Adaptive data-driven selection of sequences of biological and cognitive markers in clinical diagnosis of dementia

Patric Wyss[a], David Ginsbourger[b], Haochang Shou[c], Christos Davatzikos[d,e], Stefan Klöppel[a], Ahmed Abdulkadir[d,e,*], for the ISTAGING Consortium, the Alzheimer's Disease Neuroimaging Initiative[1], and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing[2]

[a] University Hospital of Old Age Psychiatry and Psychotherapy, University of Bern, Bern, Switzerland

[b] Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland

[c] Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, USA

[d] Center for Biomedical Image Computing and Analytics, University of Pennsylvania, Philadelphia, USA

[e] Department of Radiology, University of Pennsylvania, Philadelphia, USA

* Correspondence to: Ahmed Abdulkadir <ahmed.abdulkadir@pennmedicine.upenn.edu>, 3710 Hamilton Walk, 6th Floor, Center of Biomedical Image Computing and Analytics, University of Pennsylvania Philadelphia PA, 19104, https://www.med.upenn.edu/cbica/

Keywords: multi-variate mixed-effects model | neutral zone classifier | sequential classification | dementia diagnosis

[1] Part of the data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

[2] Part of data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database (www.loni.usc.edu/ADNI). The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at https://www.aibl.csiro.au.

Abstract

Combining the right—potentially invasive and expensive, markers at the appropriate time is critical to obtain reliable yet economically sustainable decisions in the preclinical diagnosis of dementia. We propose a data-driven analytical framework to individualize the selection of prognostic biomarkers that balance accuracy, costs of opportunity due to delaying the decision, and cost of acquisition depending to prescribed cost parameters. We compared sequential and non-sequential decision strategies based on a linear mixed-effects classification model that integrates irregular, multi-variate longitudinal data. The framework was applied to separate participants that progress to Alzheimer's disease from the ones that do not within a time interval of three years. As expected, the highest accuracy was obtained by combining all available data from 20.9 measurements per subject on average that were acquired over 4.8 years on average. The proposed sequential algorithm empirically outperformed alternative methods by having lowest costs for a range of tested cost parameters. With the default cost parameters, the sequential algorithm reached an accuracy of 0.84, specificity of 0.86, and sensitivity of 0.82 (0.89, 0.91, and 0.88 with all available data, respectively) while requiring only 2.9 measurements on average (86 percent less observations than all available data) and a time interval of half a year on average (89 percent shorter than all time points). Our sequential algorithms established the decision based on individualized sequences of measurements with reduced process costs compared to non-sequential classification strategies while maintaining competitive accuracy.

Clinical diagnosis and prognosis of dementia due to Alzheimer's disease (AD), a progressive, neurodegenerative disorder, is based on a panel of cognitive assessments, invasive molecular markers, and non-invasive neuro-imaging signatures. Invasive markers such as Aβ_{1-42} cerebro-spinal fluid (CSF) markers (1) impose a high burden on patients and high monetary costs. These shortcomings are balanced by high sensitivity of the diagnosis. Conversely, cognitive assessments such as the Mini-Mental-State Examination (MMSE) (2) or Rey Auditory Verbal Learning (RAVLT) test (3) have a lower economic cost and patient burden, but also a lower sensitivity in early stages of the disease. Non-invasive magnetic resonance imaging (MRI) provides machine-learning derived measures of AD-like atrophy (SPARE-AD) (4) that have intermediate cost of acquisition and intermediate sensitivity/specificity.

To date, a typical diagnostic decision of dementia is based on a panel of cross-sectional or a sequence of repeatedly measured (longitudinal) markers from multiple modalities such as MRI and cognitive testing (5–7). The temporal trajectory of the biomarkers is also an indicator of disease state and progression. There is currently, however, no consensus or systematic approach to individualize the selection of panels and sequences of markers to acquire. It is expected that the accuracy increases with more variables and/or repeated measurements over time. However, the benefit of higher accuracy that more data brings to the table is accompanied by added costs of acquisition, a delay of the definitive diagnosis, and higher patient burden. The goal of the process for clinical diagnosis of dementia is to reach a conclusion under an acceptable level of uncertainty for an individual while globally optimizing the use of resources and patient burden.

Conceptually, sequential algorithms based on neutral zone classifiers (8–10) address the use case in which a definitive decision is taken after considering a sequence of measurement by adding d measurements only if an objective criterion for a confident decision is not met. Different algorithms were presented in recent years and successfully applied to clinical data (8, 9, 11). While some only depend on fixed decision boundaries like the posterior class probabilities (11), other approaches consider expected cost reductions of additional measurements in the decisions (8). However, to our knowledge, the sequential neutral zone classifiers published so far do not allow the selection of variables or skipping time points.

We propose an algorithm that uses knowledge about an underlying distribution of markers to prospectively decide if more observations are needed for a definitive classification and which variable should be acquired next. To estimate the distributions, we embedded mixed-effects models in a classification task to allow for more flexibility and more parsimonious modelling of longitudinal data than other generative approaches such as traditional discriminant models (12–16). The proposed framework integrates irregularly sampled repeated (longitudinal) measures with varying number of observations and derives an adaptive sequential expansion of panel of markers for classification (see **Fig. 1**). The experimental evaluation of a flexible clinical decision process is reflected in multiple aspects including accuracy, resource usage, and delay of diagnosis. The overall performance is a trade-off resulting from the model design and associated cost parameters injected *a priori*. We applied the framework to predicting a conversion from sub-clinical to clinical stage of AD.

Costs of decision processes as objective evaluation criterion

A decision process d incorporates a subset of all available data for its decision according to a decision rule and—in case of sequential classifiers, also a sequence selection strategy. The cost of the decision process is an empirical metric of performance (lower is better) that is the sum of multiple costs each reflecting a different desirable quality. We implement binary decision processes that predict whether a participant $i \in \{1, .., n\}$ will progress from a sub-clinical stage of mild cognitive impairment (MCI) to the manifest stage of dementia due to Alzheimer's disease (AD) within a time interval of approximately three years. The total costs $c_{d,i}^{\mathcal{T}}$ of a process d are defined as the sum of misclassification costs $c_{d,i}^{\mathcal{C}}$ and measurement costs $c_{d,i}^{\mathcal{M}}$ consisting of the costs of delaying the decision $c_{d,i}^{\mathcal{D}}$ and costs of acquisition $c_{d,i}^{\mathcal{A}}$, such that

$$c_{d,i}^{\mathcal{T}} = c_{d,i}^{\mathcal{C}} + c_{d,i}^{\mathcal{M}} = c_{d,i}^{\mathcal{C}} + c_{d,i}^{\mathcal{D}} + c_{d,i}^{\mathcal{A}}.$$

See Supplementary Materials for details about the composition of the costs.

If the sequence $M_{p,i}$ of a of a process is fixed, it includes all pre-defined assessments independently of the evidence given by the previously assessed measurements. The adaptive strategies sequentially weighing the expected accuracy from acquired evidence against expected gain in accuracy and costs of acquiring new data and delaying the decision.

Sequence selection strategies

We considered different processes implementing either fixed or sequential decision strategies to choose the set $M_{p,i}$. The implemented fixed decision strategies were categorized as (a) univariate (measuring only one marker) or multivariate (measuring multiple markers) and (b) cross-sectional (using only one measurement per marker at baseline) or longitudinal (using all repeated measurements of markers over time). We implemented two different sequential strategies that stepwise included one additional observation until a class is chosen. The *greedy sequential strategy* always choses the earliest observation for which a cost reduction is expected, while the *exhaustive sequential strategy* selects the observation for which the expected cost reduction is the highest. Since the resulting sequences differ in terms of number, type, and time of assessment between decision strategies we compared the processes using a diverse set of performance metrics characterizing different facets.

Multi-objective evaluation of decision processes

The prognostic decision processes were compared using multiple performance metrics including the area under the receiver operator characteristics curve (AUC), specificity, sensitivity, and accuracy. Of note, from these metrics only the AUC is cost independent. The misclassification costs calibrate the severeness of false positive and false negative cases and hence influence sensitivity, specificity, and accuracy (see Eq. S5 in the Supplementary Materials for details). Since a correct diagnosis of MCI-converter can either be made the pre-clinical MCI state or after the conversion to AD already occurred, we also considered a metric to assess suitability of disease prognosis that counts prognoses after the conversion to AD as misclassification. We defined the pre-conversion sensitivity as the portion of MCI-converter that were correctly classified before the conversion occurred. Moreover, given the

arbitrariness in the assignment of cost parameters, multiple sets of cost parameters were prescribed to assess the robustness of the performance.

