Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression

Robert Küffner[1,2,19], Neta Zach[3,19], Raquel Norel[4], Johann Hawe[2], David Schoenfeld[5,6], Liuxia Wang[7], Guang Li[7], Lilly Fang[8], Lester Mackey[9], Orla Hardiman[10], Merit Cudkowicz[11], Alexander Sherman[11], Gokhan Ertaylan[12], Moritz Grosse-Wentrup[13], Torsten Hothorn[14], Jules van Ligtenberg[15], Jakob H Macke[16], Timm Meyer[13], Bernhard Schölkopf[13], Linh Tran[17], Rubio Vaughan[15], Gustavo Stolovitzky[4] & Melanie L Leitner[3,18]

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease with substantial heterogeneity in its clinical presentation. This makes diagnosis and effective treatment difficult, so better tools for estimating disease progression are needed. Here, we report results from the DREAM-Phil Bowen ALS Prediction Prize4Life challenge. In this crowdsourcing competition, competitors developed algorithms for the prediction of disease progression of 1,822 ALS patients from standardized, anonymized phase 2/3 clinical trials. The two best algorithms outperformed a method designed by the challenge organizers as well as predictions by ALS clinicians. We estimate that using both winning algorithms in future trial designs could reduce the required number of patients by at least 20%. The DREAM-Phil Bowen ALS Prediction Prize4Life challenge also identified several potential nonstandard predictors of disease progression including uric acid, creatinine and surprisingly, blood pressure, shedding light on ALS pathobiology. This analysis reveals the potential of a crowdsourcing competition that uses clinical trial data for accelerating ALS research and development.

ALS, also known as Lou Gehrig's disease, is a progressive neurodegenerative disorder affecting upper and lower motor neurons. Symptoms include muscle weakness, paralysis and eventually death, usually within 3 to 5 years from disease onset. Approximately 1 out of 400 people will be diagnosed with, and die of ALS[1,2], and modern medicine is faced with a major challenge in finding an effective treatment[2,3]. Riluzole (Rilutek) is the only approved medication for ALS, and has a limited effect on survival[4].

One substantial obstacle to understanding and developing an effective treatment for ALS is the heterogeneity of the disease course, ranging from under a year to over10 years. The more heterogeneous the disease, the more difficult it is to predict how a given patient's disease will progress and thereby to demonstrate the effect of a potential therapy, making clinical trials especially challenging. In addition, the uncertainty surrounding prognosis is an enormous burden for patients and their families. A more accurate way to anticipate disease progression, as measured by a clinical scale (ALS Functional Rating Scale: ALSFRS[5], or the revised version, ALSFRS-R[6]), can therefore lead to meaningful improvements in clinical practice and clinical trial management, and increase the likelihood of seeing a future treatment brought to market[7,8].

In an effort to address the important issue of variability of ALS disease progression, we took advantage of two tools: a large data set of clinical, longitudinal, patient information and the vast knowledge and new computational approaches obtainable through crowdsourcing.

Pooled clinical trial data sets have proven invaluable for researchers seeking to unravel complex diseases such as multiple sclerosis, Alzheimer's and others[9–12]. With that in mind, Prize4Life and the Neurological Clinical Research Institute (NCRI) at Massachusetts General Hospital created the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT, www.ALSdatabase.org/) platform with funding from the ALS Therapy Alliance and in partnership with the Northeast ALS Consortium. The vision of the PRO-ACT project was to accelerate and enhance translational ALS research by designing and building a data set that would contain the merged data from as many completed ALS clinical trials as possible. Containing >8,600 patients, PRO-ACT was launched as an open access platform for researchers in December 2012.

We turned to crowdsourcing[13] to facilitate an unbiased assessment of the performance of diverse methods for prediction and to

[1] Institute of Bioinformatics and Systems Biology, German Research Center for Environmental Health, Munich, Germany. [2] Department of Informatics, Ludwig-Maximilians-University, Munich, Germany. [3] Prize4Life, Tel Aviv, Israel and Cambridge, Massachusetts, USA. [4] IBM T.J. Watson Research Center, Yorktown Heights, New York, USA. [5] MGH Biostatistics Center, Massachusetts General Hospital, Boston, Massachusetts, USA. [6] Harvard Medical School, Charlestown, Massachusetts, USA. [7] Sentrana Inc., Washington, DC, USA. [8] Latham&Watkins LLP, Silicon Valley, California, USA. [9] Department of Statistics, Stanford University, Stanford, California, USA. [10] Department of Neuroscience, Beaumont Hospital and Trinity College Dublin, Dublin, Ireland. [11] Neurological Clinical Research Institute, Massachusetts General Hospital, Charlestown, Massachusetts, USA. [12] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch/Alzette, Luxembourg. [13] Max Planck Institute for Intelligent Systems, Tübingen, Germany. [14] Institute of Social- and Preventive Medicine, University of Zürich, Zürich, Switzerland. [15] Orca XL Problem Solvers, Amsterdam, the Netherlands. [16] Max Planck Institute for Biological Cybernetics and Bernstein Center for Computational Neuroscience, Tübingen, Germany. [17] Berkeley School of Public Health, University of California, Berkeley, California, USA. [18] ALS Innovation Hub, Biogen Idec, Cambridge, Massachusetts, USA. [19] These authors contributed equally to this work. Correspondence should be addressed to R.K. (robert.kueffner@helmholtz-muenchen.de) or N.Z. (nzach@prize4life.org).

Received 17 January; accepted 23 September; published online 2 November 2014; doi:10.1038/nbt.3051

raise awareness of the research potential of this new data resource. To address the question of the variability in the progression of ALS, a subset of the PRO-ACT data was used before its public launch for an international crowdsourcing initiative, The DREAM-Phil Bowen ALS Prediction Prize4Life. The prize for the challenge was $50,000 to be awarded for the most accurate methods to predict ALS progression. The challenge was developed and run through a collaboration between the Dialogue for Reverse Engineering Assessments and Methods (DREAM) initiative and Prize4Life using the InnoCentive Platform. In this challenge, solvers were asked to use 3 months of individual patient level clinical trial information to predict that patient's disease progression over the subsequent 9 months.

The challenge resulted in the submission of 37 unique algorithms from which two winning entries were identified. Overall, the best-performing algorithms predicted disease progression better than both a baseline model and clinicians using the same data. Clinical trial modeling indicates that using the algorithms should enable a substantial reduction in the size of a clinical trial required to demonstrate a drug effect. Finally, the challenge uncovered several clinical measurements formerly unknown to be predictive of disease progression, which may shed new light on the biology of ALS.

