. ALS studies have been tackling two major problems. The first is related to patients' diagnosis, studying the impact of diagnostic delay [10], the heterogeneity in ALS subtypes [11], or the diagnostic relevance of certain clinical features, such as axial muscles weakness [12,13]. The second concerns prognostic predictors, usually associated to ALS survival. The most studied prognostic factors are respiratory tests, such as the forced vital capacity (FVC) [14–18] and the maximal inspiratory and expiratory pressures (MIP/MEP) [14]. Some clinical features have also been identified as critical for prognosis, such as the site of onset (bulbar onset is generally associated with worse prognosis), weight loss and disease duration at diagnosis [17,19,20], the functional decline as assessed by the ALS Functional Rating Scale (ALSFRS) [16–18], muscle strength [15], age [17,19,20] and, possibly, gender [19,20]. The prognostic value of recent tests such as the phrenic nerve motor response [21], the respiratory subscore of the ALSFRS [16], as well as other respiratory tests [16,22], have also been explored.

In this scenario, this work proposes a new prognostic prediction approach allowing to answer a very important clinical question: "Given the patient's current condition (patient snapshot), will he/she be in respiratory insufficiency (RI) after a given period of time (time window)?". Our contribution is thus an innovative prognostic model able to evaluate the patient's current condition and, according to it, infer whether this specific patient will/will not require NIV in a given time window.

We first create patient snapshots using a new strategy based on hierarchical clustering with constraints, outperforming the current method based on pivot dates. We then compute learning examples based on the chosen time windows and use them to build classification models able to predict progression to assisted ventilation. To evaluate the proposed prognostic models, we use clinical data containing respiratory tests and neurophysiological data for 517 ALS patients, followed in the ALS clinic of the Translational Clinical Physiology Unit, Hospital de Santa Maria, Lisbon, during a period over 10 years. Since the medical appointments typically occur with a 90 days time interval, we build models for predicting the need for NIV for three time windows: (a) 90 days, the next medical appointment (short term); (b) 180 days, spanning two medical appointments from the current time (medium term), and (c) 365 days (long term). Promising results, as shown by an area under the receiver operating characteristics curve (AUC) value of approximately 79% for the three time windows, highlight the potential for such prognostic models to predict disease progression in clinical practice.

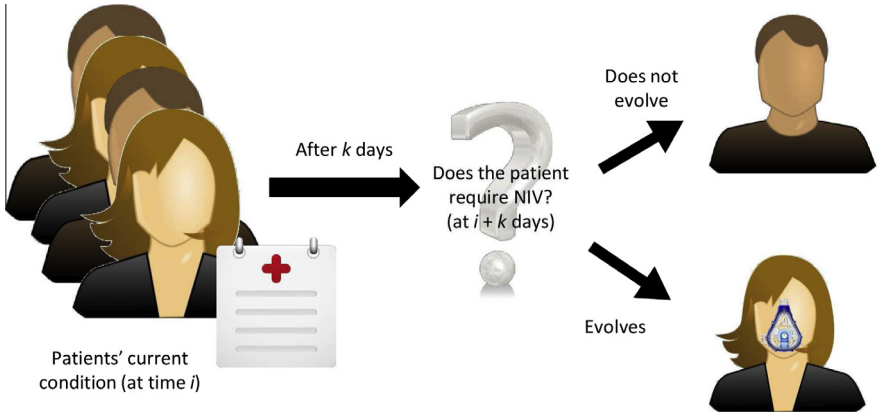

Fig. 1. Problem formulation: can we predict if a given patient will require non-invasive ventilation (NIV) after k days, using the patient's current condition?

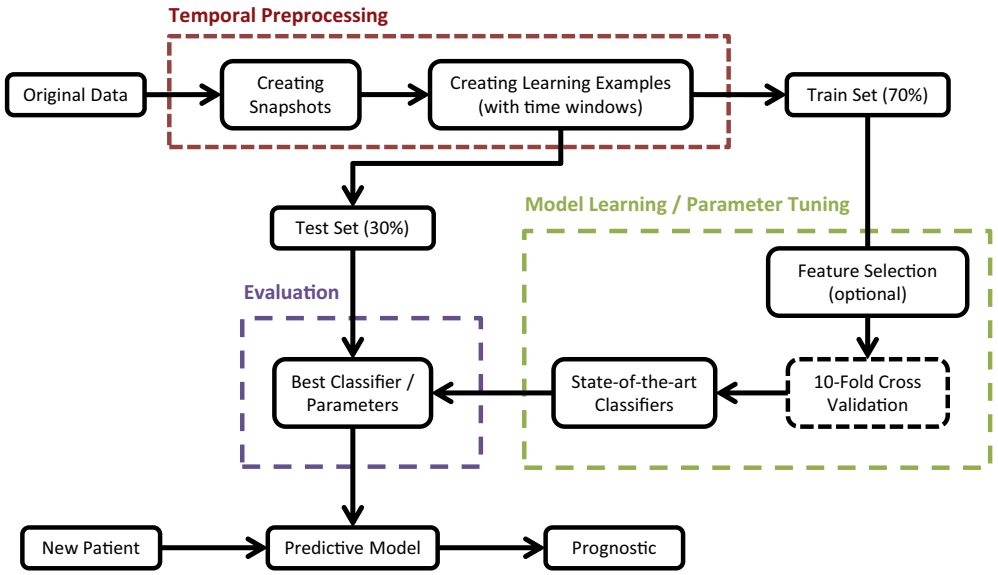

Fig. 2. Workflow of the proposed methodology for ALS prognostic prediction. The model is built based on a stratified 5 × 10-fold cross validation (CV) scheme (5 repetitions with different random seeds) using the training set (70% of patient snapshots). The performance is assessed using the test set (remaining 30%). The train-test stratified partition, as well as the 10 folds for CV, were constructed in order to guarantee that no patient has snapshots in both train and test sets, or in different CV folds.

2. Prognostic models using patient snapshots and time windows

Fig. 1 illustrates the problem addressed in this work: predict whether a given patient will require assisted ventilation k days from time i (time window), using data describing his/her condition at time i (patient snapshot).

Fig. 2 shows the workflow of the proposed supervised learning approach. The original data is transformed by creating patient snapshots and learning examples considering time windows. Then, classifiers are used to predict the evolution of the patient from a stage where he/she does not require NIV (at time i) to a stage where NIV should be started (k days from i). The time k (in days) corresponds to the considered time window. Follows the details of the individual steps.

2.1. Creating snapshots and learning examples considering time windows

2.1.1. Creating patient snapshots by clustering temporally-related tests

The data to be analyzed consists of static information (demographic data) and temporal information (the results of a set of clinical evaluations of specific tests) in the form of multivariate time series. After each appointment, a set of recommended tests is prescribed for the patient. Since he/she is not able to perform all the necessary tests in a single day, we have to deal with their temporal distribution. Therefore, we aim at computing snapshots of the patient's condition by grouping tests performed in a time interval, assumed to be long enough to allow the patient to perform all the prescribed tests.

The majority of researchers follows a simple approach to build snapshots from this type of clinical time series: the use of a pivot date, typically the date associated with a critical event or test. The grouping follows one simple rule: every evaluation of a test held between two pivot dates is grouped into the cluster containing the left-most pivot date. A typical example is to consider the hospitalization date as the pivot date [23]. In this setting, the first test evaluation after this date is included in the snapshot (even if there is more than one evaluation of the same test between consecutive hospitalizations [23]). In this work we use a key test and the date at which it was performed as pivot (see example in Fig. 3, using a key test A* and additional tests B and C). This standard approach to create snapshots has, however, several limitations: a patient may not have any evaluation of the key test, or a given test might have been performed more than once between two pivot dates (or evaluations of the key test). This would mean that many test evaluations would be discarded or that sparser snapshots (with more missing values) would be obtained.

To overcome these drawbacks we propose a new strategy to create the patient snapshots based on bottom-up (agglomerative) hierarchical clustering (HC) with constraints, where a single-linkage metric is used. The constraints are straightforward to implement, and not disease-specific: (1) two evaluations of the same test cannot belong to the same snapshot, since they are part of two different batches of tests; and (2) all tests in a snapshot must be coherent in terms of a given feature of interest (in this case, the NIV requirement). In our case, we know, at the time of each test, whether the patient already required NIV or not (we have the date of NIV start). Thus all tests in a given patient snapshot must have the same NIV status. Under these constraints, validated by the clinicians, we compute a single feature representing the NIV status for the snapshot which has two possible values: 1 and 0, representing, respectively, the requirement/no requirement of NIV. The constraints are verified at cluster merging: the two closest clusters are only merged if all constraints are met.