Performance of decision processes

The sequential algorithms were designed to select a sequence of observations with low expected total costs given a set of cost parameters. AUC and accuracy were less sensitive to the prescribed cost parameters whereas other performance measures (such as specificity and especially pre-conversion sensitivity) were highly dependent on prescribed cost parameters (see **Fig. 3** and **Tab. S1** and **Tab. S2** in the Supplementary Material). For a large range of cost parameter, both—greedy and exhaustive, sequential algorithms had lower mean total costs than all the fixed strategies (see **Tab. 1** below and **Tab. S1** and **Tab. S2** in the Supplementary Materials). Sequential strategies were competitive in terms of misclassification costs, accuracy, and sensitivity while saving measurement costs compared to longitudinal or the multivariate cross-sectional strategy (see **Tab. 1**, **Fig. 2**, **Fig. 3** and **Tab. S1** and **Tab. S2** in the Supplementary Materials). Moreover, the sequential strategies adapted to the cost parameters shown by the varying mean number of observations, mean follow-up time and accuracy. Specifications of the cost parameters especially of the costs of delaying influenced performance metrics especially specificity and pre-conversion sensitivity (see **Fig. 3** and **Tab. S1** in the Supplementary Material). If not specified otherwise, the following results are based on the following default cost parameters: *Cost of misclassification of MCI-stable=100; Cost of misclassification of MCI-converter=100; Cost for MRI acquisition=2; Cost of acquisition of* $A\beta_{1-42}$*-CSF=4; Cost of acquisition of cognitive test=1; Cost for waiting one year=2.* We chose small costs of delaying and acquisition as compared to misclassification costs to favor accuracy over measurement costs aiming to assess reductions in delaying and number of measurements of sequences that approach accuracy or AUC of classifications based on all available observations. The multi-objective evaluation revealed that both sequential strategies were competitive in terms of number of observations and the greedy strategy also for follow-up time (mean time needed for a decision of around 6 month) and resulting pre-conversion sensitivity compared to other strategies with lower accuracy and AUC (see **Fig. 2** and **Tab. 1**).

The exhaustive sequential strategy had the lowest mean total costs whereas the multivariate longitudinal strategy had the highest mean total costs driven by the highest measurement costs (see **Tab. 1**). Highest accuracy, AUC and sensitivity were achieved with the multivariate longitudinal strategy, while the longitudinal strategy using all MMSE measures was the most specific. Lowest accuracy, AUC, and sensitivity was achieved when using the first MMSE and lowest specificity when using the first $A\beta_{1-42}$-CSF measure for classification. For the multivariate cross-sectional strategy, the pre-conversion sensitivity was the highest, whereas longitudinal strategies showed the lowest pre-conversion sensitivity. Integrating the repeated measures over time led to higher specificity in case of multivariate or (univariate) cognitive markers and higher sensitivity in case of the structural marker.as compared to the corresponding cross-sectional strategy. As expected, the pre-conversion sensitivity was the same as the sensitivity for cross-sectional strategies and low for longitudinal strategies between 0.01 and 0.03 because the decision was made after conversion. The multivariate cross-sectional strategy showed improvements in accuracy, AUC and sensitivity compared to univariate cross-sectional strategies. The exhaustive sequential strategy showed smaller mean total costs than the greedy strategy driven by the higher accuracy. Also, the exhaustive strategy showed a higher AUC,

specificity and sensitivity. The pre-conversion sensitivity was higher for the greedy than for the exhaustive strategy. For the greedy strategy 12 percent of MCI-converters where correctly classified after conversion while this percentage was 43 for the exhaustive strategy. Both sequential strategies outperformed all univariate cross-sectional strategies in terms of accuracy, AUC, sensitivity, and specificity, while all these measures were lower than the one achieved by some longitudinal strategies. The multivariate cross-sectional strategy was outperformed by both sequential strategies in terms of accuracy and specificity while sensitivity was higher (AUC in between: lower than for the exhaustive and higher than for the greedy). Both sequential strategies--and the greedy strategy in particular, led to more frequent pre-conversion diagnosis of AD than longitudinal strategies, but the level of the pre-conversion sensitivity of most cross-sectional strategies could not be retained when using the default cost parameters. In summary, there were strategies that outperformed others in multiple criteria (e.g., total costs as well as accuracy measures) but no strategy dominated all others in all considered criteria. Most fixed strategies were either specific (longitudinal) or prognostic (univariate cross-sectional), while the sequential strategies made a trade-off in between.

In **Fig. 2(a)** decision processes were evaluated relative to the performance of using all available information. While the univariate cross-sectional strategy showed reductions for all classification-based performances compared to the multivariate longitudinal strategy, the multivariate cross-sectional approached the level of sensitivity and sequential strategies approached accuracy and specificity. Longitudinal MMSE measures had similar AUC and accuracy with higher specificity but lower sensitivity. All cross-sectional and sequential strategies had a lower number of observations and lower follow-up time. Over the whole sample, 70 percent of subjects had at least two measurements, while the invasive $A\beta_{1-42}$-CSF was included in 30percent of cases (**Tab. 2**). Besides the MMSE (first measurement after around 9 months and last around1.5 years in average), all other markers were assessed early (last observation before 4 months in average). Descriptive measures of other sequential strategies (with exhaustive selection criteria or other cost parameters) can be found in the Supplementary Material (see **Tab. S3 - Tab. S27**)

An evaluation of decision strategies in terms of total, misclassification (which is equal to the percentage of false classifications for the considered cost parameters) and measurement costs is presented in **Fig. 2(b).** Univariate cross-sectional strategies had higher mean misclassification costs than other strategies. All univariate longitudinal strategies--except the one including $A\beta_{1-42}$-CSF measures, were characterized by lower mean misclassification costs and higher mean measurement costs than all other non-sequential strategies. Sequential strategies had the lowest mean total costs on average and had lower misclassification costs than all cross-sectional and some longitudinal strategies. There was no strategy that consistently dominated all other strategies in all objectives. The greedy sequential strategy dominated the multivariate cross-sectional and univariate longitudinal strategies involving either all MRI, $A\beta_{1-42}$-CSF or RAVLT measures. The exhaustive strategy additionally showed similar misclassification costs as the longitudinal strategy using all MMSE measures in all three objectives and was dominated in terms of misclassification costs only by the multivariate longitudinal strategy. Both sequential strategies had higher mean measurement costs than the univariate cross-sectional strategies. Moreover, the sequential strategies were outperformed by some longitudinal strategies in terms of the cost independent AUC, univariate cross-sectional strategies in terms of both

lower costs of acquisition and delaying. and the multivariate cross-sectional in terms of costs of delaying only (but lower costs of acquisition) (**Fig. 2 (c)-(e)**).

Survival curves estimating the portion of subject that did not convert as a function of the conversion time are shown in **Fig. 2 (f)**. Observations of participants labelled as MCI-stable were considered as right-censored and the time until the last observation was used for the analysis. The multivariate cross-sectional strategy showed more distinct survival curves of positive and negative labelled cases than the one using only the first MRI measure. Adding all repeated measures of the markers (multivariate longitudinal strategy) further increased the difference between curves. Sequential strategies based on the default cost structure scored in between the two fixed multivariate strategies. When using all MMSE measures (most specific strategy as shown in **Tab. 1**) for classification the survival curves of positively labelled cases are the steepest of all considered decision strategies, but the curves of cases with predicted negative labels were steeper than for multivariate cross-sectional and longitudinal and sequential strategies as well.

All considered cost parameters presented here assumed the same misclassification costs such that accuracy, specificity, and sensitivity can be compared (see **Tab. S1** and **Tab. S2** in the Supplementary Material for results with other misclassification costs and additional variations of costs of aquistion and time). Costs either differed in terms of costs of waiting, the cost of assessments or both from the default cost structure. In **Fig. 3 (a)-(b)** AUCs of strategies are contrasted to the resources of time or number of considered measurements showing gains of performance when taking up more resources. All sequential strategies had lower AUCs as the strategy using all MMSE measures or the multivariate longitudinal and higher AUCs as univariate cross-sectional strategies. Depending on the specified cost parameters sequential strategies showed higher or lower AUCs than the multivariate cross-sectional strategy while always using lower number of observations but longer follow-up times in average. As shown in **Fig. 3 (c)** both specificity and sensitivity of sequential strategies were positioned between the fixed strategies by being always less or equal sensitive than the multivariate cross-sectional and always less sensitive than the multivariate longitudinal or the longitudinal strategy involving the RAVLT measures only as well as less specific than the multivariate longitudinal and the longitudinal strategy involving MMSE measures only. Both sequential strategies showed higher sensitivities and higher or equal specificities than all univariate cross-sectional and some univariate longitudinal strategies. The trade-off between specificity and pre-conversion sensitivity of different sequential strategies is illustrated in **Fig. 3 (d)-(f)**. Sequential strategies were bounded by the specificity of the multivariate longitudinal and the pre-conversion sensitivity of the multivariate cross-sectional strategy (**Fig. 3 (d)**). Greedy strategies were less specific but more prognostic than exhaustive strategies, while increasing the costs of waiting led to lower follow-up times and consequently higher pre-conversion sensitivities for the prize of lower specificities (**Fig. 3 (c)-(d)**).