RESULTS

Challenge design and participation statistics

As part of the 7th DREAM initiative, the DREAM-Phil Bowen ALS Prediction Prize4Life (referred to henceforth as the ALS Prediction Prize) solicited computational approaches for the assessment of the progression of ALS patients using clinical trial data from the PRO-ACT data set (Online Methods and **Supplementary Note 1**). The challenge offered a $50,000 award for the most reliable and predictive solutions.

Solvers were asked to use 3 months of ALS clinical trial information (months 0–3) to predict the future progression of the disease (months 3–12). The progression of the disease was assessed by the slope of change in ALSFRS (a functional scale that ranges between 0 and 40). The solvers were given 12 months of longitudinal data for the development and training of algorithms, and were asked to submit their algorithm to be evaluated on a separate data set not available for the development or training of the algorithms.

For evaluation, algorithms were run by InnoCentive on the InnoCentive servers. Algorithms were fed data from the first 3 months of a given patient's participation in a clinical trial. Data from the subsequent 9 months were not supplied. The performance results against a test data set were presented on a leaderboard. Finally, participants were assessed on a third, fully blinded and previously unseen validation set to prevent overfitting. The prize was awarded according to performance on this validation set (**Fig. 1**).

The challenge lasted 3 months, from July 13 to October 15, 2012. It drew 1,073 registrants from >60 countries. Following the challenge, a survey of registrants was conducted. The survey revealed a diverse audience comprising academic (58%) and industry (30%) professionals as well as others (12%). Notably, 80% of the solvers had almost no familiarity with ALS. No fewer than 93% expressed interest in participating in a future challenge (comprehensive survey results appear in the **Supplementary Note 2** and **Supplementary Tables 1** and **2**). However, as is typical for crowdsourcing challenges, only a small fraction of the registrants submitted an algorithm for testing. During the challenge a total of 37 teams submitted an algorithm to be tested through the leaderboard and 10 teams made valid final submissions. In order to be valid, the submitted R[14] code was required to be executable within InnoCentive's system and to predict ALSFRS slopes for all given patients.

Method performance and assessment

We evaluated and compared the ten final submissions provided by the solvers, as well as an eleventh method designed by the challenge organizers. The latter method is referred to as the baseline method, as it was used to establish the baseline performance that best-performing teams would need to outperform. The solvers' methods and the baseline algorithm are described in **Supplementary Note 3**, **Supplementary Figure 1** and **Supplementary Tables 2–8**; the full sets of predictions are provided in **Supplementary Predictions**. As a separate archive (**Supplementary Software**), we provide the source code of six teams that may be used in compliance with the algorithms' own copyright statements. Method performance was assessed by root mean square deviation (r.m.s. deviation) and Pearson's correlation (PC) to compare predictions of the ALSFRS slope against the actual slope derived from the data. Although the r.m.s. deviation can be directly interpreted as estimation error in units of ALSFRS, PC is useful to assess the correct prediction of trends. For visual inspection of the performance see **Supplementary Results 1** and **Supplementary Figures 2** and **3**.

Solver teams and their methods were assessed based on their r.m.s. deviation scores on the separate validation data set (**Fig. 2**), which was crucial to guarantee robustness (for differences between validation and leaderboard performance, see **Supplementary Results 2**, **Supplementary Figs. 4–8** and **Supplementary Table 9**). The top six teams (ranked at positions 1–6) exhibited an r.m.s. deviation lower than the baseline method. The ten solvers used a variety of methodological approaches, but it is interesting to observe that four out of the six top-ranking teams employed variants of the random forest machine learning approach. Two other approaches that ranked at position 1 and 6 were based on Bayesian trees and nonparametric regression, respectively. Simple regression methods (ranks 8 to 10) performed substantially worse than the baseline method. Method 7 is

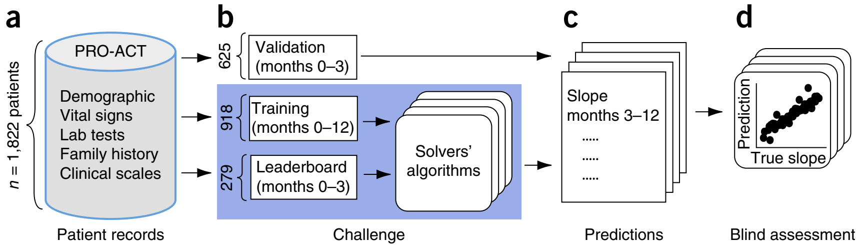

Figure 1 Challenge outline. (**a**) The data for the challenge comprised 1,822 patients from ALS clinical trials from the PRO-ACT data set. Data types included demographics, clinical and family history information, laboratory tests and vital signs. (**b**) We divided the data into three subsets: training data provided to solvers in full, leaderboard and validation data reserved for the scoring of the challenge. Leaderboard and validation data were only available to the challenge managers for the testing of the algorithms submitted by the solvers. Algorithms were fed with data from the first 3 months to perform predictions, and evaluated based on the subsequent 9 months of data. (**c**) At the end of the challenge, solvers submitted their algorithms to be tested by the challenge organizers on the validation data set. (**d**) The predictions obtained in **c** were then assessed by the judges for accuracy.

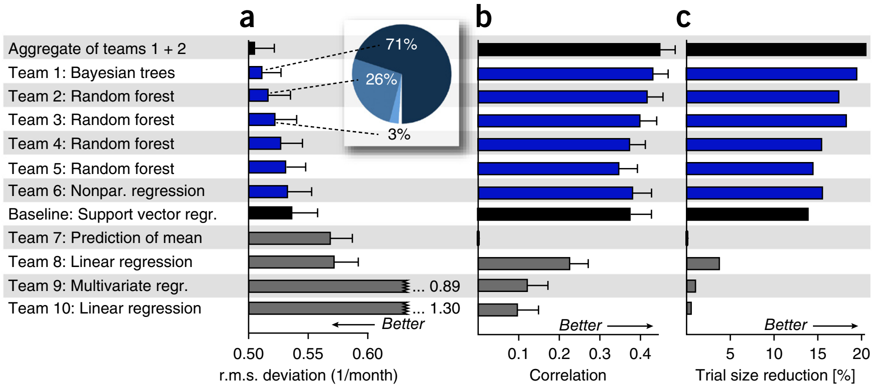

Figure 2 Performance of methods. (**a**,**b**) We compared the approaches of the ten teams that submitted executable R code in the final phase of the challenge and a baseline approach designed by the challenge organizers. All the solvers' algorithms had to be compatible with R version 2.13.1. The teams are numbered according to their ranking in r.m.s. deviation performance (**a**) or Pearson Correlation (**b**). They are colored blue if they performed better than, and gray if they performed worse than, the baseline. In addition, an aggregate of the predictions of teams 1 and 2 is shown. Whiskers indicate bootstrapped s.d. (inset). The frequency with which methods were ranked first is estimated across different bootstrap samples of patients. Teams 1 and 2 were ranked first in 71% and 26%, respectively, of the bootstrap samples (percentage rounded to the nearest integer). (**d**) By simulation, we estimated to what extent clinical trials can be reduced in size by each of the participating approaches corresponding to their improved prediction of disease progression.

a naive predictor that calculates the average slope of the training set patients and predicts the value of this slope for any further patients. We will refer to the resulting deviation as base r.m.s. deviation because it provides a good estimate for the difficulty of prediction of a given patient set, thus achieving a PC of 0. Except for this method, the performance rankings determined using r.m.s. deviation and those determined using PC were quite consistent.