Fig. 4 shows an example where, although test A_0 (test A evaluated at time 0) and test B_1 are the first candidates for merging, this does not happen since the NIV status is different. Thus, the first cluster consists of tests $\{B_1, C_3\}$, followed by the cluster $\{B_1, C_3, A_8\}$, including the second evaluation of test A, at time 8. These results obtained with HC are more consistent than those obtained when the approach with pivot dates is used. This would yield the cluster $\{A, B, C\}$, with different NIV status for tests in the snapshot, when considering A the key test.

After the HC step, the resulting dendrogram can be cut at different levels, returning different sets of clusters (snapshots). The cutting level is directly related to the considered snapshot duration (the time interval between the snapshot's initial and final tests). Algorithm 1 shows the pseudo code of our approach. Fig. 5 shows an example of the output of creating snapshots from the original data.

Algorithm 1. Hierarchical clustering with constraints to create patient snapshots

```
Input: patient.tests; /* tests for a given patient with
  evaluation date and NIV status */
  maxDuration; /* maximum snapshot duration
  (dendrogram cut-level) */
Output: clusters; /* computed patient snapshots
  (clusters of tests) */
clusters ← init(patient.tests) /* initialize each cluster
  with a different test evaluation (one cluster –
  one test evaluation) */;
numberOfCandidatePairsToMerge = −1
while clusters.size > 1 AND
  numberOfCandidatePairsToMerge ≠ 0 AND
maximumDuration(clusters) ⩽maxDuration /* there are more
  clusters to merge (and they are valid candidate
  pairs), and the maximum snapshot duration was not
  yet reached */ do
    parwiseDist ← pairwiseDistance(clusters); /* compute
     pairwise distance between clusters: single-linkage
     between dates in clusters */
     numberOfCandidatePairsToMerge ← number of pairs of
     clusters;
     mergeDone ← FALSE;
   while numberOfCandidatePairsToMerge > 0 AND
      mergeDone = FALSE do
        candidatePair = (cluster1, cluster2) ← closestPair
      (pairwiseDist);
      if cluster1.tests ≠ cluster2.tests AND cluster1.NIV _ status=
        cluster2.NIV _ status then/*the same test cannot be
      more than once in the same snapshot, and the
      snapshot has to be consistent regarding a given
      attribute (NIV status in this example) */
          newCluster ← merge(candidatePair);
          newCluster.date ← median(dates(newCluster.tests));
          clusters.remove(cluster1,cluster2);
          clusters.add(newCluster);
          mergeDone ← TRUE;
      end
      else
          Remove candidatePair from candidate pair list;
          numberOfCandidatePairsToMerge;
        = numberOfCandidatePairsToMerge – 1;
      end
   end
end
return clusters;
```

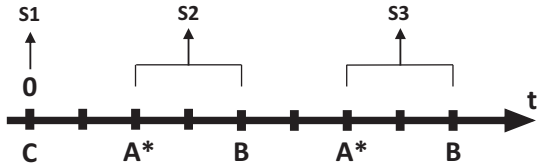

Fig. 3. Creation of patient snapshots using the pivot date approach (in this example we use the evaluation date of key test A∗). *Si* is the *i*th snapshot.

2.1.2. Creating learning examples with time windows

After creating the patient snapshots using the original follow-up data, it is necessary to create the learning examples to be used by the predictive models. These depend on the changes in NIV status between the snapshots in the specified time window of *k* days. Since the aim of this temporal analysis is to predict if, after a certain predefined period of time (such as three months) the patient will be in RI, we created the class *Evolution* (E), with two possible values: 1, the patient initiates NIV within a time window of *k* days; and 0, the respiratory condition does not change in that interval. Fig. 6 illustrates the snapshot labeling process. Clinically, this prognostic approach based on time windows is very relevant, since it allows clinicians to identify the patients with higher risk of developing hypoventilation. In this work, we tested *k* equal to 90, 180 and 365 days (3, 6 and 12 months). These values were validated by the clinicians, and correspond to multiples of the average amount of time between two appointments (∼3 months), as recommended elsewhere [4].

Since the date of NIV start is known, the labeling is performed as follows. A snapshot (with date *i*) belonging to a patient that

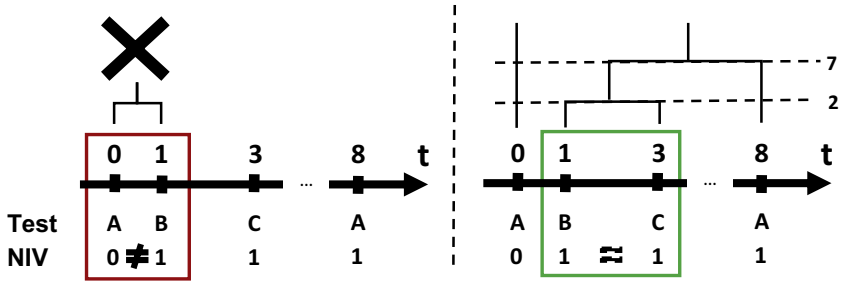

Fig. 4. Example of snapshot creation based on hierarchical clustering with constraints. A, B and C represent different tests evaluated at the time represented in the horizontal axis, and X_i represents the evaluation of test X at time *i*. First candidates for merging are A_0 and B_1, which do not have the same NIV status (left). The merged cluster is thus $\{B_1, C_3\}$ (right). The dendrogram can be cut at different levels, dependent on the considered snapshot duration: 2 time points, yielding snapshots $\{A_0\}$, $\{B_1, C_3\}$, $\{A_8\}$, or 7 time points, resulting in the snapshots $\{A_0\}$, $\{B_1, C_3, A_8\}$.

Original Data

ID	Date 1	TestA 1	TestB 1	...	Date N	TestA N	TestB N	NIV Date
1	01/01/2010	27	1		10/07/2011	12	0	20/04/2011
2	15/03/2010	30	1		25/06/2011	24	1	Not Applied
				...				
P	21/02/2010	28	1		08/10/2011	13	0	14/08/2011

Snapshots

	ID	Date (Snapshot Median)	Test A	Test B	NIV
Patient 1 Evaluations	1	15/01/2010	27	1	0
	1	20/07/2011	12	0	1
		...			
Patient P Evaluations	P	25/02/2010	28	1	0
	P	16/10/2011	13	0	1

Fig. 5. An example of the output of transforming the original data into patient snapshots by grouping batches of tests together. If a test was not performed in a given snapshot a missing value is generated (empty cell). The output is a data file with one patient snapshot (corresponding to the results of tests held in a given time period), per row. The snapshot date is the median of the dates (in which tests were performed) included in the specific snapshot. The original data includes the date of initiation of non-invasive ventilation (NIV), which is transformed into a binary attribute representing if the patient already required (NIV = 1)/not required (NIV = 0), assisted ventilation at the snapshot time.

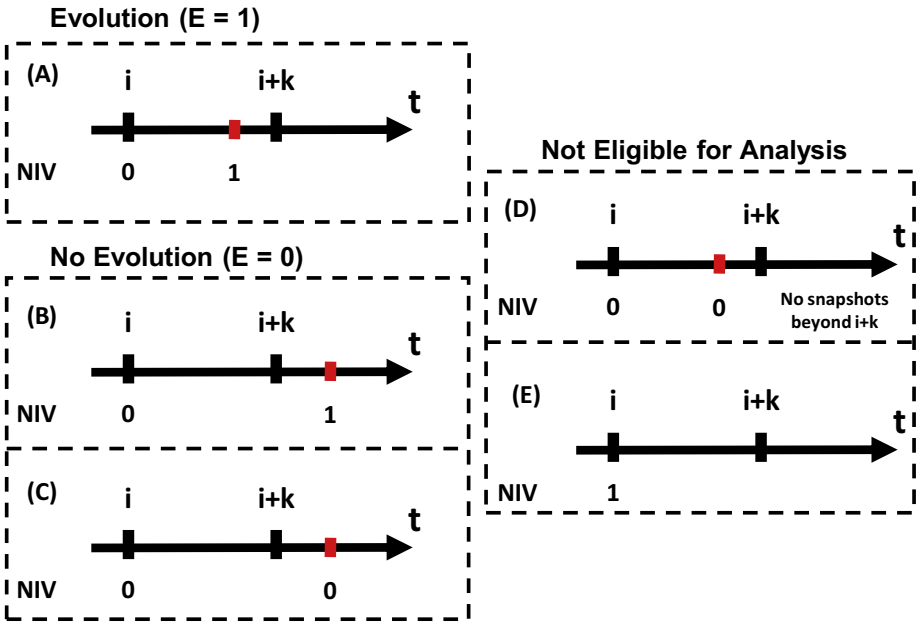

Fig. 6. Definition of class Evolution (E) according to the patient's requirement of NIV in the interval of k days. i is the median date of the snapshot. NIV represents the need (NIV = 1)/not need (NIV = 0) of assisted ventilation at time t.