Discussion

Personalized medical care tailors the diagnosis and treatment to individual patients based on circumstantial evidence and options. Motivated by a use case of pre-clinical diagnosis of AD—the most frequent cause of severe cognitive decline at old age, we developed a general statistical framework to objectify the selection of sequence and time of diagnostic markers to include in a decision process. As

previous studies indicated the benefit of combining multiple, diverse markers (17), cognitive, brain anatomical, and molecular markers were included here. The performance of the non-sequential strategies consistently improved when multiple cross-sectional and multiple longitudinal measurements were combined. This confirmed that the implemented longitudinal discriminant model was able to integrate multi-variate and repeated measurements. This was further demonstrated in the time-to-event analyses in which the models with the most data consistently had highest odds ratio between the classes. The observed complementary information of different diagnostic markers has been shown and explored in the literature previously (17) and the performance achieved with all available data marks the maximum accuracy that can be expected from sequential algorithm with fewer variables or time points. However, under certain conditions, a single or very few markers may be sufficient for a definitive diagnosis and it's the objective of sequential neutral zone classifiers to balance misclassification and acquisition costs. The results of the sequential classification strategies showed the expected benefits of reaching similar accuracy (2 to 9 percent reduction) with fewer acquisitions and shorter follow-up interval. With the default cost parameters, the waiting time until a decision was around half a year in average, whereas 30.1 percent of the participants were definitively classified based on a single measurement and in about 69.7 percent of the cases, no Aβ_{1-42}-CSF measure was included in the panel of biomarkers, considerably reducing the administration of invasive procedures.

The cost parameters influenced the performance, the cost independent measure AUC was between 0.85 and 0.92 and the accuracy between 0.80 and 0.87. The influence of the prescribed parameter for the costs of time on AUC and accuracy was stronger than the one of the acquisition costs because delayed decisions (when specifying low costs of time) had higher specificity. When increasing costs of time, more converters were definitively diagnosed before the actual conversion occurred, making the setting better suited for prognosis. Interestingly, higher costs of acquisition did not result in drop in AUC or accuracy. The implemented greedy sequential strategies based on high acquisition costs considerably reduced number of observations especially for Aβ_{1-42}-CSF (between 0 and 40.4 percent of subjects with Aβ_{1-42}-CSF assessment for varying cost of acquisition) and instead chose to assess the cognitive MMSE scale after a longer follow-up time when gains in accuracy pay out against the high acquisition costs. All these considerations are relevant when aiming to prescribe cost parameters in clinical diagnosis.

Academic research of biomarkers is often focussed on methods to build good predictive model with a single or a few general-purpose modalities (18). While such an approach is adequate to maximize population-level classification accuracy, it is not useful to design a decision process. Acquisition of a complete set of markers for all participants with high accuracy on population-level might still be insufficient in clinical diagnosis since it does not consider individual uncertainty of prediction, nor does it optimize the use of resources. An individualized sequence of markers and their measurement time points that is chosen sequentially follows the precision diagnostic principle of "use it only when and in whom it is most needed" while optimizing resources and impose patient burden only when that is countered with a prognostic value.

There is a rich literature in medical statistics covering the modelling of irregular univariate or multivariate repeated (longitudinal) data with mixed-effects models (19–30) as well as their

application to derive more flexible approaches for discriminant analyses for clinical data (12–16, 31–35). Also, in recent years there have been studies that applied mixed-effects models to infer underlying structures of multivariate change or to derive flexible or even dynamic classification approaches (11, 16, 22, 24, 27) also in diagnosis of neurodegeneration (22, 32, 35). The approaches for sequential classification with mixed-effects models in resent studies (11, 16) relied purely on acquired evidence without considering potential benefits of measurements. In the field of neutral zone prediction multiple approaches were presented in the last years (8–10, 36). Application based on such neutral zone classifiers that enable sequential classification were derived and applied to clinical data. Existing sequential neutral zone classifiers (8, 9, 11) are designed for multi-stage classification Our algorithm can skip observations and can choose which biomarker to select next. Due to the flexibility of mixed-effects models to adapt model structures to the data, task, and sample size, our framework of non-sequential and sequential classification based on mixed-effects models may be of interest in different applications and research fields. The specific implementation here was limited to linear model and fixed prevalence, but the concept is applicable to non-linear mixed models and conditional prevalence.

The classification task was defined by a clinically motivated, yet arbitrary, thresholds of follow-up and conversion time. Moreover, in the two studies, patients with significant neurological disorders and most psychiatric disorders were excluded, limiting the validity to a prospective clinical population as shown previously in applications of machine-learning methods to data from clinical routine (37, 38). Here we evaluated fixed cost parameters. In a clinical application, the costs could depend on the visit or on other factors. For instance, as already implemented in many clinical workups, initial suspicion of dementia due to AD requires a confirmatory MRI to exclude other neurological disorders. In our framework, this would lead to a cost penalty of zero in case of a suspected case of AD or worded differently: "no definitive diagnosis of AD without structural MRI". The data included in this study was inadequate to evaluate more complex cost structures that better cater to needs in the clinical routine. Furthermore, the simplified modelling of the distribution underlying the predictive model may limit the performance of the model and its application to clinical populations. While effective and computationally light, we implemented a greedy approach selecting a single measurement in a sequence which does not guarantee to find the globally optimal next step. A globally optimal algorithm would need to consider all possible future sequences which is computationally intensive and intractable for sufficiently high number of markers. A better approach might be derived by adapting the multi-stage approach presented in (8) in a way that allows to choose which observation should be taken next.

While the proposed framework does not alleviate the necessity of choosing cost parameters, it demonstrated the ability to adapt in out-of-sample classification to consistently lower costs compared to cross-sectional predictions or the ones using all available information of one or multiple markers we have for a subject. Sequential strategies that were designed to lower overall decision costs also showed to be competitive in individual performance measures. The proposed statistical framework constitutes a promising element for precision diagnostic that makes the panel of diagnostic tools conditional on past and potential future evidence, thus specifically objectifying the acquisition of the panel of markers after each visit.

Materials and Methods

Decision rules for classification

The decision rule governs the mapping from a test statics derived from observations to a binary decision. Suppose we have a class label $z \in \{1; 2\}$ a random binary variable with Bernoulli distribution with unconditional success probability π_0 (base rate) and $\boldsymbol{y} \in \mathbb{R}^m$, is a random vector taking continuous values, $\boldsymbol{y}$ has density $\phi^{(1)}$ for the population conditioning on class $z = 1$, and density $\phi^{(2)}$ for the population when class $z = 2$. Furthermore, we assume that π_0, $\phi^{(1)}$ and $\phi^{(2)}$ are known and the densities of multivariate normal distributions with a common covariance matrix. Thus,

$$\boldsymbol{y}|z \sim N_m(\boldsymbol{\mu}^{(z)}, \boldsymbol{\Sigma}) \tag{1}$$

with mean vectors $\boldsymbol{\mu}^{(z)} = E(\boldsymbol{y}|z)$ $(z \in \{1; 2\}, \boldsymbol{\mu}^{(1)} \neq \boldsymbol{\mu}^{(2)})$ and the common covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{(1)} = \boldsymbol{\Sigma}^{(2)} = \mathrm{Var}(\boldsymbol{y}|z)$. The posterior probability of $z = 2$ conditioning on $\boldsymbol{y}$ is given by:

$$\pi = P(z = 2|\boldsymbol{y}) = \frac{\pi_0 \phi^{(2)}(\boldsymbol{y})}{\pi_0 \phi^{(2)}(\boldsymbol{y}) + (1 - \pi_0)\phi^{(1)}(\boldsymbol{y})} \tag{2}$$

