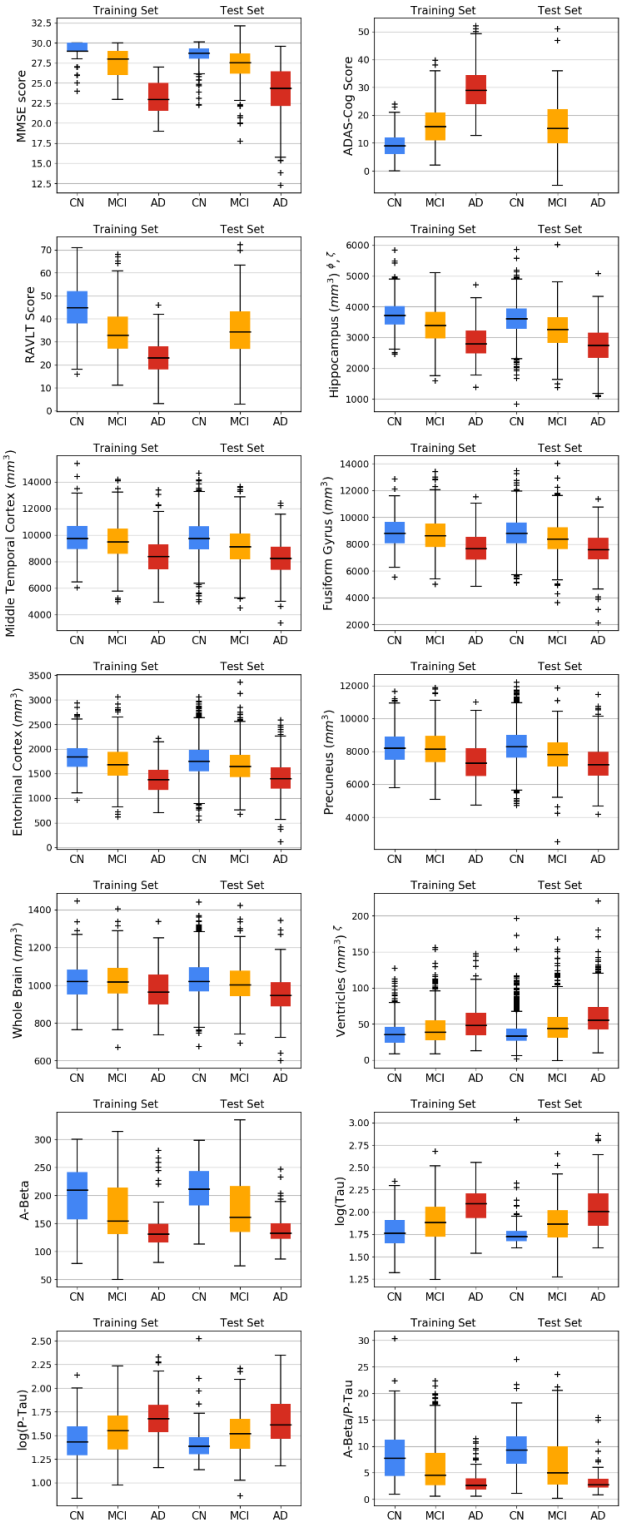
Supplementary Material

Multi-study validation of data-driven disease progression models to characterize evolution of biomarkers in Alzheimer's disease.

Damiano Archetti, Silvia Ingala, Vikram Venkatraghavan, Viktor Wottschel, Alexandra L. Young, Maura Bellio, Esther E. Bron, Stefan Klein, Frederik Barkhof, Daniel C. Alexander, Neil P. Oxtoby, Giovanni B. Frisoni, Alberto Redolfi, for the Alzheimer's Disease Neuroimaging Initiative and for the EuroPOND consortium.

This supplementary material has been provided by the authors to give readers additional information about their work.

SF1. Comparison of biomarkers between groups and data cohorts



Box plots of biomarkers for each diagnostic category (CN, MCI and AD) for subjects from ADNI and Test datasets. Lower and upper hinges of each boxplot correspond to 25th and 75th percentiles of data. Φ: no significant difference (p-value≤0.05) between biomarkers of MCI subjects from training and test datasets; ζ: no significant difference (p-value≤0.05) between biomarkers of AD subjects from training and test datasets.