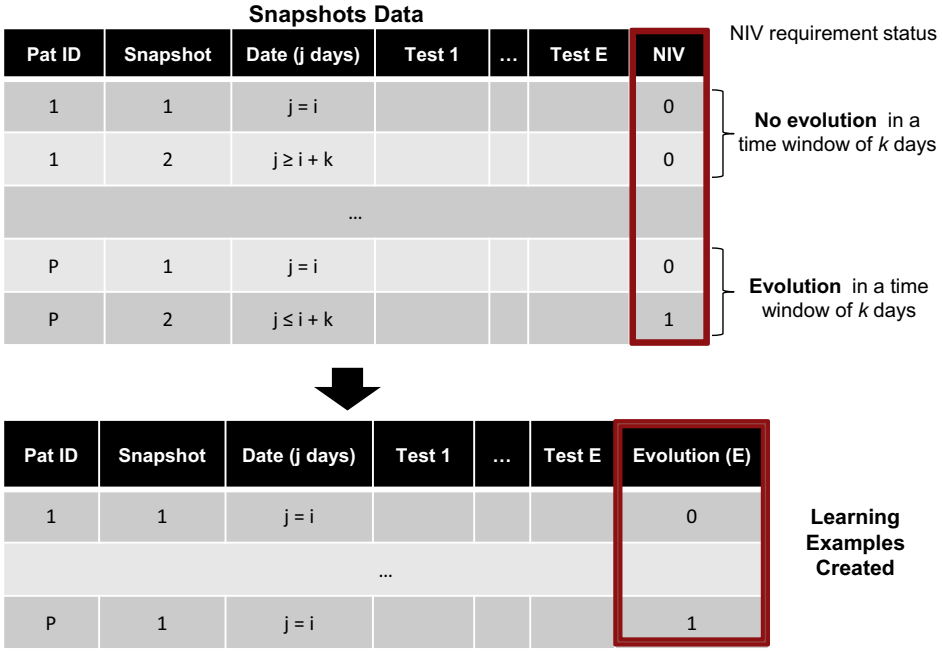

Fig. 7. Creating learning examples. The Evolution class is dependent on the changes of the NIV status in a given time window of k days after i. Following the criteria in Fig. 6: snapshot 1 of patient 1 is labeled with $E = 0$ since snapshot 2 is after the time window of k days and has NIV = 0. Snapshot 1 of patient P is labeled with $E = 1$, since snapshot 2 is within the time window of k days with NIV = 1. Snapshot 2 of patients 1 and P are not eligible because there is no NIV status after these snapshots.

started NIV between i and $i + k$ (inside the temporal window) is labeled with $E = 1$ (situation A). The snapshots where NIV starts in a snapshot after $i + k$ days (outside the temporal window) are labeled with $E = 0$ (situation B). In case the patient never started NIV (all snapshots have NIV = 0), their snapshots are labeled with $E = 0$, provided there is at least one snapshot after $i + k$ days (situation C). The snapshots with no information of NIV status after $i + k$ days (situation D) are not eligible for further analysis. In this situation it is impossible to ensure that there was/was not progression in the respiratory condition at the end of the time window. Finally, the snapshots for patients that at time i already needed NIV (situation E) are also not eligible since the only relevant predictions occur when the patient still has not started NIV (NIV = 0 at time i). Fig. 7 shows an example of creating learning examples with time windows. Note that such an approach based on time windows and using the changes of a given patient's condition can be applied in many different clinical problems, especially for those where patient follow-up is crucial, such as in neurodegenerative diseases.

2.2. Learning the predictive model

The previous step resulted in one dataset of learning examples for each of the time windows considered (90, 180 and 365 days, in our case). The designed experimental setup (Fig. 2) consists in using each of these datasets as input to different classifiers, while using a stratified 5 × 10-fold cross validation (CV) scheme [24] with the training set (70% of total examples). We used classifiers available in Weka [25], including the kNN with `IBK` implementation, Naïve Bayes (NB), Decision Tree (DT) with `J48` algorithm as well as Random Forest (RF), Logistic Regression (LR), and Support Vector Machines (SVM) using `SMO` implementation with polynomial (SVM P) and Gaussian (SVM G) kernels.

We note that, for each of the 5 repetitions, the 10 CV folds were created so that different snapshots from the same patient were not present in both test and training folds. This was also taken into consideration when partitioning the data into stratified training and test sets, using the proportion 70%/30%. We have also performed a grid search to find the best set of parameters for each classifier. Table 1 shows the parameters and corresponding ranges considered for different classifiers. We chose the parameters associated to the best average AUC across the 5 × 10-fold CV classification results for each classifier.

We resorted to two feature selection (FS) strategies to reduce the set of features before training the classifiers [26,27] (applied only on the training set): minimum redundancy maximum relevance (mRMR) [26], from the "filter" family of FS algorithms, and a "wrapper" approach where the selected features are chosen based on the classifiers' performance. mRMR sorts the features according to the maximum of the mutual information with the target class minus the mutual information with the previously listed features, to avoid redundancy. To choose the best threshold for feature selection, the performance of three Weka classifiers (kNN, SVM with polynomial kernel and NB) was evaluated for different thresholds. For the "wrapper" approach, we used the method of successive removal (sucRem) of the worst feature, as evaluated by a given classifier. Similarly to what was done for mRMR, we applied three Weka classifiers: kNN, SVM with polynomial kernel and NB. We then validated the selected features with the clinicians, for both mRMR and sucRem. The selected features from SVM were chosen as the most appropriate for this problem. Table 2 shows the selected features in each time window. Some features are consistently selected, such as gender, the functional scale ALSFRS, or the pattern of disease progression such as upper to lower limbs progression. As expected, many respiratory features were selected, including partial concentration of oxygen (SpO2, PO2), maximal inspiratory and expiratory pressures (MIP/MEP), the vital capacity (VC) and forced vital capacity (FVC).

Table 1
Parameters and corresponding ranges tested for different classifiers: Decision Tree (DT), *k*-Nearest Neighbor (kNN), Support Vector Machines (SVM), with Polynomial and Gaussian Kernels, Naïve Bayes (NB), Random Forest (RF) and Logistic Regression (LR).

Classifier	Parameter	Range
DT	Confidence factor	{0.15, 0.20, 0.25, 0.30}
kNN	Number of neighbors (*k*)	{1, 3, 5, 7, 9, 11}
SVM poly/Gaussian	Complexity	$\{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$
SVM poly	Polynomial degree	{1, 2, 3}
SVM Gaussian	Gamma	$\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$
NB	Kernel	{True, false}
RF	Number of trees	{5, 10, 15, 20}
LR	Ridge factor	$\{10^{-9}, 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}\}$

Table 2
Features selected with methods of minimum redundancy maximum relevance (mRMR) and wrapper approach with successive removal (sucRem), for the three time windows. Features include age at onset, sex, body mass index at the first visit (BMI), family history of motor neuron disease (MND), onset form (limb vs. bulbar), El Escorial reviewed criteria for ALS diagnosis, specific pattern of disease progression (Evol. patt.) and first region affected (Env. segment), and the time interval between first symptoms and first visit (1st symptoms). Functional evaluation includes the ALS Functional Rating Scale (ALSFRS), the extended version including a respiratory subscore (ALSFRS-R), and a subscore of this scale (ALSFRSb). Respiratory function parameters include the respiratory subscore of ALSFRS-R (R), forced vital capacity (FVC), vital capacity (VC), maximal inspiratory/expiratory pressures (MIP/MEP), partial gas concentrations (P0.1, PO2 and PCO2), values of mean (or under 90%) oxygen saturation (SpO2mean and SpO2 < 90%) on percutaneous oximetry, the number of depressions in oximetry saturation overnight and its pattern (Pattern), either in absolute (Dips3%, Dips4%) or per hour (Dips/h < 3%,Dips/h < 4%), and motor response amplitude, area and latency on phrenic nerve stimulation (PhrenMeanAmpl, PhrenMeanArea and PhrenMeanLat, respectively).