Suppose $\boldsymbol{y}_k \in \mathbb{R}^{m_k}$ are past observations, and $\boldsymbol{y}_l \in \mathbb{R}^{m_l}$ are optional future observations and without loss of generality, let $\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_k \\ \boldsymbol{y}_l \end{pmatrix}$. The posterior probability π_k is given by the observed values $\boldsymbol{y}_k$ and its assumed density functions for both population $\phi_k^{(1)}$ and $\phi_k^{(2)}$ (plugged in in Eq. 2) and is called the current evidence. Based on the current evidence, a forced choice classifier $\delta_{FC,k}$ definitively assigns one of the classes in a binary classification. The decision rule that minimizes misclassification costs of the forced choice classifier $\delta_{\mathrm{FC},k}$ is

$$\delta_{\mathrm{FC},k} = \begin{cases} 1 & \text{if } \pi_k < b \\ 2 & \text{if } \pi_k \geq b \end{cases}, \qquad b = \frac{c_2^{(1)}}{c_2^{(1)} + c_1^{(2)}} \tag{3}$$

where $c_2^{(1)}$ are the costs of false positive and $c_1^{(2)}$ the costs of false negative classifications. The expected misclassification costs EC_k for a forced-choice classification after y_k was assessed can be computed with the current evidence π_k as:

$$EC_k = \begin{cases} \pi_k c_1^{(2)}, & \delta_{FC,k} = 1 \\ (1 - \pi_k) c_2^{(1)}, & \delta_{FC,k} = 2 \end{cases}$$

Neutral zone classifiers add a no decision class NZ to the set of possible predicted outcomes and associated costs c_{NZ}. The no decision class is picked when the test statistic is within a neutral zone. For the descriptive neutral zone classifier denoted by $\delta_{DNZ,k}$ based on the assessed $\boldsymbol{y}_k$, the classification boundaries for the current evidence π_k depend only on the pre-specified costs parameters. The classifier $\delta_{DNZ,k}$ is given by

$$\delta_{DNZ,k} = \begin{cases} 1, & \pi_k \leq b_1 \\ NZ, & b_1 < \pi_k < b_2, \\ 2, & \pi_k \geq b_2 \end{cases} \qquad b_1 = \frac{c_{NZ}}{c_1^{(2)}}, \quad b_2 = \frac{c_2^{(1)} - c_{NZ}}{c_2^{(1)}} \tag{4}$$

In case $b_1 < b_2$, the classifier in Eq. 4 minimizes the costs, if not, we end up with the forced-choice classifier in Eq. 3 as minimum cost classifier (adapted by (11)). Since $c_{NZ} = b_1 c_1^{(2)}$ and $c_{NZ} =$

$(1-b_2)c_2^{(1)}$, the descriptive neutral zone classifiers predict the label NZ in case the expected misclassification costs for both classes are higher than the neutral costs c_{NZ}.

The descriptive neutral zone classifier does not consider potential value of the future observations, that is, how much future observations may contribute in reducing the process costs. The decision process of the descriptive neutral zone classifier may be improved by postponing the decision only if the expected future costs given by $\boldsymbol{y}_l$ conditional on the value of the already assessed $\boldsymbol{y}_k$ denoted by $EC_{l.k}$ are lower than the ones expected by a forced-choice decision based on $\boldsymbol{y}_k$ denoted by EC_k. To this end, the measurement costs $c^{\mathcal{M}}$ associated by postponing the decision and assessing the remaining markers $\boldsymbol{y}_l$ are considered. In such a case, the neutral costs are $c_{NZ} = EC_{l.k}$ given by the costs measurement costs and the expected future misclassification cost (for the descriptive neutral zone classifier we can think of neutral costs that are given by the measurement costs, i.e. $c_{NZ} = c^{\mathcal{M}}$) Consequently, for $EC_{l.k}$ the false positive and false negative rate of a forced choice classification based on the joint distribution of $\boldsymbol{y}_l$ and z conditional on $\boldsymbol{y}_k$ is needed. In linear discriminant models, the misclassification rates depend solely on the misclassification costs $c_2^{(1)}$ and $c_1^{(2)}$, current evidence π_k and the prospective discriminability denoted by $\Delta_{l.k}$ (see **Fig. 4** for an illustration and the Supplementary Material for the derivation of Eq. S10). The prospective discriminability is defined as $\Delta_{l.k} = \left\|\boldsymbol{\mu}_{l.k}{}^{(1)} - \boldsymbol{\mu}_{l.k}{}^{(2)}\right\|_{\Sigma_{l.k}} = \sqrt{(\boldsymbol{\mu}_{l.k}{}^{(1)} - \boldsymbol{\mu}_{l.k}{}^{(2)})^T \boldsymbol{\Sigma}_{l.k}{}^{-1}(\boldsymbol{\mu}_{l.k}{}^{(1)} - \boldsymbol{\mu}_{l.k}{}^{(2)})}$ where $\boldsymbol{\mu}_{l.k}{}^{(z)}$ are the expectations and $\boldsymbol{\Sigma}_{l.k} = \boldsymbol{\Sigma}_{l.k}^{(1)} = \boldsymbol{\Sigma}_{l.k}^{(2)}$ the covariance matrix of the distributions $\boldsymbol{y}_l|\boldsymbol{y}_k, z \sim N_{m_l}(\boldsymbol{\mu}_{l.k}{}^{(z)}, \boldsymbol{\Sigma}_{l.k})$ (see Eq. S6 in the supplement for the equations of mean vectors and covariance matrix) and it quantifies the expected added value of including $\boldsymbol{y}_l$. We denote the false positive rate by $FP(\pi_k, \Delta_{l.k})$, the false negative rate by $FN(\pi_k, \Delta_{l.k})$, the specificity by $SE(\pi_k, \Delta_{l.k})$ and the sensitivity by $SE(\pi_k, \Delta_{l.k})$ for a classification based on the joint distribution $\boldsymbol{y}_l, z|\boldsymbol{y}_k$. The expected costs $EC_{l.k}$ can be written as:

$$EC_{l.k} = c^{\mathcal{M}} + FP(\pi_k, \Delta_{l.k})(1-\pi_k)c_2^{(1)} + FN(\pi_k, \Delta_{l.k})\pi_k c_1^{(2)} \tag{5}$$

The exact formula (closed form solution) for the expected misclassification rates $FP(\pi_k, \Delta_{l.k})$ and $FN(\pi_k, \Delta_{l.k})$ can be found in the Supplementary Material (see Eq. S10). Given the expected costs we can derive a modified neutral zone classifier that assigns the label NZ only in case $EC_k > EC_{l.k}$, i.e., when a forced-choice classification including the future measurements $\boldsymbol{y}_l$ given the already assessed $\boldsymbol{y}_k$ is smaller than the expected costs for a forced-choice classification given solely by the already assessed $\boldsymbol{y}_k$. The prospective neutral zone classifier $\delta_{PNZ,k}$ is given by:

$$\delta_{PNZ,k} = \begin{cases} 1, & \pi_k c_1^{(2)} < \min(EC_{l.k}, (1-\pi_k)c_2^{(1)}) \\ NZ, & EC_{l.k} < \min(\pi_k c_1^{(2)}, (1-\pi_k)c_2^{(1)}) \\ 2, & (1-\pi_k)c_2^{(1)} < \min(EC_{l.k}, \pi_k c_1^{(2)}) \end{cases} = \begin{cases} 1, & \pi_k \leq b_1(\Delta_{l.k}) \\ NZ, & b_1(\Delta_{l.k}) < \pi_k < b_2(\Delta_{l.k}) \\ 2, & \pi_k \geq b_2(\Delta_{l.k}) \end{cases} \tag{6}$$

Where the second part of Eq. 6 only holds in case $b_1(\Delta_{l.k}) < b_2(\Delta_{l.k})$ and the boundaries $b_1(\Delta_{l.k})$ and $b_2(\Delta_{l.k})$ cannot be computed in a closed form (see Eq. S11 and Eq. S12 and their description in the Supplementary Material for more information). Of note, we did not compute the boundaries and used the first part of Eq. 6 for classification.

Fig. 4 illustrates the forced-choice, descriptive neutral zone, and prospective neutral zone classifiers.