SS1. Mathematical modelling details

According to EBM and DEBM approaches, each biomarker is considered as either *normal* or *abnormal* and its probabilistic transition from the normal to the abnormal state is defined as *event*. The aim is to define in a data-driven manner the sequence of events that describe the most probable cascade of symptoms that characterize the transition from the healthy state to the full-blown disease spectrum.

SS1.1 Event Based Model (EBM)

In EBM various event sequences are sampled via a Markov Chain Monte Carlo (MCMC) process that aims at the maximization of the likelihood $P(X|S)$ in *SEq.1*:

$$P(X|S) = \prod_{i=1}^{N}\left[\sum_{k=1}^{N_B}\left(P(k)\prod_{j=1}^{k}P(x_{ij}|E_j)\prod_{j=k+1}^{N_B}P(x_{ij}|\neg E_j)\right)\right] \qquad (SEq.1)$$

where X denotes the entire data set, N_B is the number of biomarkers, $P(k)$ is the probability of being at stage k, x_{ij} is the j-th biomarker of subject i and $P(x_{ij}|E_j)$ and $P(x_{ij}|\neg E_j)$ are the likelihoods of measurement x_{ij} given that biomarker j has or has not become abnormal respectively, implying that events $E_1 \dots E_k$ already occurred and events $E_{k+1} \dots E_{N_B}$ still have to occur. The central ordering is therefore the ordering for which $P(X|S)$ is maximum, or equivalently the ordering that best fits X.

Sequences are sampled via an MCMC process where at each Monte Carlo step a new sequence S' is sampled as a random swap between two biomarkers of the benchmark sequence S. If the likelihood of S' is greater than the likelihood of S, then S' is considered as the benchmark sequence for the following MCMC step. The transition to a new state can also happen if the likelihood of S' is less than the likelihood of the benchmark sequence, and in this case the transition occurs with probability:

$$p = e^{P(X|S')-P(X|S)} \qquad (SEq.2)$$

so that event sequences can be chosen hierarchically as benchmark sequences based on their likelihood.

The normal and abnormal states for each biomarker are defined by a gaussian mixture model (GMM), where the populations of CN and AD subjects are described respectively by the normal distributions $N_j(\mu_{CN}, \sigma_{CN})$ and $N_j(\mu_{AD}, \sigma_{AD})$. To avoid the possibility that biomarkers will not show a clear bimodal distribution, the standard deviations of $P(x|E_j)$ and $P(x|\neg E_j)$ must be less or equal to the distributions of biomarkers from AD and CN subjects respectively. The mixture model distribution for each biomarker j is then found as the distribution that minimizes:

$$C_j = \sum_{i=1}^{N}\log\left(\vartheta_j P\left(x_{ij}\middle|E_j, N_j(\mu_{AD}, \sigma_{AD})\right) + (1-\vartheta_j)P\left(x_{ij}\middle|\neg E_j, N_j(\mu_{CN}, \sigma_{CN})\right)\right) \qquad (SEq.3)$$

along $\mu_{AD}, \sigma_{AD}, \mu_{CN}, \sigma_{CN}$ and ϑ_j. The summation in *SEq.3* is intended over all subjects. The parameter ϑ_j is a mixing parameter between 0 and 1 weighs the CN and AD distributions for the j-th biomarker.

SS1.2 Discriminative Event Based Model (DEBM)

The approach of DEBM model for the calculation of the central ordering, on the other hand, is a two-step process where first (i) a specific ordering S_i is calculated for each subject by sorting the posterior probability that biomarker x_{ij} has become abnormal and then (ii) computing the central ordering S as the event sequence that minimizes the sum of modified Kendall's tau distances (see *SS1.4*) between itself and all the subject-wise orderings S_i

As the posterior probability is influenced by the physiological variability of biomarkers, DEBM assumes that single subject orderings S_i are noisy estimates of the central ordering S.