mRMR			
90 days	Age at onset	Sex	ALSFRS-R
	R	MEP	VC
	SpO2mean	PhrenMeanAmpl	PhrenMeanArea
180 days	Sex	MND	Onset form
	El Escorial	Evol. patt.	R
	MIP	SpO2 < 90%	Pattern
365 days	Sex	MND	1st symptoms
	Onset form	El Escorial	Evol. patt.
	Env. segment	ALSFRS	ALSFRSb
	R	MEP	VC
	P0.1	SpO2mean	SpO2 < 90%
	Dips3%	Dips4%	Dips/h < 3%
	Dips/h < 4%	Pattern	PhrenMeanAmpl
	PhrenMeanArea	PhrenMeanLat	
sucRem			
90 days	Sex	Env. segment	ALSFRS-R
	R	VC	FVC
	PO2	SpO2 < 90%	PhrenMeanAmpl
180 days	Sex	BMI	MND
	1st symptoms	Evol. patt.	ALSFRS-R
	R	MIP	VC
	PCO2	SpO2mean	Pattern
	PhrenMeanAmpl		
365 days	Age at onset	Sex	BMI
	MND	1st symptoms	Onset form
	ALSFRS	ALSFRS-R	MIP
	SpO2mean	PhrenMeanAmpl	

We followed the most standard strategy for missing value imputation (MVI) [28], using a Weka filter responsible for replacing missing values with the mean/mode of the numerical/nominal attributes (already performed internally for classifiers such as the SVMs and LR). In our case, the missing values were 36.49%, 35.78% and 35.41% for the datasets corresponding to 90, 180 and 365 days, respectively.

Since discretization had already proved useful in previous work [29], we studied the impact of using a domain-knowledge data discretization (provided by the clinicians who collected the original data) on the classification performance. Preliminary results with Weka 3.6.9 [25] supervised discretization showed no benefits in classification performance.

Other discretization, FS and MVI techniques can be tested according to the data and problem at hand.

As a baseline for comparison regarding the prognostic model, we assessed the performance of Cox proportional hazard regression models [9], with implementation available in IBM SPSS Statistics 22, Release Version 22.0.0.0, using both the original data, with information regarding the time until each patient started NIV, and the data regarding the three time windows. For the latter setting, we used a maximum time to event corresponding to the time window. For example, if a patient took 120 days from the first snapshot

until NIV was started, for the window of 90 days that snapshot would have a time of 90 days with no event recorded (since at this time NIV was not required yet). On the other hand, if it took 45 days for the patient to initiate NIV, we would register 45 days with recorded event. Since with Cox proportional hazard regression models we can only use the first snapshot for each patient, and given the population based estimation of the survival curve, we maximized the size of the train set by avoiding the 5×10-fold CV, and testing only with the held-out test set. Nonetheless, we note that a preliminary analysis using the model regarding the available time to event information, only returned predicted values for the hazard function for approximately 70% of the patients, resulting in very low AUC (under 50%). For this reason, further supporting the use of temporal windows, we show only the results for the regression models using the three time windows.

3. Results and discussion

In this Section, we first compare the quality of the patient snapshots created using the proposed strategy using HC with constraints, with the standard approach based on pivot dates. Then, we present the results of the stratified 5×10-fold CV [24] scheme in the training set (70% of the original set of snapshots), used to assess the impact on AUC of the proposed approach for creating snapshots and different preprocessing techniques. Follows the results of the prognostic models obtained using data versions 90 d, 180 d and 365 d when applied to the test set (30% of the patient snapshots).

Our results are based on the analysis of a cohort of 517 ALS patients, followed in a single center for over 10 years, containing detailed clinical information and a complete set of respiratory tests. In this population, all patients were evaluated by the same clinician with a standardized approach and the respiratory tests were performed in the same way. As NIV is provided without personal costs and the process of NIV adaptation is tried many times to achieve good tolerance, less than 5% of patients declined NIV, a number we believe has no impact on results. The publically available PRO-ACT database [18] contains medical records of over 8500 ALS patients who participated in industry clinical trials. However, this dataset lacks information regarding NIV and/or RI, which are at the core of the addressed clinical question, as well as important features such as the ones related to phrenic nerve stimulation. Thus, we did not try to use other population database for testing our model, as we are not aware of any other large ALS population dataset with the required characteristics.

3.1. Learning the predictive model

Tables 6–8 show the results of the stratified 5×10-fold CV in the training set for the time windows of 90, 180 and 365 days and different preprocessing techniques. For clarity sake, we show only the results for AUC, sensitivity (the proportion of actual positives ($E = 1$) which are correctly classified) and specificity (the proportion of negatives ($E = 0$) correctly identified as such).

3.1.1. Preprocessing techniques

Techniques such as knowledge-based discretization (Disc), missing value imputation (MVI) and FS (both mRMR and sucRem methods) can have significant impact on both classification performance and model simplification. In what concerns MVI, Table 6 shows that, regarding mean AUC, we obtain in general better performance when using an imputated version of the data (MVI) (recall that both SVM P and SVM G, as well as LR, already perform this imputation internally). In fact, there is a statistical significant improvement in the AUC when using MVI for the three time windows ($p = 0.028$, Wilcoxon Signed-Ranks Test [30]).

When studying the impact of knowledge-based discretization, Table 6 shows that the results of AUC suggest that, in general, the discretization does not improve the classification performance, exception made to the kNN classifier, which also presents an increased specificity (Table 8). In fact, there is no statistical significant difference in AUC when using knowledge-based discretization for the three time windows ($p = 0.664$, Wilcoxon Signed-Ranks Test [30]). For further analyses, and considering the overall results, we chose to proceed with the real-valued version of the dataset.

Regarding FS, the AUC results in Table 6 show that, in general, most classifiers benefit from at least one of the FS methods. If not reflected in metrics such as the AUC, the benefit of FS can be observed in the obtained models, which are simpler due to a reduced set of features. This can be very important in the clinical practice, since the most critical features can have priority for evaluation, and more expensive (although not critical) tests, can be discarded or at least postponed. The mRMR method does not present statistical significant differences from using all features ($p > 0.05$ for all time windows, Wilcoxon Signed-Ranks Test [30]), whereas the sucRem method showed significant improvement in AUC for the time window of 180 days ($p = 0.018$, Wilcoxon Signed-Ranks Test [30]). Considering the overall results, FS can be worth applying, although using a reduced set of features is dependent on the final chosen classifier. The wrapper method (sucRem) seems to perform better in more models.

Finally, when comparing the different classifiers the AUC results make clear that NB is generally better for most preprocessing techniques and time windows. We performed the Friedman test (as suggested by Demšar [30]) in IBM SPSS Statistics 22, Release version 22.0.0.0, concluding that there are statistical significant differences between the AUC values across classifiers. Followed the analysis of pairwise comparisons, with significance values corrected for multiple testing. NB was significantly better than most classifiers ($p < 0.05$), excepting RF and LR, which were also found to be significantly better than DT, KNN, SVM P, SVM G ($p < 0.05$). Nonetheless, NB showed the highest mean rank of the set. Since NB can output a probability value for prognosis for each time window (very useful for supporting clinical decision) and NB is virtually independent of parameters (no prior correction was used, since we use stratified CV and class proportions are considered representative of the population), we chose NB as the best classifier for this problem. LR also showed promising results and studying the returned odds ratios is also possible.

3.2. Creating snapshots: Pivot dates vs. hierarchical clustering with constraints

We analyzed the quality and predictive value of resulting snapshots using both the standard approach based on pivot dates, and the proposed approach based on bottom-up HC with constraints. Table 3 summarizes the statistics regarding the built snapshots, after correcting for outlier values and discarding evaluations with unknown NIV status, which support our expectations. We used a maximum snapshot duration of 100 days, a value chosen based on a preliminary analysis [29], considering the number and quality of the snapshots retrieved (number of missing values, for example). This value was later confirmed by the clinicians as clinically relevant since it corresponds, approximately, to the usual amount of time between appointments for most patients (3 months).

Results show that the number of snapshots is higher when using the pivot dates. So is the number of missing values in the snapshots. This results in sparser snapshots, due to the aforementioned drawbacks of the pivot dates strategy, such as the existence

Table 3
Comparison of statistics regarding snapshots obtained using the standard approach based on pivot dates and our proposed strategy based on bottom-up hierarchical clustering (HC) with constraints. We show the total percentage of missing values (MV), and the percentage of MV in specific attributes with high diagnostic value in the clinical practice, such as the ALS-FRS (Functional Rating Scale) and a respiratory subscore (R).