A multi-variate longitudinal discriminant model for sequential classification

In this study we implemented classification framework that can deal with irregularly acquired repeated and multi-variate measurements. Mixed effects model-based estimation was embedded into linear (or quadratic, see Supplementary Material) discriminant models to account for inter-subject differences (12–15, 34, 35, 39). The vector $\boldsymbol{y}_i$ from participant $i \in \{1; 2; \dots; n\}$ consists of longitudinal measurements from multiple time points $y_{i,j}$ acquiring one of four response variables. We assumed for a subject i with measurements $\boldsymbol{y}_i$and unknown label z_i that

$$\begin{aligned} z_i &\sim Bernoulli(\hat{\pi}_0) \\ \boldsymbol{y}_i | z &\sim N_{m_i}(\hat{\boldsymbol{\mu}}_i^{(z)}, \hat{\boldsymbol{\Sigma}}_{\mathrm{i}}) \end{aligned} \tag{7}$$

The prevalence (unconditional probability) $\pi_0 = p(z = 2)$ was estimated by the relative frequency $\hat{\pi}_0$ of subjects with diagnosis 2 in the training set. The population level estimators of the mean vector $\hat{\boldsymbol{\mu}}_i^{(z)}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_i$ were derived using linear mixed-effects models trained on labelled measurements denoted by $y_{i,j}^{(z_i)}$ respectively $\boldsymbol{y}_i^{(z_i)}$. The LMM included age at baseline a_i, the time $t_{i,j}$ since baseline of observation *j* and four dummy variables for identifying the response variable denoted by $v_{h,i,j}$ ($h \in H$). We used a model adapted from (19, 29, 30) where

$$y_{i,j}^{(z_i)} = \sum_{h \in H} v_{h,i,j} \left(\beta_{h,1}^{(z_i)} + \beta_{h,2}^{(z_i)} a_i + \beta_{h,3}^{(z_i)} t_{i,j} + \zeta_{h,i,1} + \zeta_{h,i,2} t_{i,j} + \rho_h \epsilon_{i,j} \right), \tag{8}$$

whereas $\beta_{h,1}^{(z)}$, $\beta_{h,2}^{(z)}$ and $\beta_{h,3}^{(z)}$ ($z \in \{1; 2\}$) are the diagnosis and response variable specific fixed effects and $\boldsymbol{\beta}^{(z)}$ the vector containing all diagnosis specific fixed effects (population-level), $\zeta_{h,i,1}$ and $\zeta_{h,i,2}$ the response specific random effects (subject-level, same for both labels) and$\epsilon_{i,j}$ the (scaled) residuals which are multiplied with response specific intra-subject variance components ρ_h (same for both labels). The distribution of the vector ζ_i containing all random effects (for the intercept and time for all variables) is given by $\boldsymbol{\zeta}_i \sim N_{2r}(0, \boldsymbol{\Psi})$. The scaled residuals were assumed to be independent from each other and the random intercept and slopes and standard norm distributed i.e., $\epsilon_{i,j} \sim N(0,1)$. The distribution of the unscaled residuals $\varepsilon_{i,j}$ varies between response variables and is given as. $\varepsilon_{i,j} \sim N(0, \sum_{h \in H} v_{h,i,j} \rho_h)$. We denote with $\boldsymbol{\rho} = (\rho_h)_{h \in \mathrm{H}}$ the vector containing all response-type specific intra-subject variances. All model parameter $\boldsymbol{\theta} = \left[\pi_0; \boldsymbol{\beta}^{(1)}; \boldsymbol{\beta}^{(1)}; \boldsymbol{\Psi}; \boldsymbol{\rho}\right]$ (prevalence for the diagnose 2 and model parameters of the LMM) for the discriminant models in Eq. 7 were estimated on training data with labelled observations using a 20-fold cross validation framework. The estimation of the parameters of the LMMs were obtained with the restricted maximum likelihood approach using the library *nlme* (see (40) for a detailed description) from the statistical programming language R 4.1.1 (41).

Given the 20 trained models cross-validated out-of-sample predictions of $\hat{\boldsymbol{\mu}}_i^{(z)}$ and $\hat{\boldsymbol{\Sigma}}_i$ from all m_i observations of subject were conducted. With the estimated prevalence, mean vectors and covariance matrix we computed all previously derived quantities (posterior probabilities, expected miss-classification rates, expected costs and classifiers) by plugging in the estimates $\hat{\pi}_0, \hat{\boldsymbol{\mu}}_i^{(z)}$ and $\hat{\boldsymbol{\Sigma}}_i$ of

subject i for the true population values in the equations. Moreover, we derived a sequential algorithm that stepwise adds new observations based on decision and selection rules. The sequential classifier predicts for a subject i at the step k $(1 \leq k \leq m_i)$ one of the possible diagnoses or stay in the neutral zone (decision rule) and selects which future measurement should be included next (selection rules). The sequential classifier (Eq. S17) consisted of the application of the prospective neutral zone classifier as given by Eq. 6 to individual future observations. We can think of applying the prospective neutral zone classifier for every left-over observation separately and assigning the label NZ if at least for one observation the prospective neutral zone classifier reveals the label NZ as outcome (see the Supplementary material for more information). In case the label NZ is chosen a selection rule is applied to choose which (single) observation is included next for the prediction. We used two different selection rules i.e., the greedy rule where the earliest observation with expected cost reduction or the exhaustive rule where the observations with highest expected cost reduction is chosen as the next observation. In every step all quantities needed to compute the decision as well as selection rules are given by the estimated means and covariances contained in $\hat{\boldsymbol{\mu}}_i^{(1)}$ respectively $\hat{\boldsymbol{\mu}}_i^{(2)}$ and $\hat{\boldsymbol{\Sigma}}_i$. The code for **P**r**O**spective **SE**quent**I**al **D**iagn**O**sis with **N**eutral zones (POSEIDON) is available at https://git.upd.unibe.ch/openscience/POSEIDON.

Study data and empirical evaluation

Longitudinal data from individuals from Alzheimer's Disease Neuroimaging Initiative (ADNI) (42) and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) (43) were included. AIBL study methodology has been reported previously (43). The ADNI was launched in 2003 with the primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information and data access see https://www.adni-info.org and https://adni.loni.usc.edu.

We included biological as well as cognitive markers to separate patients with MCI that do not convert to AD over a follow-up time of at least 2.75 years (MCI-stable, label $z = 2$) or convert to manifest AD within 3.25 years since study entry (MCI-converter, $z = 1$). As structural biomarker we used the SPARE-AD score (4) computed from regional brain volumes obtained from standard structural MRI with a multi-atlas segmentation algorithm (44) that captures how "AD-like" the structure of the brain of a subject is. The SPARE-AD score is the average decision value of an ensemble of linear support vector machine classifiers trained to discriminate cognitively healthy subjects from patient suffering from AD based on regional anatomical brain volumes. The ensemble was trained on multi-centric data from the iSTAGING Consortium. We include $A\beta_{1-42}$ levels in the CSF (1) as invasive, AD-specific marker. Cognitive markers were scores given by either the MMSE or RAVLT. All four markers were irregularly measured with differing time points and number of observations between subjects. We implemented the described linear mixed-effects model based classification approach to perform non-sequential and sequential classification. We fitted linear mixed-effects models considering the longitudinal measurements of all variables and included only subjects that match the definition either of MCI-stable or MCI-converter and had at least 8 observations independently of the variable. Eventually, a sample with 336 MCI-stables and 269 MCI-converters were used to fit the 20 models for

cross-validated predictions. For the comparison of the decision strategies, we only used the data of subjects that had at least one observation of each variable within 3 months after the first observation, leading to a sample of 234 MCI-stables and MCI-converters 162 (see **Fig. S1** in the Supplementary Material for more information about the selection of the data set).

Acknowledgments

The iSTAGING consortium is a multi-institutional effort funded by NIA through RF1 742 AG054409. ADNI (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012) are funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

1. Shaw LM *et al.* (2009) Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. *Annals of Neurology* 65:403–413.
2. Khanfer R *et al.* (2013) in *Encyclopedia of Behavioral Medicine*, eds Gellman MD, Turner JR. (Springer New York, New York, NY), pp 1248–1249.
3. Bean J (2011) in *Encyclopedia of Clinical Neuropsychology*, eds Kreutzer JS, DeLuca J, Caplan B. (Springer New York, New York, NY), pp 2174–2175.
4. Davatzikos C, Xu F, An Y, Fan Y, Resnick SM (2009) Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index. *Brain : a journal of neurology* 132:2026–2035.
5. McKhann GM *et al.* (2011) The diagnosis of dementia due to Alzheimer's disease. Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia : the journal of the Alzheimer's Association* 7:263–269.
6. Albert MS *et al.* (2011) The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia* 7:270–279.

7. Sperling RA *et al.* (2011) Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia* 7:280–292.

8. Kim H, Jeske DR (2017) Truncated SPRTs with application to multivariate normal data. *Sequential Analysis* 36:251–277.

9. Jeske DR, Zhang Z, Smith S (2020) Construction, visualization and application of neutral zone classifiers. *Statistical methods in medical research* 29:1420–1433.