An initial estimate of the distributions of non-diseased and diseased subjects for each biomarker is performed using values from subjects at the very opposite sides of the disease spectrum, as defined by a Bayesian classifier which is trained to remove outliers and wrongly labelled data. It generates truncated Gaussian functions that neglect the tails of the distributions of the two populations, thus reducing the value of the standard deviations of the two distributions with respect to those of the whole population distribution. This allows to separate efficiently the two normal distributions for the j-th biomarker $N_j(\mu_{CN}, \sigma_{CN})$ and $N_j(\mu_{AD}, \sigma_{AD})$. With this method the resulting distributions are biased estimates of the expected distribution, characterized by smaller variance and a mean that is greater than the expected one for the distribution with the larger mean and minor for the distribution that has the smaller mean.

The biased distributions are then refined including data from all subjects via a GMM that has constraints based on the aforementioned relationships between the expected and the biased distributions, where the objective function for optimization of biomarker j is the same as that for EBM (*Eq.3*). The optimization of C_j is performed by alternatively optimizing the gaussian parameters $\mu_{CN}, \sigma_{CN,}\mu_{AD}, \sigma_{AD}$ and the mixing parameter ϑ_j until the latter converges. The mixing parameters of the Bayesian classifier are used as prior probabilities for the class they represent, i.e. pathological or non-pathological.

SS1.3 Subject staging

Specific methods for staging subjects on the basis of the event sequences are available in both EBM and DEBM original formulations. For the sake of simplicity, and in order to have a common staging system for both models, the method from EBM was employed in this work. This method stages each subject on the central event sequence, with the inclusion of stage 0 where no biomarker is abnormal, and assigns each individual the stage σ_i defined as:

$$\sigma_i = \mathrm{argmax}_k P(X_i|S,k) = \mathrm{argmax}_k P(k) \prod_{j=1}^{k} P(x_{ij}|E_j) \prod_{j=k+1}^{N_B} P(x_{ij}|\neg E_j) \qquad (SEq.4)$$

The stage σ_i is the k-th step of the optimal sequence S that maximizes the probability that all events up to k already occurred for subject i and events from $(k+1)$ to N_B have not occurred yet given

the biomarker set X_i. In case of a missing biomarker the probability of the biomarker to be in abnormal state was set to ½ .

SS1.4 Modified Kendall's Tau distance

Traditional Kendall's Tau distance is often used in order to measure quantitative differences between sequences, and it can be defined as:

$$K(S,S') = \sum_{n=1}^{N-1} V_n(S,S') \qquad (SEq.5)$$

Where S and S' are the two sequences, N is the total number of events and $V_n(S,S')$ is the number of adjacent swaps needed so that event n of sequence S is at the same position in both sequences. In a nutshell, Kendall's tau distance computes the total number of adjacent swaps that are needed to transform the sequence S' into sequence S. In DEBM model the estimates of the individual sequences are based on rankings of posterior probabilities of biomarkers being abnormal, therefore it is convenient to define a model version of Kendall's Tau distance that takes into account of posterior probabilities where swaps between events for which the difference of probability is large is penalized. In this case the number of swaps $V_n(S,S')$ is replaced by $\hat{V}_n(S,S')$ that is computed as:

$$\hat{V}_n(S,S') = \sum_{l=n+1}^{k} p_n - p_l \qquad (SEq.6)$$

Where n and k denote the positions of the same biomarker in S and S' respectively and p_i denotes the probability that biomarker at position i has become abnormal. After $\hat{V}_n(S,S')$ has been calculated for a single biomarker S' is updated by swapping event at position k with event at position n.