In addition, we employed bootstrapping to assess the robustness of performance (Online Methods). Here, we evaluated the probability that a given method would achieve the best overall performance on different subsets of patients. Teams 1 and 2 achieved the best r.m.s. deviation in 71% and 26%, respectively, of the patient samples generated by bootstrapping (**Fig. 2b**). We thus concluded that the algorithms of these two teams provide both the most robust and the most reliable predictions. Therefore, teams 1 and 2 were declared the best performers of the challenge and received an award of $20,000 each. Team 3 (ref. 15) won a third place prize of $10,000. As in previous installments of the DREAM challenges[16], the aggregation of predictions across teams 1 and 2 further reduced the prediction error. Bootstrapping further allowed us to estimate that a statistically significant improvement over the baseline algorithm would correspond to an r.m.s. deviation of 0.5, which was not achieved by any method.

The two top-performing methods as well as the baseline algorithm were then applied to predict the disease progression of patients in the full PRO-ACT data set. The algorithms maintained their ability to predict disease progression reliably (r.m.s. deviations: team 1, 0.544; team 2, 0.559; baseline, 0.559; **Supplementary Results 3** and **Supplementary Figs. 9** and **10**). That performance was slightly lower than during the challenge can be explained by the greater variability in data across a larger number of trials, which also increased the base r.m.s. deviation from 0.566 to 0.610.

Besides the selection of an appropriate predictive approach, performance was also influenced by the processing of the clinical measurements. Static features (those with one value per patient, e.g., gender or age) could be exploited as is. In contrast, the remaining features were 'time-resolved' and could not, therefore, be incorporated directly into standard machine learning frameworks because time points and number of measurements varied between patients. Generally, teams converted each type of time-resolved data per patient into a constant number of static features by applying various statistics. For instance, linear regression was applied to represent a set of measurements by the slope and intercept (e.g., baseline method). Another approach was to select designated measurements as features, such as the minimum and maximum of the values. The latter approach was successfully applied

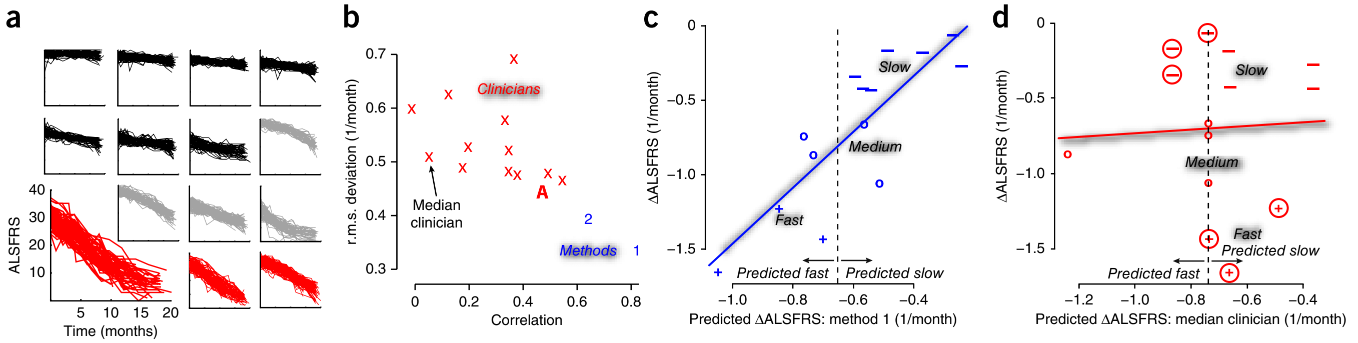

Figure 3 Prediction and classification by algorithms and clinicians. (**a**) ALSFRS slopes were partitioned into 14 clusters of commonly occurring disease progression profiles via k-means. Clusters predominantly contain slow (black), average (gray) or fast patients (red). Patients closest to the center of each cluster were selected to yield 14 representative patients. (**b**) Performance of 12 clinicians (red, "A" indicates the performance of their aggregated predictions) and two algorithms (blue) is assessed based on r.m.s. deviation (ordinate) and Pearson's correlation (abscissa). Here, r.m.s. deviation and correlation are calculated based on the 14 representative patients. (**c**,**d**) The slope predictions (abscissa) as generated by algorithm 1 (**c**) and the median clinician regarding r.m.s. deviation (**d**). Individual patients were classified as slow (–), medium (o) or fast (+) according to the true progression (ordinate). The predicted classifications were assessed relative to a threshold (dashed line). Patients left or right of the threshold are assumed to be predicted fast or slow, respectively. Circles highlight incorrectly classified patients (**d**).

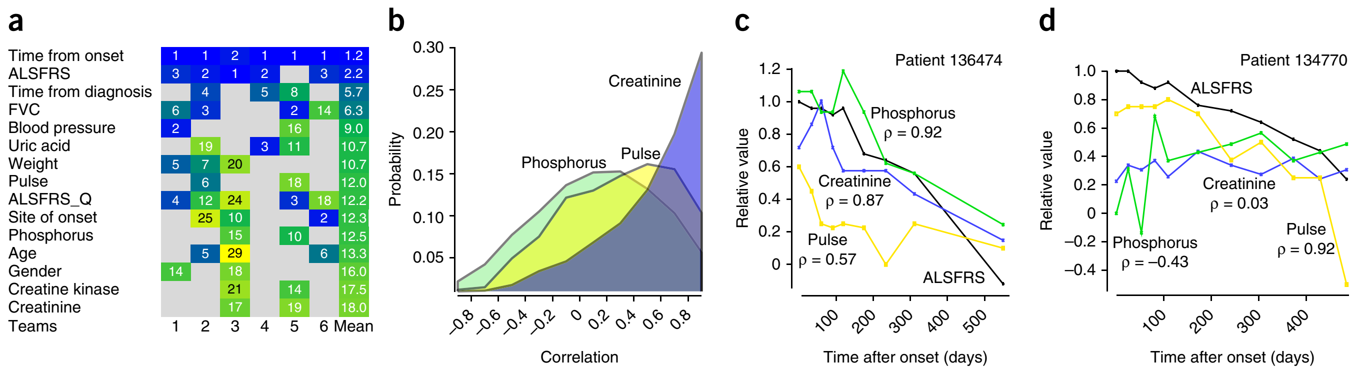

Figure 4 Analysis of predictive features. (**a**) The heat map depicts the features that were identified (ranked from 1–30, illustrated by colors from blue to yellow) by at least two of the algorithms among their 30 most predictive. The column termed mean is the average across the top six solvers. ALSFRS_Q denotes the usage of at least one of the individual ALSFRS questions in contrast to the usage of just their sum, ALSFRS. Three features not previously reported in the literature, phosphorus, creatinine and pulse, are analyzed in the remaining panels. (**b**) The probability of correlation between a feature and ALSFRS distributed across patients. (**c**,**d**) For two example patients, the time progression of these predictive features. Note: to show different measures in one diagram, we normalized them to relative values based on quartiles (Online Methods).