10. Jeske DR, Smith S (2018) Maximizing the usefulness of statistical classifiers for two populations with illustrative applications. *Statistical methods in medical research* 27:2344–2358.

11. Zhang X, Jeske DR, Li J, Wong V (2016) A sequential logistic regression classifier based on mixed effects with applications to longitudinal data. *Computational Statistics & Data Analysis* 94:238–249.

12. Lix LM, Sajobi TT (2010) Discriminant analysis for repeated measures data: a review. *Frontiers in psychology* 1:146.

13. Marshall G, Barn AE (2000) Linear discriminant models for unbalanced longitudinal data. *Statist. Med.* 19:1969–1981.

14. Marshall G, La Cruz-Mesía R de, Quintana FA, Barón AE (2009) Discriminant analysis for longitudinal data with multiple continuous responses and possibly missing data. *Biometrics* 65:69–80.

15. Lisa Tomasko, Ronald W. Helms, Steven M. Snapinn (1999) A discriminant analysis extension to mixed models. *Statistics in medicine* 18:1249–1260.

16. Hughes DM, Komárek A, Czanner G, Garcia-Fiñana M (2018) Dynamic longitudinal discriminant analysis using multiple longitudinal markers of different types. *Statistical methods in medical research* 27:2060–2080.

17. Palmqvist S *et al.* (2021) Prediction of future Alzheimer's disease dementia using plasma phospho-tau combined with other accessible measures. *Nat Med* 27:1034–1042.

18. Rathore S, Habes M, Iftikhar MA, Shacklett A, Davatzikos C (2017) A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage* 155:530–548.

19. Doran HC, Lockwood JR (2006) Fitting Value-Added Models in R. *Journal of Educational and Behavioral Statistics* 31:205–230.

20. Thum YM (1997) Hierarchical Linear Models for Multivariate Outcomes. *Journal of Educational and Behavioral Statistics* 22:77–108.

21. Fieuws S, Verbeke G (2006) Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics* 62:424–431.

22. Gordon BA *et al.* (2018) Spatial patterns of neuroimaging biomarker change in individuals from families with autosomal dominant Alzheimer's disease: a longitudinal study. *The Lancet Neurology* 17:241–250.

23. Beckett LA, Tancredi DJ, Wilson RS (2004) Multivariate longitudinal models for complex change processes. *Statistics in medicine* 23:231–239.

24. Gao F *et al.* (2017) Estimating correlation between multivariate longitudinal data in the presence of heterogeneity. *BMC medical research methodology* 17:124.

25. Verbeke G, Fieuws S, Molenberghs G, Davidian M (2014) The analysis of multivariate longitudinal data: a review. *Statistical methods in medical research* 23:42–59.

26. Shah A, Laird N, Schoenfeld D (1997) A Random-Effects Model for Multiple Characteristics With Possibly Missing Data. *Journal of the American Statistical Association* 92:775.

27. Adjakossa EH, Sadissou I, Hounkonnou MN, Nuel G (2016) Multivariate Longitudinal Analysis with Bivariate Correlation Test. *PloS one* 11:e0159649.

28. Fieuws S, Verbeke G (2004) Joint modelling of multivariate longitudinal profiles: pitfalls of the random-effects approach. *Statistics in medicine* 23:3093–3104.

29. Goldstein H (2011) *Multilevel statistical models* (Wiley, Hoboken (N.J.)).

30. MacCallum RC, Kim C, Malarkey WB, Kiecolt-Glaser JK (1997) Studying Multivariate Change Using Multilevel Models and Latent Curve Models. *Multivariate behavioral research* 32:215–253.

31. La Cruz-Mesía R de, Quintana FA (2007) A model-based approach to Bayesian classification with applications to predicting pregnancy outcomes from longitudinal beta-hCG profiles. *Biostatistics (Oxford, England)* 8:228–238.

32. Brant LJ, Sheng SL, Morrell CH, Zonderman AB (2005) Data from a longitudinal study provided measurements of cognition to screen for Alzheimer's disease. *Journal of clinical epidemiology* 58:701–707.

33. Fieuws S, Verbeke G, Maes B, Vanrenterghem Y (2008) Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics (Oxford, England)* 9:419–431.

34. Liu D, Albert PS (2014) Combination of longitudinal biomarkers in predicting binary events. *Biostatistics (Oxford, England)* 15:706–718.

35. Sheng SL, Brant LJ (2004) in *Encyclopedia of Statistical Sciences*, eds Kotz S, Read CB, Balakrishnan N, Vidakovic B, Johnson NL. (John Wiley & Sons, Inc, Hoboken, NJ, USA).

36. Benecke S, Jeske DR, Reugger P, Borneman J (2013) Bayes Neutral Zone Classifiers With Applications to Nonparametric Unsupervised Settings. *JABES* 18:39–52.

37. Klöppel S *et al. (*2015) Applying automated MR-based diagnostic methods to the memory clinic. A prospective study. *Journal of Alzheimer's disease : JAD* 47:939–954.

38. Stephan KE *et al. (*2017) Computational neuroimaging strategies for single patient predictions. *NeuroImage* 145:180–199.

39. Lu Z, Leen TK, Kaye J (2011) Kernels for longitudinal data with variable sequence length and sampling intervals. *Neural computation* 23:2390–2420.

40. Pinheiro JC, Bates DM (20) *Mixed-Effects Models in S and S-PLUS* (Springer International Publishing, Cham).

41. R Core Team (2017) *R: A Language and Environment for Statistical Computing* (Vienna, Austria). https://www.R-project.org/.

42. Mueller SG *et al. (*2005) The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America* 15:869-77, xi-xii.

43. Ellis KA *et al. (*2009) The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *International Psychogeriatrics* 21:672–687.

44. Doshi J *et al.* (2016) MUSE: MUlti-atlas region Segmentation utilizing Ensembles of registration algorithms and parameters, and locally optimal atlas selection. *NeuroImage* 127:186–195.