ST1. EBM and DEBM comparative table

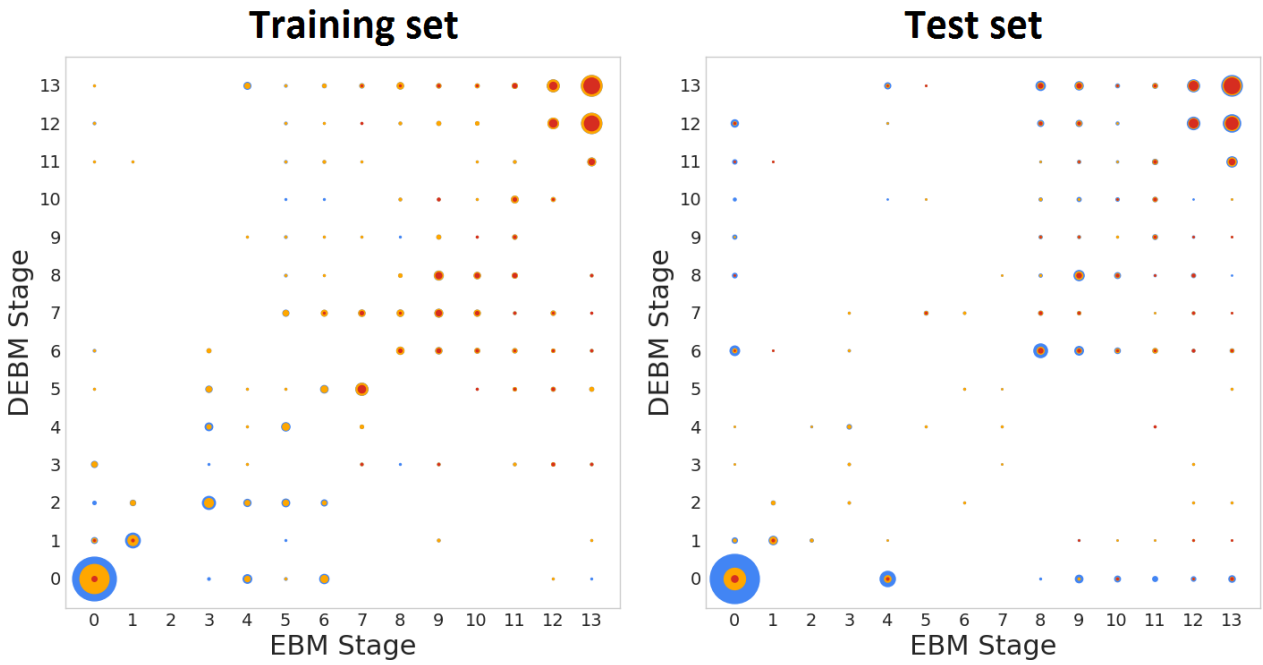
FEATURES	EBM	DEBM
Input files	-training data set (.csv format) -test data sets (optional, .csv format)	-training data set (.csv format) -test data sets (optional, .csv format)
Outputs	-optimal ordering of biomarkers -mixture model biomarker distributions -subject staging (optional)	-optimal ordering of biomarkers -mixture model biomarker distributions -event proximity -subject staging (optional)
Mixture model specs	-in order to ensure bi-modality the standard deviations of the distributions of events are bounded to be less or equal than the standard deviations of distributions of biomarkers of AD and CN subjects	-mixture model is built starting from distributions from easy controls and easy diseased subjects -optimization of model parameters performed alternatively between gaussian parameters and mixing parameter
Assumptions	-optimal ordering is the sequence that best fits the training data set	-optimal ordering is the average of subject-wise optimal sequences
Strengths	-based on cross sectional data sets -no a priori assumptions about biomarkers distributions -method explored and validated in 8 peer-reviewed works -possibility to define abnormality cut points for biomarkers	-based on cross sectional data sets -no *a priori* assumptions about biomarkers distributions -ease of calculation and short computation times (10 minutes for calculation of a sequence from 1500 subjects) -estimation of distance between events
Weaknesses	-long computation times -may incur in overfitting	-Novel approach that requires further validation
Algorithm validation	-performed only on well curated research data sets (ADNI, MAGNIMS, GENFI, TRACK-HD) and on synthetic data	-performed only on ADNI and on synthetic data

Acronyms: csv: comma-separated values; ADNI: Alzheimer's Disease neuroimaging Initiative; MAGNIMS: Magnetic Resonance in Multiple Sclerosis; GENFI: GENetic Frontotemporal dementia Initiative; HD Huntington's Disease.