	Pivot date	HC with constraints
Total # patients	499	506
Total # snapshots	2988	2694
Total % MV	44.13	41.42
ALS-FRS % MV	23.39	15.03
R % MV	28.65	20.86

Table 4
Class distribution for time windows of $k = 90$, 180 and 365 days.

k	Snapshots	Evolution ($E = 1$)	No evolution ($E = 0$)	Not eligible
90	1487	337 (22.66%)	1150 (77.34%)	1207
180	1410	518 (36.74%)	892 (63.26%)	1284
365	1277	754 (59.04%)	523 (40.96%)	1417

Table 5
Data characteristics (number of patients/number of snapshots) for the training and test sets and time windows of 90, 180 and 365 days.

Time window	# Patients	# Snapshots	Evolution ($E = 1$)	No evolution ($E = 0$)
Train set (70%)				
90	311	1037	237 (22.85%)	800 (77.15%)
180	302	988	370 (37.35%)	618 (62.55%)
365	279	891	532 (59.71%)	359 (40.29%)
Test set (30%)				
90	133	450	100 (22.22%)	350 (77.78%)
180	129	422	148 (35.07%)	274 (64.93%)
365	120	386	222 (57.51%)	164 (42.49%)

of similar tests between pivot dates/evaluations of the key test. This also means that we end up with less patients than in our approach, since some patients end up with no valid snapshots (no coherence of the NIV requirement status, for example). These facts show the effectiveness of the new method in addressing some of the issues in grouping sets of temporally-related tests. Note that this method is not exclusive for this problem in ALS.

Table 4 shows the class distribution resulting from the different values of k (window size in days) with snapshots based on HC with constraints. We can observe that while with smaller values of k the value $E = 0$ predominated, for larger k values, the situation is reversed. To understand these distributions, we note that, by increasing the value of k, we increase the width of the considered interval. Thus, it is more likely that the date of NIV start falls within that longer time window. In the limit case of an infinite time window, all patients still alive would eventually require NIV. In such a limit case, the class distribution would be 0% for $E = 0$ and 100% for $E = 1$. Finally, the decrease in the number of used snapshots, as well as the increase in the number of discarded snapshots, can be explained by the fact that the snapshots that have no information after k days are not eligible for this analysis (situation D in Fig. 6). Hence, for a larger time window (higher values of k), the probability of discarding snapshots increases. In the case of an infinite time window, all snapshots belonging to patients who never required NIV are discarded. In practice, this happens only for patients whose follow-up time is short. Table 5 shows statistics of training and test sets for the three time windows considered (snapshots built using HC with constraints).

We also compared the two approaches of creating snapshots in terms of classifier performance. Table 6 shows that the snapshots created using HC with constraints (MVI, given the choice of using this technique) lead to higher AUC values than those returned by the standard approach based on pivot dates (Pivot(MVI)) for all classifiers and all time windows ($p = 0.000$, Wilcoxon Signed-Ranks Test [30]). This further supports our proposed strategy to create patient snapshots, regarding the quality of training data (number of patients and missing values) and classifier performance.

Table 6
AUC (area under the receiver operating characteristics curve) results of stratified 5 × 10-fold cross validation with the train set (70% of data snapshots) for time windows of 90, 180 and 365 days (mean value ± standard deviation). Orig is the original train set snapshots, obtained using hierarchical clustering with constraints. MVI stands for missing value imputation. Pivot is the strategy for creating snapshots based on pivot dates. Disc is the knowledge-based discretization. mRMR is the minimum redundancy maximum relevance feature selection method, and sucRem is the wrapper feature selection method. Classifiers are Decision Tree (DT), k-Nearest Neighbor (kNN), Support Vector Machine (SVM) with polynomial (P) and Gaussian (G) kernels, Naïve Bayes (NB), Random Forest (RF) and Logistic Regression (LR).

	AUC						
	DT	kNN	SVM P	SVM G	NB	RF	LR
90 d							
Orig	64.50 ± 2.30	61.65 ± 0.80	65.66 ± 0.80	64.98 ± 0.90	77.80 ± 0.40	73.37 ± 1.20	78.22 ± 1.1
MVI	66.21 ± 2.60	65.55 ± 1.00	65.66 ± 0.80	64.98 ± 0.90	79.45 ± 0.30	76.27 ± 0.60	78.22 ± 1.1
Pivot (MVI)	59.81 ± 2.00	62.26 ± 2.00	61.60 ± 0.00	60.92 ± 1.00	74.08 ± 0.00	71.97 ± 1.00	74.47 ± 0.3
Disc	62.26 ± 1.20	71.74 ± 0.30	62.89 ± 0.50	64.01 ± 0.90	78.69 ± 0.40	71.77 ± 1.10	75.84 ± 0.8
mRMR	68.79 ± 2.00	74.45 ± 0.60	62.28 ± 0.40	63.80 ± 0.90	79.15 ± 0.20	75.21 ± 0.80	79.88 ± 0.4
sucRem	66.97 ± 0.80	70.47 ± 0.70	59.81 ± 0.60	59.45 ± 0.80	77.89 ± 0.20	76.45 ± 0.20	78.32 ± 0.1
180 d							
Orig	67.94 ± 1.00	59.75 ± 0.90	69.02 ± 0.60	68.82 ± 1.00	77.09 ± 0.30	70.53 ± 1.10	75.70 ± 0.9
MVI	65.17 ± 2.20	65.44 ± 1.00	69.02 ± 0.60	68.82 ± 1.00	78.51 ± 0.40	75.78 ± 0.30	75.70 ± 0.9
Pivot (MVI)	61.70 ± 3.00	64.77 ± 1.00	65.93 ± 1.00	65.73 ± 1.00	74.65 ± 0.00	72.17 ± 1.00	74.18 ± 0.2
Disc	68.49 ± 0.80	71.96 ± 0.80	67.54 ± 0.20	67.53 ± 0.90	77.50 ± 0.20	70.36 ± 0.90	74.76 ± 1.2
mRMR	68.10 ± 1.00	66.17 ± 1.00	64.64 ± 0.50	63.79 ± 0.70	75.16 ± 0.70	69.00 ± 1.10	73.81 ± 0.9
sucRem	65.93 ± 1.20	69.28 ± 1.20	71.14 ± 0.60	70.20 ± 0.60	78.97 ± 0.50	75.91 ± 1.00	78.27 ± 1.0
365 d							
Orig	62.92 ± 1.90	61.27 ± 1.00	65.86 ± 0.70	66.11 ± 0.70	77.80 ± 0.50	70.14 ± 1.20	73.03 ± 0.8
MVI	66.93 ± 2.20	63.50 ± 0.90	65.86 ± 0.70	66.11 ± 0.70	77.52 ± 0.50	75.46 ± 1.20	73.03 ± 0.8
Pivot (MVI)	62.43 ± 2.00	61.80 ± 2.00	62.95 ± 1.00	61.71 ± 1.00	72.64 ± 1.00	69.56 ± 1.00	69.84 ± 1.0
Disc	66.94 ± 1.50	72.67 ± 1.00	67.73 ± 1.20	67.45 ± 1.30	77.43 ± 0.60	70.51 ± 2.20	73.01 ± 1.1
mRMR	66.27 ± 2.10	64.00 ± 0.70	65.99 ± 1.00	67.62 ± 0.60	76.49 ± 0.50	74.25 ± 0.40	73.74 ± 0.4
sucRem	66.40 ± 1.50	69.89 ± 1.30	67.82 ± 0.40	67.88 ± 0.80	76.16 ± 0.50	73.50 ± 1.40	75.80 ± 0.8

3.3. Evaluating the predictive models

After obtaining the final models using the optimized parameters, we evaluated them in the independent test set. These results provide an idea on what to expect from the classifiers' performance when dealing with new, unknown, data. We note that these results were not used to choose the best classifiers or parameters and recall that snapshots from the same patient were guaranteed not to be in both train/test sets. Tables 9 and 10 show these results for the windows of 90, 180 and 365 days.

We compare our proposed models with the baseline method, the Cox proportional hazards regression model, in the three time windows (90, 180 and 365 days). Given that regression models are highly sensitive to the population statistics, we didn't use CV in order to use as much data as possible. Thus, this comparison is made just in regard to the results in the test set. When comparing the prognostic models (especially the NB, RF and LR classifiers) to the Cox proportional hazards regression model, it is clear that a less population-biased approach is more suited to this problem (see Tables 9 and 10 for AUC, sensitivity and specificity).