by the best-performing team 1. Notably, this min/max approach represented the time-resolved data in a more robust way than the linear regression approach, which apparently suffered from the relatively few data points available. The treatment of features was also what distinguished the four methods based on random forest variants, that is, they used specific approaches for feature selection, missing value imputation or feature summary statistics.

To be useful clinically, algorithms should be able to maintain predictability with limited data that have either (i) fewer features (ii) or cover less than 3 months. Therefore, we tested the effects of (i) by using only the five most-predictive features and (ii) by limiting the time period of information available to the algorithms to the first month of data. For both the best-performing algorithm and the baseline methods, performance did not substantially deteriorate (**Supplementary Results 3** and **Supplementary Figs. 9** and **10**).

Predictions facilitate reduction in clinical trial size

ALS clinical trials serve to evaluate the effect of a given drug treatment on disease progression, with ALSFRS and ALSFRS-R scores serving as common outcome measures. The great variability in ALS disease progression hinders the ability to detect effects of a given treatment, necessitating larger and more costly clinical trials. The ability to more accurately predict the expected disease progression for a given patient (without treatment intervention) can therefore reduce the number of patients needed by increasing the ability to detect a drug's effect on disease (less obscured by the inherent disease variability). To quantify the potential trial size reduction engendered by use of the algorithms, we simulated trials (Online Methods). We estimated that trial size could be reduced by up to 20.4% using the aggregated predictions of teams 1 and 2 (**Fig. 2d**). As the average cost per patient in an ALS clinical trial is $30,000 (L.A. White and D. Kerr, personal communication), for a phase 3, 1,000-patient trial, this would translate into a $6-million reduction in cost.

Comparison between algorithms and clinicians

ALS prognostic prediction is challenging. Clinicians often feel they lack the necessary tools to provide their patients with accurate prognostic information. Thus, we aimed to evaluate whether the ALS prediction algorithms could help clinicians by comparing their predictive performance.

Therefore, we selected a relatively small but representative subset of patients. Using k-means clustering, we divided the 1,822 patients into 14 clusters, reflecting commonly occurring ALSFRS time courses. Clusters were distinguished by both their intersect (ALSFRS score at time 0) and the shape of their progression curves (**Fig. 3a**). Based on these 14 distinct, disease-progression patterns, we selected 14 representative patients, that is, the centroid of each cluster.

Subsequently, 12 clinicians from top ALS clinics from seven countries were asked to estimate the future disease progression of these representative patients (**Supplementary Results 4**, **Supplementary Fig. 11**, **Supplementary Table 10** and **Supplementary Data 1**), using the exact same data provided to the algorithms. The two best-performing algorithms substantially outperformed all clinician predictions, indicated by both higher PC and lower r.m.s. deviation (**Fig. 3b**). In addition, the algorithms also outperformed the aggregate of all the clinicians' predictions. The rate of progression predicted by the best-performing algorithm (team 1's)—in contrast to the rate predicted by the clinicians—and the actual rate of disease progression across the 14 cases, were well-correlated (**Fig. 3c**). These results suggest that the prediction algorithms might prove useful in helping clinicians to assess patient disease progression.

In addition to an algorithm's predictive accuracy, it is crucial that it provides a broad assessment of patient progression, that is, that it can correctly classify a given patient as having an average disease progression, or having an unusually slow or fast disease course. Therefore, we analyzed both the algorithm- and clinician-based predictions to determine to what extent slow and fast progressors were correctly classified. Three patients showing an ALSFRS slope of less than −1.1 points/month were considered fast, seven patients with a slope greater than −0.5 points/month were considered slow, and the remaining four patients were considered average.

On this limited subset of patients, the best algorithm (team 1's) discriminated perfectly between slow and fast-progressing patients (**Fig. 3c,d**). In contrast, the rate of progression predicted by a typical clinician showed substantially less correlation to the true rate and many patients were misclassified. On average, clinicians misclassified 35% of the patient cases (**Supplementary Results 4**, **Supplementary Fig. 11** and **Supplementary Table 10**).

Predictive features

To assess the importance of each feature for the different algorithms, we looked at the predictive features as they were ranked by the top six best-performing algorithms. We focused on the features that at least two solvers included within the top 30 predictive features. Sixteen such features were identified (**Fig. 4a**). The list includes several features previously reported to predict ALS progression, including time from onset, age, forced vital capacity (FVC), site of onset, gender, weight[17–20], as well as uric acid concentration in blood, a feature that has only been suggested recently as a predictor[20]. In addition, the challenge was successful in identifying nonstandard predictive features, opening the door to new insights into ALS disease mechanisms. These were pulse, blood pressure as well as the concentration of creatine kinase, creatinine and phosphorus.

We further assessed these features by determining the correlation, over time, between the relevant feature and the ALSFRS score, for each subject (**Fig. 4b**–**d**). Notably, for creatinine the distribution of correlations across the patients was skewed toward higher correlations, indicating a subset of patients that exhibit an unusually high correlation between changes in creatinine and ALSFRS score. This was also found, to a lesser extent, for creatine kinase, which is correlated with creatinine. This suggests that these features may be especially predictive for specific subgroups of patients and therefore might be useful biomarkers of the disease. A similar trend was not found for pulse, phosphorus or blood pressure (**Supplementary Results 5** and **Supplementary Figs. 12** and **13**), for which further detailed analysis, beyond the scope of this study, is needed to explore their potential predictive properties.

DISCUSSION

The current lack of robust approaches for estimating the future disease progression of ALS patients represents a major obstacle for the testing of novel therapeutic approaches in clinical trials and the understanding of disease mechanisms. As ALS is a rapidly progressing disease, the accurate estimation of progression is very important for patient care and making decisions regarding clinical interventions and assistive technology.

The unique global challenge presented here brought together the efforts of 37 participating teams to develop tools to predict disease progression in a way useful to ALS clinical trials and clinicians, and to identify new predictive features that can provide new insights into disease processes and could provide important biomarkers.