SF2. Staging of subjects from single cohorts

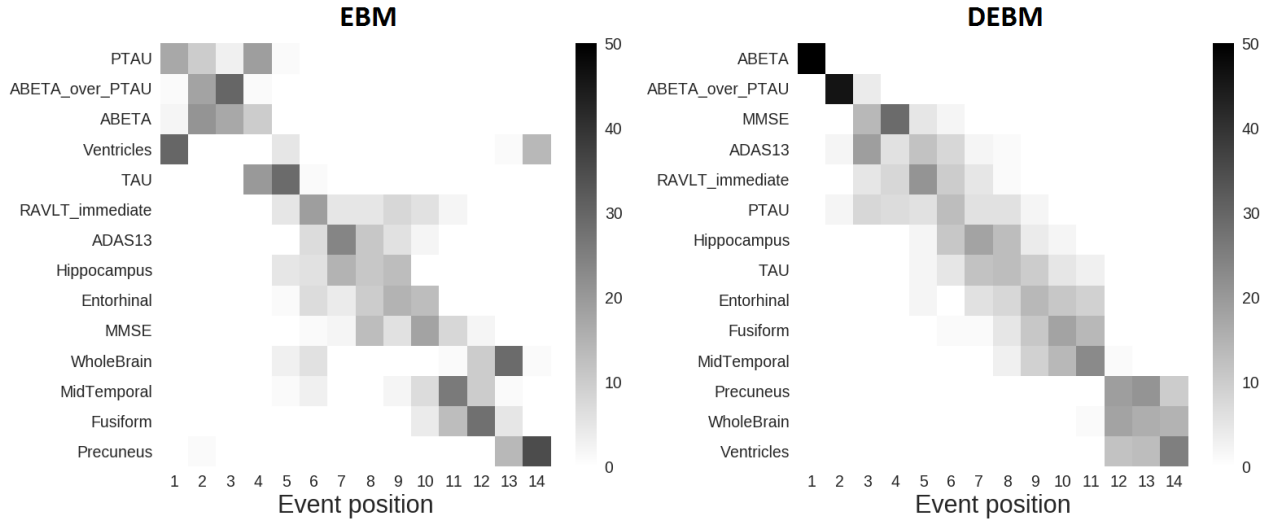
Staging based on the sequences obtained with EBM and DEBM for subjects of each test cohort. Staging of subjects from all diagnostic categories (Cognitively normal (CN) in blue, mild cognitive impairment (MCI) in orange, and Alzheimer's disease (AD) in red) are shown for the cases of (a) ADC subjects on EBM sequence; (b) ADC subjects on DEBM sequence; (c) ARWiBo subjects on EBM sequence; (d) ARWiBo subjects on DEBM sequence; (e) EDSD subjects on EBM sequence; (f) EDSD subjects on DEBM sequence; (g) OASIS subjects on EBM sequence; (h) OASIS subjects on DEBM sequence; (i) ViTA subjects on EBM sequence; (j) ViTA subjects on DEBM sequence.