Table 7
Sensitivity results of stratified 5 × 10-fold cross validation with the train set (70% of data snapshots) for time windows of 90, 180 and 365 days (mean value ± standard deviation). Orig is the original train set snapshots, obtained using hierarchical clustering with constraints. MVI stands for missing value imputation. Pivot is the strategy for creating snapshots based on pivot dates. Disc is the knowledge-based discretization. mRMR is the minimum redundancy maximum relevance feature selection method, and sucRem is the wrapper feature selection method. Classifiers are Decision Tree (DT), *k*-Nearest Neighbor (kNN), Support Vector Machine (SVM) with polynomial (P) and Gaussian (G) kernels, Naïve Bayes (NB), Random Forest (RF) and Logistic Regression (LR).

	Sensitivity						
	DT	kNN	SVM P	SVM G	NB	RF	LR
90 d							
Orig	9.87 ± 0.90	32.74 ± 1.90	44.73 ± 1.00	45.49 ± 0.80	54.51 ± 0.90	9.87 ± 0.90	40.34 ± 1.7
MVI	56.33 ± 3.20	6.67 ± 0.50	44.73 ± 1.00	45.49 ± 0.80	48.44 ± 0.80	38.82 ± 2.90	40.34 ± 1.7
Pivot (MVI)	33.60 ± 2.00	11.88 ± 1.00	31.88 ± 1.00	34.46 ± 0.00	34.32 ± 1.00	35.71 ± 2.00	39.47 ± 1.2
Disc	22.03 ± 3.00	15.27 ± 0.80	39.24 ± 1.50	40.34 ± 1.60	49.11 ± 0.60	30.89 ± 2.50	34.68 ± 1.8
mRMR	35.44 ± 1.70	29.03 ± 1.30	29.37 ± 0.70	37.13 ± 2.00	44.73 ± 0.50	40.17 ± 2.30	38.40 ± 1.3
sucRem	37.13 ± 2.40	26.67 ± 1.80	29.45 ± 1.60	30.04 ± 0.80	39.24 ± 0.80	41.86 ± 1.70	34.09 ± 1.1
180 d							
Orig	36.00 ± 3.20	81.30 ± 1.00	53.41 ± 1.10	51.89 ± 2.00	63.73 ± 0.70	44.49 ± 2.40	55.78 ± 1.1
MVI	53.35 ± 1.30	34.65 ± 1.60	53.41 ± 1.10	51.89 ± 2.00	54.49 ± 1.20	56.86 ± 1.60	55.78 ± 1.1
Pivot (MVI)	55.15 ± 3.00	50.05 ± 1.00	59.37 ± 1.00	55.87 ± 1.00	36.10 ± 1.00	58.70 ± 1.00	56.64 ± 1.2
Disc	53.95 ± 2.40	29.41 ± 0.70	51.73 ± 0.80	51.57 ± 1.00	56.16 ± 0.50	54.43 ± 1.00	52.32 ± 0.4
mRMR	50.49 ± 0.90	33.19 ± 1.20	41.30 ± 0.40	39.46 ± 1.40	42.92 ± 0.50	52.38 ± 1.60	48.38 ± 0.9
sucRem	53.03 ± 1.50	41.35 ± 1.30	56.81 ± 0.90	55.84 ± 1.20	53.78 ± 0.50	57.57 ± 0.80	55.73 ± 0.3
365 d							
Orig	84.25 ± 1.90	97.18 ± 0.20	79.96 ± 0.80	80.15 ± 0.60	76.05 ± 0.70	82.86 ± 1.60	77.74 ± 0.9
MVI	75.34 ± 1.50	72.03 ± 0.60	79.96 ± 0.80	80.15 ± 0.60	61.84 ± 0.90	81.80 ± 0.70	77.74 ± 0.9
Pivot (MVI)	78.14 ± 1.00	78.48 ± 1.00	83.54 ± 1.00	85.47 ± 1.00	77.41 ± 0.00	84.55 ± 2.00	79.15 ± 1.0
Disc	72.86 ± 1.20	56.32 ± 0.90	72.86 ± 1.30	73.20 ± 1.50	65.71 ± 1.00	78.35 ± 1.20	74.10 ± 1.0
mRMR	74.59 ± 1.70	70.79 ± 1.30	79.47 ± 0.70	80.68 ± 0.60	64.10 ± 0.80	80.75 ± 1.60	77.89 ± 0.3
sucRem	76.80 ± 2.20	80.49 ± 2.00	84.66 ± 0.40	83.76 ± 0.60	69.32 ± 0.70	80.98 ± 1.10	81.84 ± 0.5

Table 8
Specificity results of stratified 5 × 10-fold cross validation with the train set (70% of data snapshots) for time windows of 90, 180 and 365 days (mean value ± standard deviation). Orig is the original train set snapshots, obtained using hierarchical clustering with constraints. MVI stands for missing value imputation. Pivot is the strategy for creating snapshots based on pivot dates. Disc is the knowledge-based discretization. mRMR is the minimum redundancy maximum relevance feature selection method, and sucRem is the wrapper feature selection method. Classifiers are Decision Tree (DT), *k*-Nearest Neighbor (kNN), Support Vector Machine (SVM) with polynomial (P) and Gaussian (G) kernels, Naïve Bayes (NB), Random Forest (RF) and Logistic Regression (LR).

	Specificity						
	DT	kNN	SVM P	SVM G	NB	RF	LR
90 d							
Orig	98.72 ± 0.40	79.80 ± 0.80	86.80 ± 0.90	84.85 ± 1.20	83.20 ± 0.30	98.28 ± 0.40	92.18 ± 0.2
MVI	83.25 ± 0.20	96.98 ± 0.20	86.80 ± 0.90	84.85 ± 1.20	87.50 ± 0.40	92.95 ± 1.10	92.18 ± 0.2
Pivot (MVI)	87.00 ± 1.00	92.63 ± 1.00	92.15 ± 0.00	88.08 ± 0.00	88.80 ± 0.00	88.48 ± 1.00	89.72 ± 0.8
Disc	94.88 ± 0.80	98.05 ± 0.20	86.73 ± 0.70	87.53 ± 0.70	87.18 ± 0.40	90.47 ± 1.70	92.98 ± 0.3
mRMR	94.00 ± 0.60	95.45 ± 0.20	96.08 ± 0.30	90.35 ± 0.70	89.88 ± 0.30	90.68 ± 0.40	94.28 ± 0.3
sucRem	90.65 ± 0.50	94.05 ± 0.50	91.08 ± 0.50	89.33 ± 0.30	92.08 ± 0.10	89.75 ± 1.00	94.48 ± 0.4
180 d							
Orig	86.41 ± 0.50	33.59 ± 1.90	84.34 ± 1.90	85.57 ± 1.00	78.54 ± 0.40	81.68 ± 0.50	82.75 ± 1.7
MVI	75.47 ± 1.10	83.27 ± 1.80	84.34 ± 1.90	85.57 ± 1.00	84.40 ± 0.50	79.00 ± 1.50	82.75 ± 1.7
Pivot (MVI)	66.63 ± 4.00	72.17 ± 3.00	72.71 ± 2.00	75.84 ± 2.00	86.87 ± 1.00	70.93 ± 1.00	76.20 ± 1.1
Disc	82.88 ± 0.70	91.42 ± 0.40	83.17 ± 1.10	83.33 ± 0.90	80.97 ± 0.60	74.56 ± 0.70	84.60 ± 0.7
mRMR	82.36 ± 1.60	84.76 ± 1.00	87.44 ± 1.00	87.61 ± 1.50	88.09 ± 0.80	73.37 ± 1.10	84.72 ± 1.0
sucRem	77.35 ± 1.70	83.17 ± 1.50	85.31 ± 0.90	84.40 ± 0.40	84.76 ± 0.30	77.83 ± 1.40	84.72 ± 1.0
365 d							
Orig	36.38 ± 3.40	8.86 ± 1.20	51.64 ± 0.80	51.92 ± 0.80	65.40 ± 0.60	43.34 ± 1.80	53.20 ± 1.2
MVI	53.93 ± 3.40	44.90 ± 1.10	51.64 ± 0.80	51.92 ± 0.80	77.72 ± 0.70	52.81 ± 1.70	53.20 ± 1.2
Pivot (MVI)	43.65 ± 2.00	37.25 ± 3.00	42.49 ± 1.00	38.10 ± 1.00	52.28 ± 1.00	42.28 ± 1.00	44.29 ± 0.9
Disc	54.37 ± 3.00	75.93 ± 1.00	62.40 ± 1.80	61.50 ± 1.60	75.10 ± 0.70	48.30 ± 1.90	55.99 ± 0.8
mRMR	55.26 ± 4.90	48.30 ± 0.70	52.42 ± 1.20	54.48 ± 0.90	72.48 ± 1.20	49.92 ± 2.10	53.93 ± 1.1
sucRem	51.36 ± 4.80	51.92 ± 1.70	50.70 ± 0.70	51.75 ± 1.50	70.70 ± 0.80	52.76 ± 2.50	52.37 ± 0.3

As previously mentioned, when the regression model was computed using all the available time to event information, it could only predict a prognosis for about 70% of the test patients. On the other hand, when using models with maximum time of 90, 180 and 365 days, all except one patient were given a predicted value of the hazard function. This is another strong argument in favor of using temporal windows in such problems. Furthermore, the AUC values of the regression models were very low, both using time windows approach and the available time to event.