The PRO-ACT platform, the largest existing ALS clinical trial data set, has provided an unprecedented opportunity to increase our understanding of the ALS patient population and the natural history of the disease[21].

The crowdsourcing approach had several advantages. First, the challenge attracted new minds and new perspectives to a problem largely unknown outside the ALS research community[22]. Second, the format of the challenge allowed blinded side-by-side comparisons of different prediction methods, tested on a data set the solvers never saw and to which the algorithms had never been exposed. This allowed a better assessment of the robustness of the solutions.

Notably, the algorithms could be divided into two groups based on their performance, with teams using tree-based ensemble regression techniques, such as random forest or Bayesian additive regression, almost always outperforming teams using simple regression. These results suggest that tree-based ensemble regression techniques are likely suitable for clinical data in general, beyond the scope of ALS and are therefore of broad general importance for analysis of clinical trial information in the context of clinical trials as well as clinical health records. In addition, robust processing of the time-resolved clinical measurements seemed to be the key to achieving the overall most-predictive results, where simple summary statistics performed best. This may be due to the limited number of time points available and the intrinsic noise, issues prevalent in medical data.

By simulation, we estimated that predictions by the winning algorithms could lead to a 20% reduction in population size for an ALS clinical trial. This reduction stems from changes in trial design. When planning clinical trials, variability between the patients is estimated to plan for a sufficient trial size to capture the effect of the drug beyond that variability. An algorithm that gives more information about the patients, and thereby reduces the interpatient variability, can facilitate a reduction in trial size. Furthermore, a reduction of the estimated magnitude could also affect the number of medical sites required by a trial, leading to further cost savings. Prognostic methods could also lead to improvements in trial effectiveness, so assessing the financial implications of incorporating predictive algorithms into clinical trial design is not straightforward. If we limit the effect of a more quantitative understanding of disease heterogeneity only to calculating the number of patients needed, we estimate that predictions enable a 20% reduction in population size of a phase 3 trial, resulting in a $6-million dollar reduction in costs. These financial benefits need to be weighed against the potential costs of providing a lead-in period where patients are tracked before the start of the trial to determine their expected progression and other factors that affect the trial, such as patient drop-out or limited survival. The finding that the algorithms maintain their predictability using just 1 month of information (**Supplementary Results 3**), and the fact that the performance of the algorithms remained robust when tested on the larger and more diverse full PRO-ACT data set demonstrates the power of crowdsourcing, where a challenge with a monetary award of $50,000 can potentially reduce the costs of multiple future clinical trials by millions of dollars. The algorithms are currently being further tested and validated on proprietary ALS trial data. Furthermore, efforts are underway to transform the algorithms into a ready-to-use software tool for evaluation and future application in clinics and in clinical trials.

To further assess the ability of the winning algorithms to help clinicians in accurately determining the prognosis of their patients, we directly compared the algorithms' predictions with the estimates of 12 leading ALS clinicians. For all 14 patient cases examined, the algorithms outperformed all of the clinicians and the aggregate of the clinicians by a substantial margin. Clearly, predicting disease prognosis based solely on an anonymized data set in the absence of a clinical encounter certainly cannot truly reflect the wealth of information that can be gleaned by an experienced clinician through clinical observation. However, these results demonstrate how a predictive algorithm could prove helpful to clinicians when advising patients. As the patient population participating in clinical trials is not fully representative of the patient population seen in the clinic[23], steps are being taken to directly test the utility of these algorithms in a clinical setting.

Another important goal of the ALS Prediction Prize challenge was to validate features that had previously been suggested to be predictive in small studies and to identify novel predictive features. Overall, 15 different features were identified by more than one solver. Several were features reported in the literature[17,18,24–33], including age, site of disease onset, gender, the slope of disease progression so far, past ALSFRS slope, and past FVC slope, thus serving as validation of both the features and the algorithms. Unfortunately for many of the patients in PRO-ACT, FVC information was not available or other key features required to calculate FVC were missing. Weight, which has been disputed as a predictor in the ALS literature[19,34,35], was found to be predictive. Specific ALSFRS questions were found to be predictive by the different teams, but with no consensus over certain questions being more predictive of the total score than others.

Notably, the challenge also validates the predictive value of uric acid[20]. Higher than average uric acid concentrations have also been shown to correlate with slower progression in Parkinson's disease, dementia and Huntington's disease[36–39], suggesting a common pathological mechanism or by-product across neurodegenerative processes, and clinical trials based on increasing uric acid have already been initiated for several central nervous system–related conditions. The fact that the algorithms still identified uric acid as a predictor supports past findings[20]. Similarly, the challenge was able to support the predictive power of creatinine[40,41], which was not only identified by several different algorithms, but was also found to be highly correlated with the changes in ALSFRS as the disease progresses over time. Further features of interest include creatine kinase, whose levels highly correlate with creatinine levels, suggested to be predictive of prognosis in ALS[42,43]. Pulse and blood pressure, which at first glance may be surprising predictive features for ALS, are supported by a body of literature suggesting sympathetic dysfunction in ALS[44–47]. Further research and studies are needed to assess the potential of these features and elucidate their involvement in ALS pathophysiology.

In summary, the ALS prediction prize brought new minds to the field of ALS and demonstrated the benefits of crowdsourcing in fostering new approaches in ALS research. The best-performing algorithms in this challenge have the potential to reduce the population size needed to measure a drug effect by 20% and have enabled the identification of several nonstandard, potential predictive features that might shed new light on disease pathways. Lastly, the algorithms could aid clinicians in their judgment during evaluation of the patients and thereby improve the care of ALS patients. The algorithms are now being tested in clinical settings.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to the following people for their important assistance with this manuscript: the pharmaceutical companies which provided data to the PRO-ACT data set that enabled this entire endeavor, the Bowen family and D. Lautman for their generous support for this project, R. Betensky from the Harvard School of Public Health for her statistical advice and contributions to the earliest stage of developing the challenge, L. Reinhold from InnoCentive for her support and management of the challenge, our challenge sponsors, *Nature*, *The Economist* and *Popular Science*, the clinicians who participated in the clinicians' assessment and follow-on discussions about the results, L.A. White and D. Kerr from Biogen Idec for their assistance with estimating the financial impact of the algorithms and, of course, the solvers who participated in the challenge and the patients who inspired this effort.