SF3. EBM stage vs DEBM stage



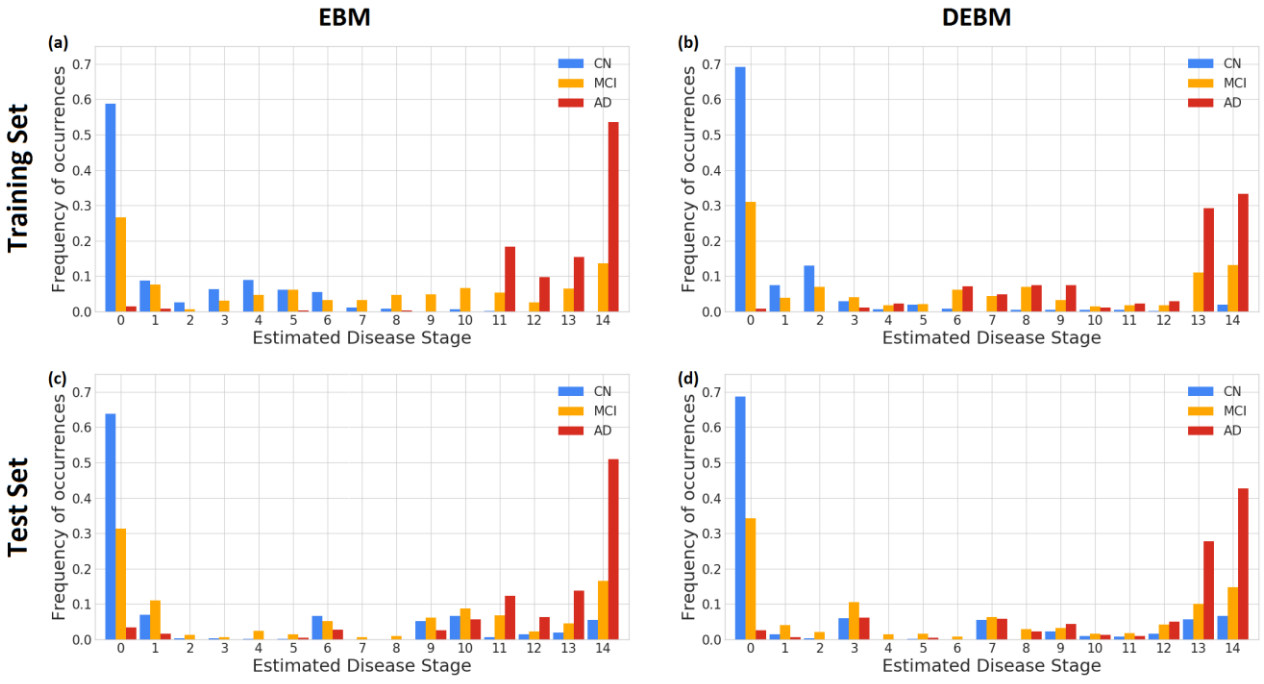
Scatter plot of DEBM stage vs. EBM stage for training (left) and test (right) subjects. Areas of annuli are proportional to the number of subjects. Linear regression resulted in slopes of 0.891 (R^2=0.802) for training subjects and 0.829 (R^2=0.680) for test subjects.

SF4. Event ordering including all biomarkers



Positional variance diagrams of Event ordering obtained with EBM and DEBM when MMSE is included in the original set of biomarkers. Both diagrams show the number of times each biomarker occurred in a certain position from a batch of 50 independent bootstrapped sequences generated form biomarkers of subjects from the training set with EBM (left) and DEBM (right) methods.