The main drawback of the Cox proportional hazard regression model when applied to personalized prediction is the fact that it is very sensitive to the population, in the sense that the shape of the hazard function is based solely on the baseline hazard function, which is dependent on the survival curve of the population.

When comparing the results on Tables 9 and 10 to the ones obtained for the training set (Tables 6–8), we can see that, although the standard deviation for the 5 × 10-fold CV with the train set is low, the AUC values for the test set are, in general, close to the mean ± standard deviation interval values obtained for the train set. In fact, some models perform better in the test snapshots, such as SVM P and RF for 90 days, DT, NB, RF and LR for 180 days, and DT, kNN, SVM P, NB and LR for 365 days. These results suggest that model overfitting is reduced, as aimed by using stratified 5 × 10-fold CV to learn and tune the models, and a held-out test set to evaluate performance. Moreover, an AUC of approximately 79% for all time windows for the chosen NB classifier, is promising to predict RI and consequent need of NIV. Although there are non-evidence based guidelines indicating when ALS patients should start NIV [4], when predicting the outcome of respiratory function the clinicians still rely on their experience in analyzing the results of respiratory tests. In this scenario, translation into clinical practice of reliable prognostic models would allow the adjustment of the next visit and promote timely medical interventions by taking into account the risk of respiratory complication in the following months.

Table 9
AUC (area under the receiver operating characteristics curve) results of the prognostic models using the test set (30% of data snapshots) for time windows of 90, 180 and 365 days. Orig is the original test set. MVI stands for missing value imputation. Disc is the knowledge-based discretization. mRMR is the minimum redundancy maximum relevance feature selection method, and sucRem is the feature selection wrapper method. Classifiers are Cox proportional hazard regression model (CoxR), Decision Tree (DT), *k*-Nearest Neighbor (kNN), Support Vector Machine (SVM) with polynomial (P) and Gaussian (G) kernels, Naïve Bayes (NB), Random Forest (RF) and Logistic Regression (LR).

	AUC							
	CoxR	DT	kNN	SVM P	SVM G	NB	RF	LR
90 d								
MVI	53.48	63.11	62.70	67.71	61.29	78.87	77.22	78.43
mRMR		64.06	78.99	60.79	55.43	78.52	81.39	78.28
sucRem		75.05	70.42	62.07	52.00	75.97	77.77	75.62
180 d								
MVI	42.02	67.46	65.64	67.52	68.17	79.11	77.42	77.91
mRMR		68.69	63.26	63.70	60.73	71.33	68.41	73.57
sucRem		71.85	77.30	66.84	64.89	78.77	79.84	78.16
365 d								
MVI	49.53	67.54	69.87	70.42	61.54	78.86	75.29	79.43
mRMR		65.50	66.32	69.04	67.65	77.52	74.26	78.69
sucRem		64.68	70.44	68.59	62.41	76.87	66.62	75.94

Table 10
Sensitivity and specificity results of the prognostic models using the test set (30% of data snapshots) for time windows of 90, 180 and 365 days. Orig is the original test set. MVI stands for missing value imputation. Disc is the knowledge-based discretization. mRMR is the minimum redundancy maximum relevance feature selection method, and sucRem is the feature selection wrapper method. Classifiers are Cox proportional hazard regression model (CoxR), Decision Tree (DT), *k*-Nearest Neighbor (kNN), Support Vector Machine (SVM) with polynomial (P) and Gaussian (G) kernels, Naïve Bayes (NB), Random Forest (RF) and Logistic Regression (LR).

	Sensitivity							
	CoxR	DT	kNN	SVM P	SVM G	NB	RF	LR
90 d								
MVI	40.00	37.00	8.00	50.00	44.00	45.00	33.00	37.00
mRMR		26.00	31.00	27.00	20.00	48.00	47.00	35.00
sucRem		47.00	21.00	33.00	12.00	40.00	38.00	31.00
180 d								
MVI	52.94	52.03	35.81	50.00	65.54	54.05	60.14	57.43
mRMR		47.97	30.41	40.54	41.89	40.54	51.35	48.65
sucRem		60.14	50.00	53.38	65.54	50.68	68.24	56.08
365 d								
MVI	65.67	75.68	72.97	81.08	76.13	63.96	82.43	78.83
mRMR		72.52	73.87	81.98	74.32	64.41	81.98	80.63
sucRem		75.23	74.77	81.08	81.53	68.92	71.17	80.63
	Specificity							
90 d								
MVI	65.29	88.57	96.86	85.43	78.57	86.29	90.86	90.00
mRMR		92.86	94.29	94.57	90.86	86.57	87.43	91.71
sucRem		87.14	95.14	91.14	92.00	88.57	90.00	93.14
180 d								
MVI	35.00	82.12	79.93	85.04	70.80	82.12	78.47	83.58
mRMR		83.94	83.21	86.86	79.56	84.67	80.29	84.67
sucRem		77.74	81.75	80.29	64.23	85.77	77.37	85.04
365 d								
MVI	45.24	49.39	53.05	59.76	46.95	78.05	58.54	60.37
mRMR		64.02	47.56	56.10	60.98	74.39	51.22	58.54
sucRem		49.39	48.78	56.10	43.29	65.85	53.66	56.71

4. Conclusions

The contributions of this work are threefold: a new strategy to cluster temporally-related tests, yielding patient snapshots; a new approach for prognostic prediction using patient snapshots and time windows; and the application of such models to predict disease progression to assisted ventilation in ALS.

Patients usually undergo a set of recommended tests between appointments. These cannot be performed within a very short period of time. Hence, clustering the temporally-related evaluations is crucial. We thus proposed a new strategy to compute patient snapshots based on HC with constraints, resulting in more consistent snapshots of the patients' condition at a given time period. These also yielded improvements in the prognostic performance when compared to a standard approach based on pivot dates. The key clinical question was to predict whether a given patient will progress to a stage where he/she requires NIV in a given time window using his/her own data (patient snapshot).

In this context, we proposed three prognostic models that predict if a patient that can breathe without help will be in need of NIV after 90, 180 or 365 days. In the construction of the prognostic models we assessed the impact of preprocessing techniques such as missing value imputation, knowledge-based discretization and feature selection, using stratified 5 × 10-fold CV in the training set (70% of all instances, or snapshots). Overall results suggested that imputated data can benefit the models while the knowledge-based data discretization did not show improved performance. Regarding FS, the main conclusion was that, even though the results did not improve significantly, the prognostic models obtained were simpler, and thus presented an important advantage, since clinicians can thus prescribe clinical tests according to their weight in the models, as well as their costs. Our models achieved AUC values of 78.87%, 79.11% and 78.86% for 90, 180 and 365 days, respectively, for NB, followed by LR and RF.

The proposed prognostic models using patient snapshots and time windows were shown to have significantly higher performance than a baseline using Cox proportional hazards regression models, which also supported the use of temporal windows in the analysis. This innovative prognostic approach is suited for ALS, especially when taking into account that the great majority of these patients present a very fast progression, which usually leads to respiratory failure in just a few months. It is however applicable in other diseases where follow-up data is available and progression is typically linear.

We stress that once the clinician decides to treat a patient with NIV, that becomes irreversible (the time of use per day generally increases with disease progression), unless the patient is intolerant or refuses treatment [4]. However, the latter cases were quite rare in our cohort (<5%), and were not eligible for this analysis. It seems impossible to compare the results of the proposed prognostic models with the clinicians' anticipation, as it depends on the previous experience and skills, which are not possible to standardize. Moreover, trying to predict the exact time of respiratory insufficiency is far from a realistic scenario, since clinicians are more interested in a predetermined time period, such as the next few appointments.