Figures and Tables

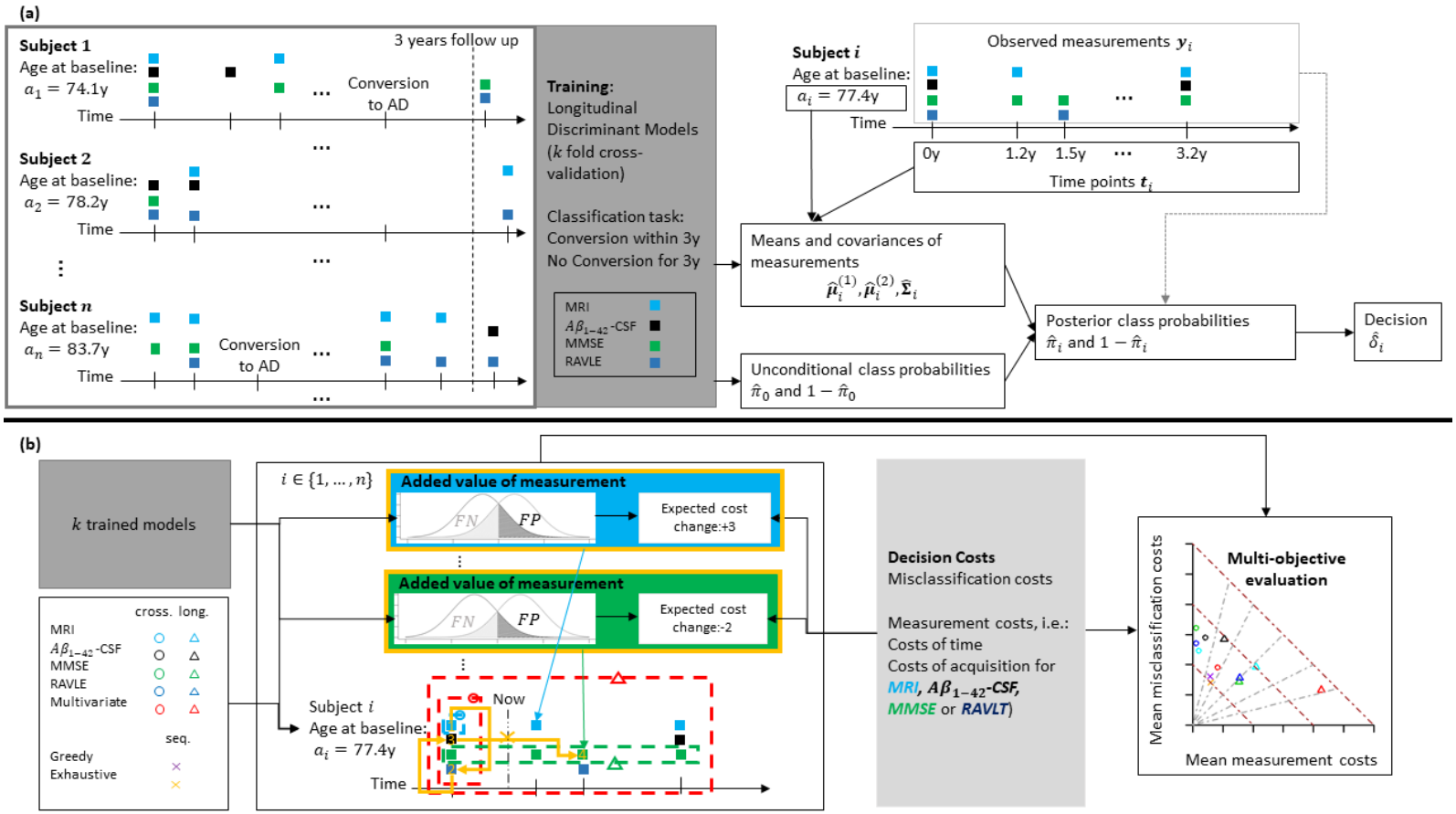

Fig. 1 Visualisation of the implemented classification framework. **(a)** Linear mixed-effects models where trained (using $k = 20$ fold cross-validation) on irregular, multi-variate longitudinal data to derive classifiers separating patients with mild cognitive impairment that either converted to AD within approximately three years or stayed stable for approximately three years. Distributions (means and covariances) of the markers (MRI, $A\beta_{1-42}$-CSF, MMSE or RAVLT) were estimated using the age at baseline and the time points at which the observations occurred (time since baseline) as predictors. **(b)** Fixed cross-sectional (using only the measurement(s) at baseline) or longitudinal (using measurements over time) and sequential decision strategies were applied to perform the classification task. The strategies were evaluated from a multi-objective perspective using a set of performance metrics such as e.g. the mean misclassification or measurement costs.

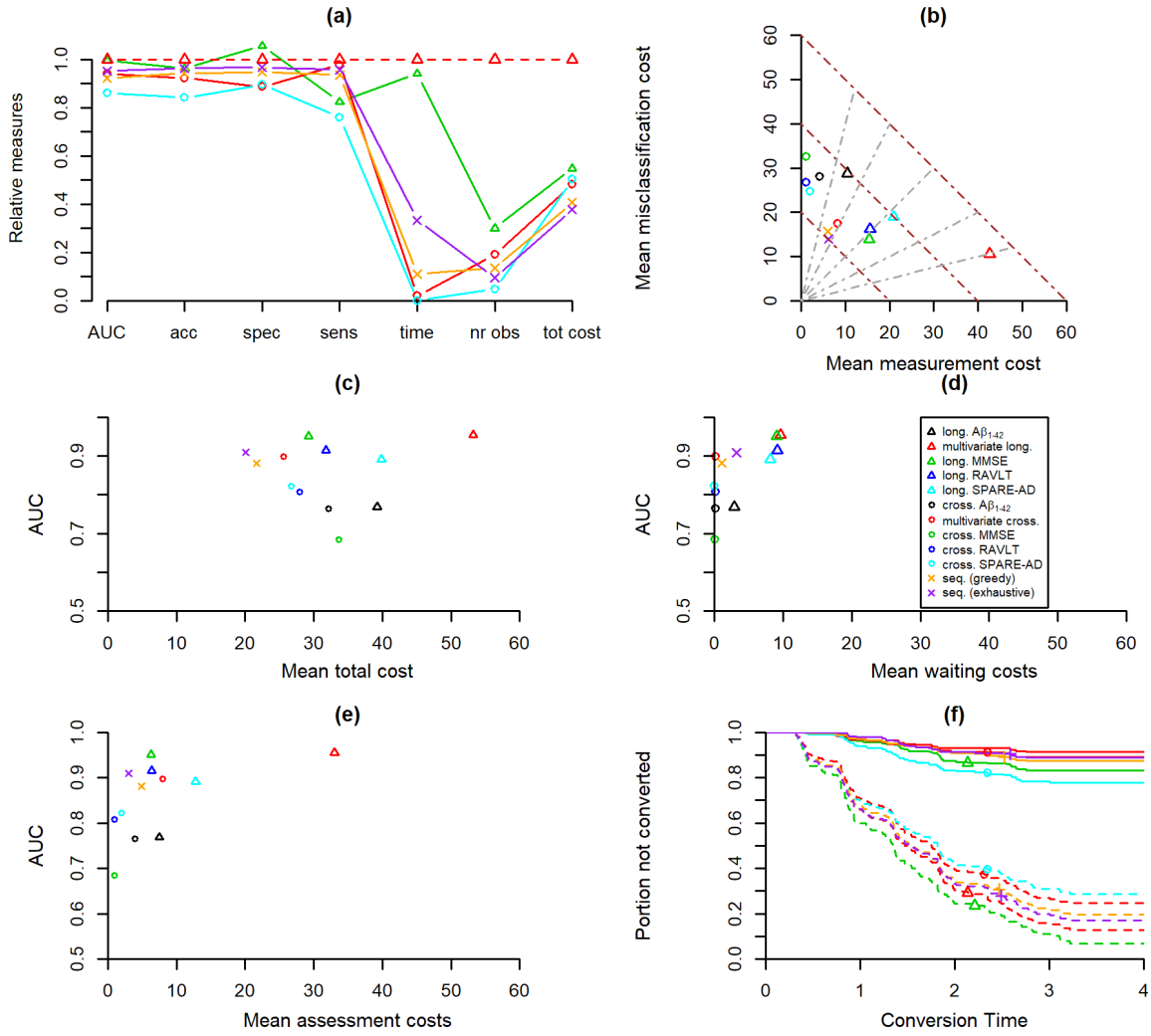

Fig. 2: Comparison of decision strategies using the default cost parameters. **(a)** Strategies in relation to the one using all available information (for some selected strategies). On one hand, accuracy measures (AUC, accuracy, specificity, and sensitivity) of different strategies divided with the one from the multivariate longitudinal strategy are displayed (represent portion of retained performance). On the other hand, the ratios of mean follow-up time or number of observations to the one of the multivariate longitudinal strategy were computed (represent portion of utilized resources). The ratio of the total cost of strategies with the one using all information is also displayed in the right end of the figure. **(b)** Mean misclassification costs vs. mean measurement costs: Brown dotted lines represent points with same mean total and grey dotted lines represent the shares of misclassification and measurement costs from the total costs (middle line: same misclassification and measurement cost, rotated left: higher share of misclassification costs, rotated right: higher share of measurement costs). **(c)-(e)**: AUC vs. specific type of decision costs either the mean total **(c)**, assessment **(d)** or time **(e)** costs. **(f)**: Kaplan-Meier estimates of survival curves i.e., the portion of not converted subjects (where MCI stables are considered as right- censored) either fitted on the cases predicted as MCI-stable (straight lines) or MCI-converter (dashed lines) by the decision strategies.