AUTHOR CONTRIBUTIONS

R.K., N.Z., R.N., J.H., D.S., O.H., M.C., G.S. and M.L.L. designed the challenge. R.K. and J.H. prepared the data and baseline algorithm, A.S. helped with data preparation. N.Z. managed the challenge. L.W., G.L., L.F., L.M., G.E., M.G.-W., T.H., J.v.L., J.H.M., T.M., B.S., L.T. and R.V. submitted algorithms. D.S., L.W., G.L., L.F. and L.M. contributed further analysis on challenge performance. R.K. and N.Z. analyzed the results and wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Drigo, D. *et al.* The incidence of amyotrophic lateral sclerosis in Friuli Venezia Giulia, Italy, from 2002 to 2009: a retrospective population-based study. *Neuroepidemiology* **41**, 54–61 (2013).
2. Johnston, C.A. *et al.* Amyotrophic lateral sclerosis in an urban setting: a population based study of inner city London. *J. Neurol.* **253**, 1642–1643 (2006).
3. Kiernan, M.C. *et al.* Amyotrophic lateral sclerosis. *Lancet* **377**, 942–955 (2011).
4. Miller, R., Mitchell, J., Lyon, M. & Moore, D. Riluzole for amyotrophic lateral sclerosis (ALS)/motor neuron disease (MND). *Cochrane Database Syst. Rev.* CD001447 (2007).
5. Cedarbaum, J.M. & Stambler, N. Performance of the amyotrophic lateral sclerosis functional rating scale (ALSFRS) in multicenter clinical trials. *J. Neurol. Sci.* **152**, S1–S9 (1997).
6. Cedarbaum, J.M. *et al.* The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. *J. Neurol. Sci.* **169**, 13–21 (1999).
7. Grossman, I. *et al.* Alzheimer's disease: diagnostics, prognostics and the road to prevention. *EPMA J.* **1**, 293–303 (2010).
8. Tangri, N. *et al.* A predictive model for progression of chronic kidney disease to kidney failure. *J. Am. Med. Assoc.* **305**, 1553–1559 (2011).
9. Cutter, G.R. *et al.* Development of a multiple sclerosis functional composite as a clinical trial outcome measure. *Brain* **122**, 871–882 (1999).
10. Romero, K. *et al.* The coalition against major diseases: developing tools for an integrated drug development process for Alzheimer's and Parkinson's diseases. *Clin. Pharmacol. Ther.* **86**, 365–367 (2009).
11. Ontaneda, D., LaRocca, N., Coetzee, T. & Rudick, R. Revisiting the multiple sclerosis functional composite: proceedings from the national multiple sclerosis society (NMSS) task force on clinical disability measures. *Mult. Scler.* **18**, 1074–1080 (2012).
12. Rogers, J.A. *et al.* Combining patient-level and summary-level data for Alzheimer's disease modeling and simulation: a beta regression meta-analysis. *J. Pharmacokinet. Pharmacodyn.* **39**, 479–498 (2012).
13. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804 (2012).
14. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2013).
15. Hothorn, T. & Jung, H.H. RandomForest4Life: a random forest for predicting ALS disease progression. *Amyotroph. Lateral Scler. Frontotemporal Degener.* **15**, 444–452 (2014).
16. Costello, J. & Stolovitzky, G. Seeking the wisdom of crowds through challenge-based competitions in biomedical research. *Clin. Pharmacol. Ther.* **93**, 396–398 (2013).
17. Magnus, T. *et al.* Disease progression in amyotrophic lateral sclerosis: predictors of survival. *Muscle Nerve* **25**, 709–714 (2002).
18. del Aguila, M., Longstreth, W., McGuire, V., Koepsell, T. & Van Belle, G. Prognosis in amyotrophic lateral sclerosis: a population-based study. *Neurology* **60**, 813–819 (2003).
19. Paganoni, S., Deng, J., Jaffa, M., Cudkowicz, M.E. & Wills, A.M. Body mass index, not dyslipidemia, is an independent predictor of survival in amyotrophic lateral sclerosis. *Muscle Nerve* **44**, 20–24 (2011).
20. Paganoni, S. *et al.* Uric acid levels predict survival in men with amyotrophic lateral sclerosis. *J. Neurol.* **259**, 1923–1928 (2012).
21. Gomeni, R. & Fava, M. The pooled resource open-access, ALSCTC amyotrophic lateral sclerosis disease progression model. *Amyotroph. Lateral Scler. Frontotemporal Degener.* **15**, 119–129 (2014).
22. Lakhani, K.R. *et al.* Prize-based contests can provide solutions to computational biology problems. *Nat. Biotechnol.* **31**, 108–111 (2013).
23. Chiò, A. *et al.* ALS clinical trials: do enrolled patients accurately represent the ALS population? *Neurology* **77**, 1432–1437 (2011).
24. Czaplinski, A., Yen, A.A., Simpson, E.P. & Appel, S.H. Predictability of disease progression in amyotrophic lateral sclerosis. *Muscle Nerve* **34**, 702–708 (2006).
25. Pastula, D.M. *et al.* Factors associated with survival in the national registry of veterans with ALS. *Amyotroph. Lateral Scler.* **10**, 332–338 (2009).
26. Qureshi, M.M. *et al.* Analysis of factors that modify susceptibility and rate of progression in amyotrophic lateral sclerosis (ALS). *Amyotroph. Lateral Scler.* **7**, 173–182 (2006).
27. Fujimura-Kiyono, C. *et al.* Onset and spreading patterns of lower motor neuron involvements predict survival in sporadic amyotrophic lateral sclerosis. *J. Neurol. Neurosurg. Psychiatry* **82**, 1244–1249 (2011).
28. Qureshi, M. *et al.* Medications and laboratory parameters as prognostic factors in amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler.* **9**, 369–374 (2008).
29. Zoccolella, S. *et al.* Analysis of survival and prognostic factors in amyotrophic lateral sclerosis: a population based study. *J. Neurol. Neurosurg. Psychiatry* **79**, 33–37 (2008).
30. Turner, M.R. *et al.* Pattern of spread and prognosis in lower limb-onset ALS. *Amyotroph. Lateral Scler.* **11**, 369–373 (2010).
31. Kollewe, K. *et al.* ALSFRS-R score and its ratio: a useful predictor for ALS-progression. *J. Neurol. Sci.* **275**, 69–73 (2008).
32. Traynor, B., Zhang, H., Shefner, J., Schoenfeld, D. & Cudkowicz, M. Functional outcome measures as clinical trial endpoints in ALS. *Neurology* **63**, 1933–1935 (2004).
33. Vender, R.L., Mauger, D., Walsh, S., Alam, S. & Simmons, Z. Respiratory systems abnormalities and clinical milestones for patients with amyotrophic lateral sclerosis with emphasis upon survival. *Amyotroph. Lateral Scler.* **8**, 36–41 (2007).