Although the discussion on the performance of each method is very relevant to the scientific community, it is crucial to consider the interpretability of the models. Future work should include the thorough interpretation of the proposed models, selected features and risk factors, always in close relationship with the clinicians. Returning interpretable prognostic information that clinicians can apply should have major impact in the care of ALS, reducing costs, prolonging survival and improving quality of life.

5. Author contributions

All authors listed are justifiably credited with authorship. Collected data: SP MC; Conceived and designed the experiment: SCM MC PT; Preprocessed the data: AVC, PMTA; Performed the experiments: AVC PMTA; Wrote the paper: AVC; Supervised the work: SCM MC PT; Revised the manuscript: SCM MC.

Conflict of interest

The authors declare that they have no conflict of interest.

6. Summary Table

What was already known on this topic.

1. Amyotrophic Lateral Sclerosis (ALS) is a devastating neurodegenerative disease. Associated rapidly progressive motor weakness usually leads to death in a few years by respiratory failure. Correctly predicting respiratory insufficiency would be a major improvement in the patient management.
2. Related work in ALS is mostly restricted to a population based approach, focusing on the study of common features significantly associated to reduced survival, relying on conventional statistical tests, such as Kaplan–Meier survival tables and multivariable Cox proportional hazard regression models.

What this study added to our knowledge.

1. An innovative strategy is presented to cluster temporally-related tests, thus building snapshots of the patients' condition (patient snapshots). We show the advantages of this approach when compared to a naïve one based on pivot dates. We then propose prognostic models based on patient snapshots and time windows able to perform short, medium and long term predictions.
2. The proposed prognostic models are applied to answer a very important clinical question in ALS: "Given the patient's current condition (patient snapshot), will he/she be in respiratory insufficiency after a given period of time (time window)?". This allows the prediction of disease progression to assisted ventilation for a particular ALS patient using his/her data.

Acknowledgments

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT), under projects UID/CEC/50021/2013, NEUROCLINOMICS: Understanding NEURODegenerative diseases through CLINical and OMICS data integration (PTDC/EIA-EIA/111239/2009), and a doctoral grant SFRH/BD/82042/2011 to AVC.

References

[1] J. Stein, T. Schettler, B. Rohrer, M. Valenti, Environmental Threats to Healthy Aging With a Closer Look at Alzheimer's Parkinson's Diseases, Greater Boston Physicians for Social Responsibility and Science and Environmental Health Network, 2008.
[2] M. Cudkowicz, M. Qureshi, J. Shefner, Measures and markers in amyotrophic lateral sclerosis, NeuroRx: J. Am. Soc. Exp. NeuroTherapeut. 1 (2004) 273–283.
[3] C. Heffernan, C. Jenkinson, T. Holmes, H. Macleod, W. Kinnear, D. Oliver, N. Leigh, M.-A. Ampong, Management of respiration in MND/ALS patients: an evidence based review, Amyotroph. Lateral sclerosis Off. Publ. World Federat. Neurol. Res. Group Motor Neuron Diseas. 7 (2006) 5–15.
[4] P.M. Andersen, S. Abrahams, G.D. Borasio, M. de Carvalho, A. Chio, P. Van Damme, O. Hardiman, K. Kollewe, K.E. Morrison, S. Petri, Efns guidelines on the clinical management of amyotrophic lateral sclerosis (mals)-revised report of an efns task force, Eur. J. Neurol. 19 (2012) 360–375.
[5] S.C. Bourke, M. Tomlinson, T.L. Williams, R.E. Bullock, P.J. Shaw, G.J. Gibson, Effects of non-invasive ventilation on survival and quality of life in patients with amyotrophic lateral sclerosis: a randomised controlled trial, Lancet Neurol. 5 (2006) 140–147.
[6] A. Pinto, Home telemonitoring of non-invasive ventilation decreases healthcare utilisation in a prospective controlled trial of patients with amyotrophic lateral sclerosis, J. Neurol., Neurosurg. Psychiat. 81 (2010).
[7] F. Azuaje, Bioinformatics and Biomarker Discovery: omic Data Analysis for Personalized Medicine, 2010.
[8] P. Villoslada, L. Steinman, S.E. Baranzini, Systems biology and its application to the understanding of neurological diseases, Ann. Neurol. 65 (2009) 124–139.
[9] D.R. Cox et al., Regression models and life tables, J. Roy. Stat. Soc. B 34 (1972) 187–220.
[10] P.K. Gupta, S. Prabhakar, S. Sharma, A. Anand, A predictive model for amyotrophic lateral sclerosis (ALS) diagnosis, J. Neurol. Sci. (2011).
[11] M.R. Turner, J. Scaber, J.A. Goodfellow, M.E. Lord, R. Marsden, K. Talbot, The diagnostic pathway and prognosis in bulbar-onset amyotrophic lateral sclerosis, J. Neurol. Sci. 294 (2010) 81–85.
[12] M. Carvalho, S. Pinto, M. Swash, Paraspinal and limb motor neuron involvement within homologous spinal segments in ALS, Clin. Neurophysiol.: Off. J. Int. Federat. Clin. Neurophysiol. 119 (2008) 1607–1613.
[13] M. Carvalho, S. Pinto, M. Swash, Motor unit changes in thoracic paraspinal muscles in amyotrophic lateral sclerosis, Muscle Nerve 39 (2009) 83–86.
[14] F. Baumann, R.D. Henderson, S.C. Morrison, M. Brown, N. Hutchinson, J.a. Douglas, P.J. Robinson, P.a. McCombe, Use of respiratory function tests to predict survival in amyotrophic lateral sclerosis, Amyotroph. Lateral Sclerosis: Off. Publ. World Federat. Neurol. Res. Group Motor Neuron Diseas. 11 (2010) 194–202.
[15] C. Armon, M.C. Graves, D. Moses, D.K. Forté, L. Sepulveda, S.M. Darby, R.A. Smith, Linear estimates of disease progression predict survival in patients with amyotrophic lateral sclerosis, Muscle Nerve 23 (2000) 874–882.
[16] P. Kaufmann, G. Levy, J. Thompson, M. DelBene, V. Battista, P. Gordon, L. Rowland, B. Levin, H. Mitsumoto, The ALSFRSR predicts survival time in an ALS clinic population, Neurology 64 (2005) 38–43.
[17] K. Kollewe, U. Mauss, K. Krampfl, S. Petri, R. Dengler, B. Mohammadi, ALSFRS-r score and its ratio: a useful predictor for ALS-progression, J. Neurol. Sci. 275 (2008) 69–73.
[18] N. Atassi, J. Berry, A. Shui, N. Zach, A. Sherman, E. Sinani, J. Walker, I. Katsovskiy, D. Schoenfeld, M. Cudkowicz, The pro-act database design, initial analyses, and predictive features, Neurology 83 (2014) 1719–1725.
[19] M. Aguila, W. Longstreth, V. McGuire, T. Koepsell, G. van Belle, Prognosis in amyotrophic lateral sclerosis: a population-based study, Neurology 60 (2003) 813–819.

[20] W. Scotton, K. Scott, D. Moore, L. Almedom, L. Wijesekera, A. Janssen, C. Nigro, M. Sakel, P. Leigh, C. Shaw, A. Al-Chalabi, Prognostic categories for amyotrophic lateral sclerosis, Amyotroph. Lateral Sclerosis 13 (2012) 502–508.
[21] S. Pinto, A. Turkman, A. Pinto, M. Swash, M. de Carvalho, Predicting respiratory insufficiency in amyotrophic lateral sclerosis: the role of phrenic nerve studies, Clin. Neurophysiol.: Off. J. Int. Federat. Clin. Neurophysiol. 120 (2009) 941–946.
[22] A. Czaplinski, A. Yen, S. Appel, Forced vital capacity (fvc) as an indicator of survival and disease progression in an ALS clinic population, J. Neurol. Neurosurg. Psychiat. 77 (2006) 390–392.
[23] J.-H. Lin, P.J. Haug, Exploiting missing clinical data in bayesian network modeling for predicting medical problems, J. Biomed. Inform. 41 (2008) 1–14.
[24] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: IJCAI, vol. 14, 1995, pp. 1137–1145.
[25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software, ACM SIGKDD Explorat. Newslett. 11 (2009) 10.
[26] H. Peng, Fulmi Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. (2005) 1226–1238.
[27] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273–324.
[28] M. Bramer, Principles of Data Mining, Springer, 2007.
[29] P.M. Amaral, S. Pinto, M. de Carvalho, P. Tomás, S.C. Madeira, Predicting the need for non-invasive ventilation in patients with amyotrophic lateral sclerosis, in: ACM SIGKDD Workshop on Health Informatics (HI-KDD 2012), 2012.
[30] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.