SF5. Staging of subjects including all biomarkers



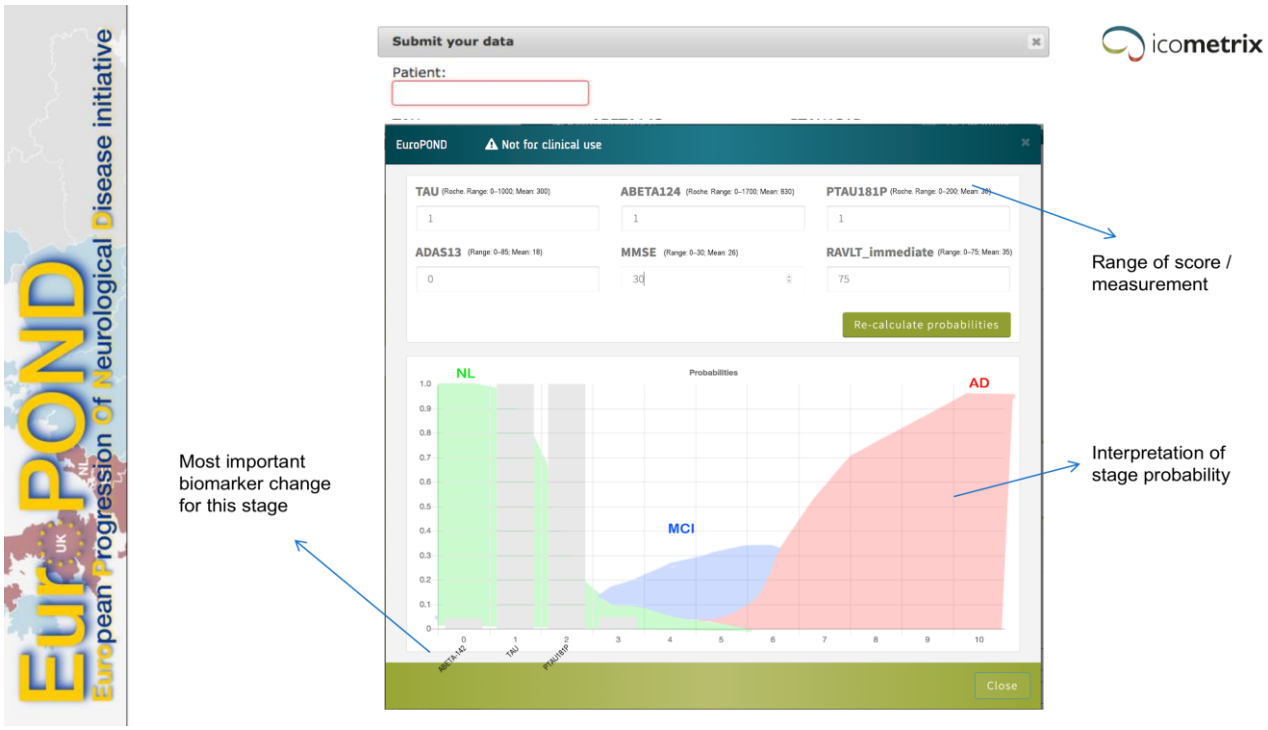
Subject staging based on the sequences obtained with EBM and DEBM methods when MMSE is included in the set of biomarkers. Staging of subjects from all diagnostic categories (Cognitively normal (CN) in blue, mild cognitive impairment (MCI) in orange, and Alzheimer's disease (AD) in red) are shown for the cases of (a) *training subjects on EBM sequence;* (b) *training subjects on DEBM sequence;* (c) *test subjects on EBM sequence and* (d) *test subjects on DEBM sequence. Histograms are normalized for each diagnostic category.*

ST2. Performance metrics of EBM and DEBM including all biomarkers

Training set	**EBM**					**DEBM**					**p-value**
	k_T	**sens**	**spec**	**BalAcc**	**AUC**	**k_T**	**sens**	**spec**	**BalAcc**	**AUC**	
AD vs CN	11	0.99	0.97	0.98	0.99*	6	0.95	0.96	0.96	0.97*	$5.66\ 10^{-2}$
AD vs MCI	11	0.72	0.97	0.84	0.85*	6	0.50	0.96	0.73	0.78	$7.64\ 10^{-2}$
MCI vs CN	7	0.97	0.48	0.73	0.75*	3	0.90	0.58	0.74	0.76*	0.671
Test set	**EBM**					**DEBM**					**p-value**
	k_T	**Sens**	**spec**	**BalAcc**	**AUC**	**k_T**	**Sens**	**spec**	**BalAcc**	**AUC**	
AD vs CN	11	0.90	0.84	0.87	0.91	3	0.70	0.97	0.84	0.89	$1.24\ 10^{-3}$
AD vs MCI	11	0.70	0.84	0.77	0.80	9	0.64	0.82	0.73	0.78	0.222
MCI vs CN	1	0.64	0.69	0.66	0.68	1	0.69	0.66	0.67	0.67	0.251

*Measurements of area under curve (AUC), sensitivity (Sens), specificity (Spec) and balanced accuracy (BalAcc) at a specific threshold (K_T) for the subject staged with EBM and DEBM methods on train and test datasets when MMSE is included in the set of biomarkers. Thresholds are chosen to maximize the balanced accuracy in each classification task. P-values of Delong test performed to compare AUCs obtained with EBM and DEBM methods are reported in the last column. AUCs of ADNI subjects denoted with * are significantly different from their analogous of test subjects (p-value ≤ 0.05)*

SF6. Single case interface of data-driven model



Single case interface (alpha release developed by Icometrix NV in the context of the EuroPOND H2020 initiative) to stage patient according to the biomarkers data availability. X-axis reports the patient's stage, Y-axis reports the probability of biomarker abnormalities.