34. Paganoni, S., Deng, J., Jaffa, M., Cudkowicz, M.E. & Wills, A.M. What does body mass index measure in amyotrophic lateral sclerosis and why should we care? *Muscle Nerve* **45**, 612 (2012).
35. Reich-Slotky, R. *et al.* Body mass index (BMI) as predictor of ALSFRS-R score decline in ALS patients. *Amyotroph. Lateral Scler. Frontotemporal Degener.* **14**, 212–216 (2013).
36. Ascherio, A. *et al.* Urate as a predictor of the rate of clinical decline in Parkinson disease. *Arch. Neurol.* **66**, 1460–1468 (2009).
37. Auinger, P., Kieburtz, K. & Mcdermott, M.P. The relationship between uric acid levels and Huntington's disease progression. *Mov. Disord.* **25**, 224–228 (2010).
38. de Lau, L.M., Koudstaal, P.J., Hofman, A. & Breteler, M. Serum uric acid levels and the risk of Parkinson disease. *Ann. Neurol.* **58**, 797–800 (2005).
39. Euser, S., Hofman, A., Westendorp, R. & Breteler, M.M. Serum uric acid and cognitive function and dementia. *Brain* **132**, 377–382 (2009).
40. Ikeda, K., Hirayama, T., Takazawa, T., Kawabe, K. & Iwasaki, Y. Relationships between disease progression and serum levels of lipid, urate, creatinine and ferritin in Japanese patients with amyotrophic lateral sclerosis: a cross-sectional study. *Intern. Med.* **51**, 1501–1508 (2012).
41. Chen, X. *et al.* An exploratory study of serum creatinine levels in patients with amyotrophic lateral sclerosis. *Neurol. Sci.* **35**, 1591–1597 (2014).
42. Chahin, N. & Sorenson, E.J. Serum creatine kinase levels in spinobulbar muscular atrophy and amyotrophic lateral sclerosis. *Muscle Nerve* **40**, 126–129 (2009).
43. Rafiq, M., Lee, E., Bradburn, M., McDermott, C. & Shaw, P. Elevated creatine kinase suggests better prognosis in patients with amyotrophic lateral sclerosis. *J. Neurol. Neurosurg. Psychiatry* **84**, e2 (2013).
44. Murata, Y., Harada, T., Ishizaki, F., Izumi, Y. & Nakamura, S. An abnormal relationship between blood pressure and pulse rate in amyotrophic lateral sclerosis. *Acta Neurol. Scand.* **96**, 118–122 (1997).
45. Baltadzhieva, R., Gurevich, T. & Korczyn, A.D. Autonomic impairment in amyotrophic lateral sclerosis. *Curr. Opin. Neurol.* **18**, 487–493 (2005).
46. Kandinov, B., Drory, V.E., Tordjman, K. & Korczyn, A.D. Blood pressure measurements in a transgenic SOD1–G93A mouse model of amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler.* **13**, 509–513 (2012).
47. Pavlovic, S. *et al.* Impairment of cardiac autonomic control in patients with amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler.* **11**, 272–276 (2010).

ONLINE METHODS

General setup. The goal of the challenge was to predict disease progression in ALS. Participants were tasked to use the first 3 months of clinical information (months 0–3 from the beginning of the trial), to predict the change in ALSFRS over months 3 to 12 (with no overlap between the two time periods). ALSFRS drops by one point per month in average patients. We chose to focus on ALSFRS because it is a common primary outcome measure in ALS clinical trials, and a measure that reflects a variety of the patient's everyday function. Other important factors, such as patient survival, were not chosen because survival is a point measure and reflects parameters beyond disease progression (such as the patient's decision regarding tracheotomy and gastrostomy, and the patient's background diseases).

Part of the clinical data was time-resolved (such as ALSFRS, functional vital capacity, vital signs and laboratory tests) and part was static (such as gender, demographics and family history of ALS). The challenge used a subset of the PRO-ACT data set comprising 1,822 patients, derived from four clinical trials, completed in the past 15 years. ALSFRS was shown to be independent of the effective date of trials[48]. All of the subject records used in the challenge included 12 months of completed assessment and contained either ALSFRS or ALSFRS-R scores. Atassi *et al.* calculated basic statistics for the database[49].

The data records were subdivided into a training ($n = 918$), a test ($n = 279$) and a validation set ($n = 625$ unique patient records). For the development and training of computational approaches, solvers received full access to the training data set (challenge phase 1). So that the performance of the solvers' algorithms on the test and validation sets could be assessed, teams had to submit R[14] code to the challenge manager, who then ran the code on the data set, consisting of only 3 months of information. During the course of the challenge, at their discretion, solvers could upload their algorithms to be evaluated by InnoCentive blindly and automatically on the test data and the results were shown on a leaderboard (challenge phase 2). Although the performance on the test set had no influence on the ultimate determination of best performer, the leaderboard served to provide important feedback to the solvers. During the last phase of the challenge (challenge phase 3), the test set was provided to the solvers. They then had to submit their final code, which was again assessed blindly and automatically against the never-before-seen set of subject records comprising the separate validation set.

Assessment. *Computing the actual slopes*. The goal of the challenge was to predict the expected decline of the value of the ALSFRS (slope) during months 3 to 12. ALSFRS is a score composed of ten questions, each contributing a value between 0 and 4. For patients where only ALSFRS-R (12 questions) was available, we discarded two questions not contained in ALSFRS. This was done to simplify the prediction task, although it may remove some information. See **Supplementary Note 1** for more details. Subsequently, we removed ALSFRS and ALSFRS-R values for measurements where not all 10 required questions were available, resulting in a final scale ranging from 0 to 40.

To select subjects eligible for prediction, we removed all subject records with fewer than two visits during the first 3 months. To determine the slope, we assigned the first visit after month three of participation in the clinical trial as *m1*. If there were visits through month 12, we assigned the first such visit after month 12 as *m2*. Otherwise, if there were only visits through the eleventh month we use the final visit as *m2*. If there was no such visit, the subject was removed from consideration. The slope is then calculated as

$$slope = \frac{ALSFRS(m2) - ALSFRS(m1)}{m2 - m1}$$

Performance assessment. Two ALSFRS values were available for each patient, namely s_i from the actual slope and p_i from slope prediction. To assess the prediction performance, the sets of computed slopes S and predicted slopes P were compared across patients using the r.m.s. deviation as well as the Pearson's correlation. The r.m.s. deviation measures absolute deviations between N corresponding slope pairs so that smaller values correspond to a smaller prediction error by

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |s_i - p_i|^2}$$

In contrast, Pearson's correlation ρ is a relative measure that evaluates how well a prediction method is able to recover ALSFRS trends across patients. Better predictions lead to a higher value of the correlation, up to 1.0 for the perfect prediction. It is calculated by

$$\rho_{S,P} = \frac{cov(S,P)}{\sigma_S \sigma_P}$$

where cov is the covariance of S and P and σ_S and σ_P are the s.d. of S and P, respectively.

Aggregation of predictions. Given a variety of different predictive methods, aggregation aims to combine their divergent predictions resulting in a single prediction that is often more accurate than any of the underlying individual ones. Here, we aggregated predictions across teams participating in the challenge. For a given patient, each prediction contributes an estimate of the future ALSFRS slope. The aggregate ALSFRS slope for that patient is then calculated as the average of the ALSFRS values predicted by the algorithms to be aggregated.

Bootstrapping. In the course of this challenge, a single set of predictions was obtained from each participating team, so that the application of one of the above-mentioned scoring schemes results in a single performance estimate per team. In order to approximate the empirical performance distribution for each team, and thus to obtain the expected variance of the performance, we applied bootstrapping. For each team-specific bootstrap, 100 bootstrap samples were generated by sampling (from the validation set, $n = 625$ patients) with replacement, so that some patients appear several times as prediction targets whereas others might be omitted. Subsequently, the performance scoring metric r.m.s. deviation was calculated for each of the bootstrap samples to estimate the bootstrapped average μ and s.d. σ. Thereby, we described improvements over the baseline approach *base* in units of the z-score. We regard the $r.m.s.\ deviation_{thresh}$ at a z-score of $z = 2$ (corresponding to a P-value of 0.02) as a significant improvement in performance.

$$RMSD_{thresh} = \mu_{base} - z\sigma_{base}$$

Clustering and selection of representative patients. We aimed to select a small number of between 10 and 20 representative patients to facilitate the manual estimation of disease progression by clinicians. Therefore, we partitioned the ALS profiles of patients by performing k-means clustering 100 times using $k = 25$, the Euclidian distance metric and randomly placed initial centroids. Out of the 100 k-means runs, the one with the smallest average distance of patients to the centroids was selected and clusters with less than ten members were discarded. One patient was selected from each cluster that was closest to the cluster centroid. K-means was applied to a matrix C of 1,822 patients by 20 months, so that the element $c_{i,j}$ contains the ALSFRS value of patient i at month j. As the time points of obtaining the ALSFRS values did not match across patients, we linearly interpolated the ALSFRS values at the beginning of each month for the construction of C.

Classification of patients. To assess the usefulness of a prediction algorithm, it should be able to correctly classify patients with unusually slow or fast disease courses. Using the following approach we compared the computed slopes to those predicted by algorithms or clinicians to assess their performance regarding patient classification. In a first step, we defined patients with computed ALSFRS slopes of less than −1.1 ALSFRS points/months as fast and greater than −0.5 ALSFRS points/months as slow. Patients in between these extremes were defined as average. In step two, we calculated the median predicted slopes of the average patients. Patients with an ALSFRS of less than or more than this median slope were assumed to be classified as fast or slow, respectively. Subsequently, we obtained the number of classification errors, that is, slow patients misclassified as fast, or fast patients misclassified as slow.

Heatmap of feature correlations. The ranking of the features were provided by the submitting solvers. Only features found by at least two solvers to be in the top 30 of the most predictive features were further analyzed. Different processing variants of the time-resolved features (e.g., average, slope) were

doi:10.1038/nbt.3051

grouped together. To assess the general importance of difference features, the rankings were averaged across solvers.

Calculation of relative values. The various clinical measurements obtained during trials exhibit markedly different ranges and units. We normalize the different measures to a common scale so that several of them can be shown in a single diagram. The normalized measures are referred to as relative values. A measure is normalized by subtracting its Quartile Q1 and dividing by the difference of the quartiles Q3-Q1 to obtain relative values. Thereby, the middle 50% of the measurements of each feature are scaled to a common range.

Estimation of the clinical trial size reduction. Two important parameters may decide whether the effectiveness of a drug treatment can be demonstrated, (i) the number of patients included in a clinical trial and (ii) the accuracy of the estimated disease progression in comparison to the disease progression observed under drug treatment. A reduced trial size leads to an increased variance in treatment effect quantification and thereby reduces the statistical power of tests employed for demonstrating drug effectiveness. Here, we determined the particular trial size reduction such that the resulting increased variance was compensated by the reduced variance stemming from the predictions of the future disease progression. Thus, the required size of clinical trials can be reduced in proportion to the accuracy of these predictions.

For this purpose, we simulated clinical trials for each method, first using the method's slope predictions as covariates in the data analysis and ignoring the predictions in a second simulation. The simulation was performed by fitting linear random effects models, routinely used to describe treatment effects on the progression of the ALSFRS slope in most current clinical trials. The setting of the simulation was a placebo-controlled trial where the treatment, which has unknown clinical effects, was compared to the placebo arm.

The model was parameterized for the time interval between 3 and 12 months, thus modeling the future disease progression after a 3 month lead-in period as assumed during the competition. Suppose $slope_i$ was the slope prediction from the model and $treatment_i$ was a variable which is zero for the control group and one for the treatment group. We then modeled the ALSFRS value a_{ij} determined for patient i during a clinical examination at time point j by

$$\begin{aligned} a_{ij} &= (\beta_1 + b_i) * time_{ij} + \beta_2(time_{ij} * treatment_i) \\ &\quad + \beta_3 * slope_i + \beta_4(time_{ij} * slope_i) + intercept_i + e_{ij} \end{aligned}$$

with the coefficients $\beta_1 \ldots \beta_4$, weighting the terms corresponding to the average slope, the treatment effect, the part of the intercept dependent on the slope, as well as the estimated future disease progression, respectively, as well as the random effect of β_1, b_i, which was included to reduce the overall error e_{ij}. We also estimated the same model except that we removed all the terms containing $slope_i$. Let s_1 be the standard error of the treatment effect β_2 in the model with $slope_i$ and s_2 be the standard error of β_2 without it. Then the percent reduction in sample size is proportional to the reduction in variance given by

$$100 * (1 - (s_1 / s_2)^2)$$

Treatment codes were simulated at random. Due to the large number of patients in the sample, different choices of treatment codes had a negligible effect.

Ethics statement. In all of the trials included in the PRO-ACT data set, study protocols were approved by the participating medical centers and all participating patients gave informed consent. De-identified data from these trials were donated to the PRO-ACT data set for research purposes only and under the explicit conditions that Prize4Life and all users of the data would maintain the anonymity of subjects and not attempt to discover the identity of any subject. In the rare cases where donated data were not already completely anonymized, donated data were further anonymized following the HIPAA de-identification conventions for personal health information: any potential patient initials and/or dates of birth were removed, new randomized subject numbers were created, and wherever possible, trial-specific information was removed in the merging of data sets, including trial center identity and location, trial dates, or other identifying information.

48. Miller, R.G. *et al.* Phase II screening trial of lithium carbonate in amyotrophic lateral sclerosis: examining a more efficient trial design. *Neurology* **77**, 973–979 (2011).

49. Atassi, N. *et al.* The PRO-ACT Database: Design, initial analyses, and predictive features. *Neurology* doi:10.1212/WNL.0000000000000951 (8 October 2014).