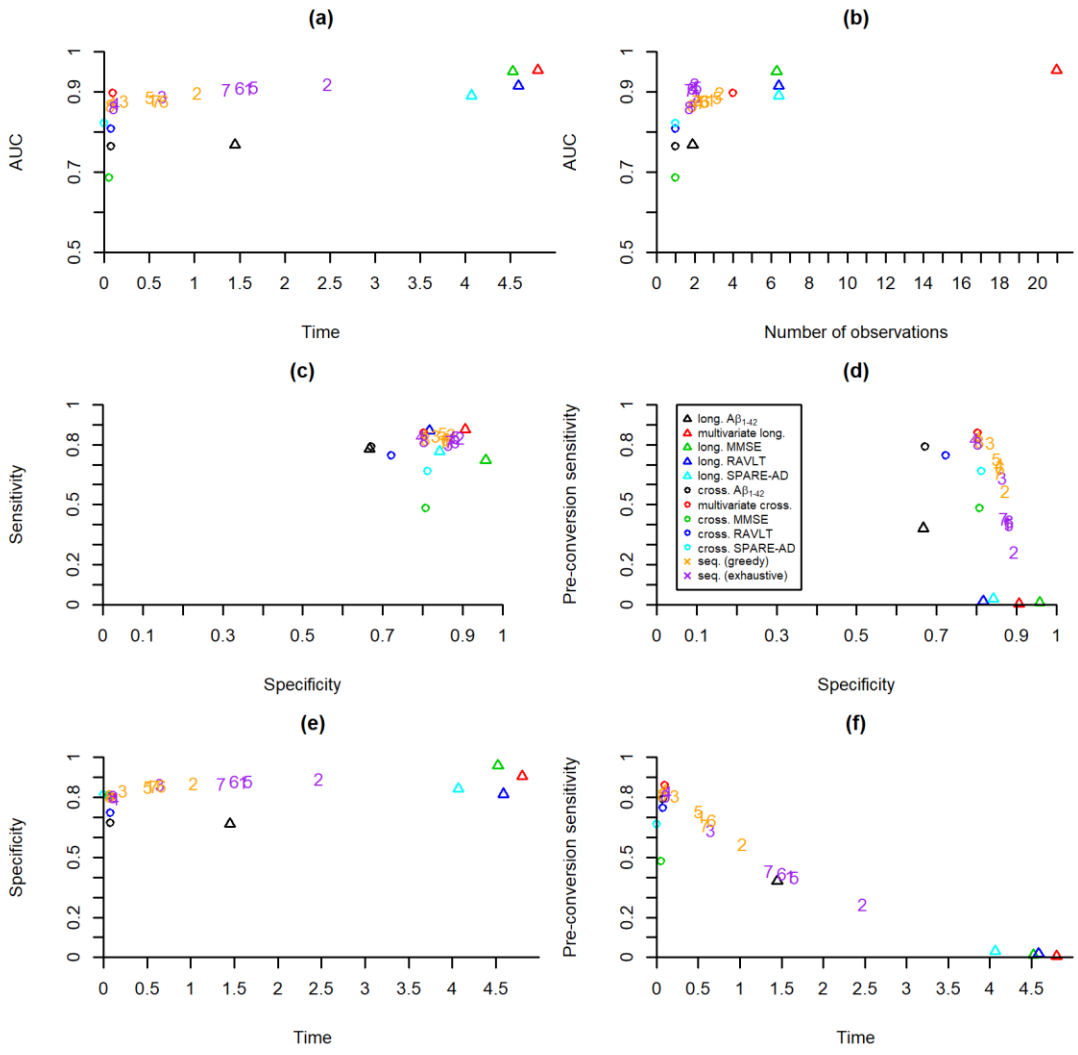

Fig. 3: Multi- objective quantitative evaluation of decision strategies comparing fixed and sequential strategies based on different cost parameters (see below for the specifications). **(a)** and **(b)**: AUC as a function of the follow up time **(a)** or number of observations **(b)** used for a decision in average. **(c)** Comparison of decision strategies in terms of specificity and sensitivity, **(d)** Comparison of decision strategies in terms of specificity and pre-conversion sensitivity (portion of correctly classified MCI-converters before conversion to AD), **(e)** and **(f)**: Specificity **(d)** or pre-conversion sensitivity **(f)** as a function of follow-up time used for a decision in average.

Costs:	**Time**	**MRI**	**$A\beta_{1-42}$-CSF**	**cognitive test**
Prescription 1	2	2	4	1
Prescription 2	1	2	4	1
Prescription 3	4	2	4	1
Prescription 4	8	2	4	1
Prescription 5	2	1	2	0.5
Prescription 6	2	4	8	2
Prescription 7	2	8	16	4
Prescription 8	8	4	8	2

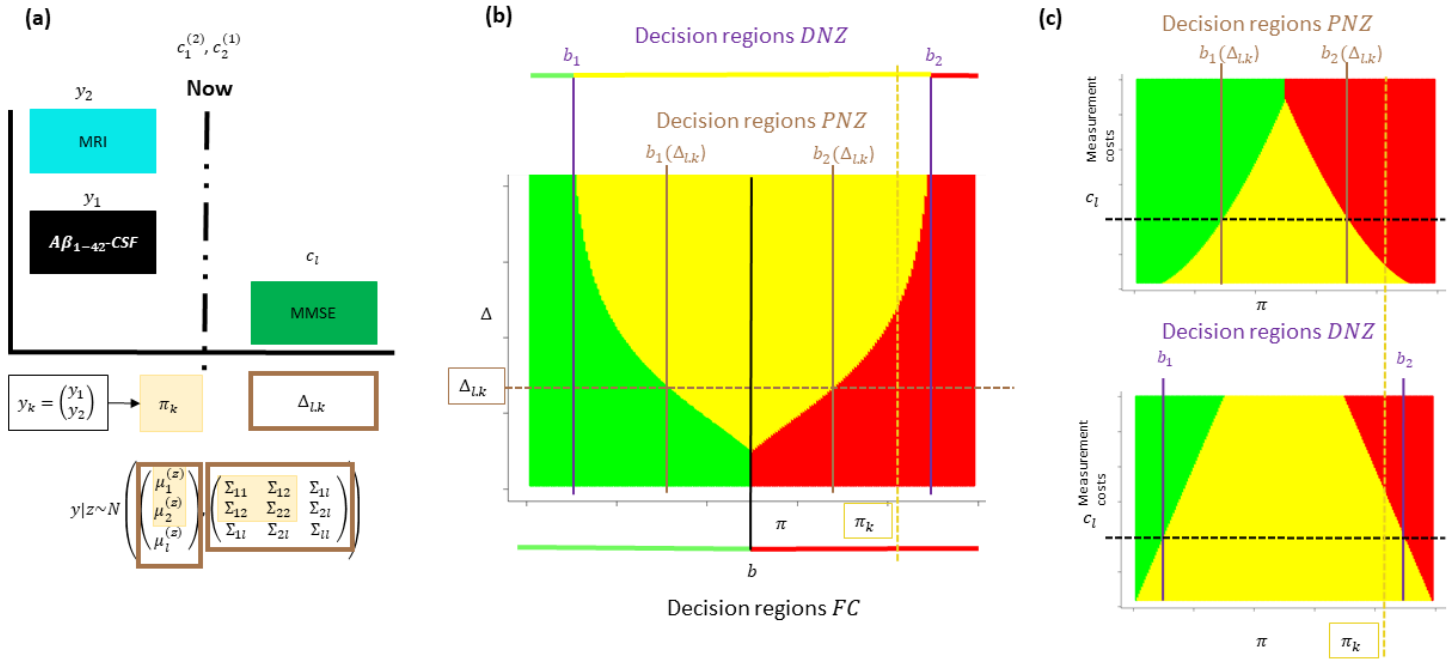

Fig. 4: Visualization of the prospective neutral zone classifier **(a)** Situation in which already a MRI (y_1) and $A\beta_{1-42}$-CSF (y_2) measure were assessed ($\boldsymbol{y}_k$ vector containing both measures) and the classification task consist of either choosing one of the possible classes (MCI-stable or MCI converter) in case there is enough evidence or postpone the decision and stay neutral (NZ) to additionally assess the MMSE (y_l) and then decide for one of the classes. The classification outcome depends on the misclassification costs $c_1^{(2)}$, $c_2^{(1)}$ and the measurement cost c_l for y_l, the current evidence π_k and the prospective discriminability $\Delta_{l.k}$ (given by the conditional distributions $y_l|z, \boldsymbol{y}_k$). **(b)** Visualizations of decision regions for constant decision costs. In contrast to the forced-choice (FC, black line) and the descriptive neutral zone classifier (DNZ, purple lines) for the prospective neutral zone classifier (PNZ) the amount of leftover diagnosis-relevant information in y_l (quantified by $\Delta_{l.k}$) determines the amount of evidence given by the assessment $\boldsymbol{y}_k$ (quantified by π_k) that is needed to make a finite decision without assessing y_l ($\pi_k < b_1(\Delta_{l.k})$ or $\pi_k > b_2(\Delta_{l.k})$, Eq. S11 and Eq. S12 in the Supplementary Material for more infornation). For the prospective neutral zone classifier, the prediction outcome for a given value of $\Delta_{l.k}$ is displayed by the coloured areas, the boundaries between the green and yellow area is $b_1(\Delta_{l.k})$ and the one between the yellow and red area is $b_2(\Delta_{l.k})$ **(c)** For fixed misclassification costs $c_1^{(2)}$, $c_2^{(1)}$ and prospective discriminability $\Delta_{l.k}$ the decision boundaries for the PNZ- and DNZ classifiers are displayed as a function of the measurement costs.

Tab. 1: Performances of non-sequential and sequential decision strategies based on the default cost structure.

Strategy	ACC[1]	SPEC[2]	SEN[3](PRE[4])	AUC	total costs[5]	measurement costs[6]	No measurement	Time
long. $A\beta_{1-42}$ CSF	0.71	0.67	0.78	0.77	39.23	10.44	1.89	1.45
long. markers	**0.89**	0.91	**0.88**	**0.96**	**53.25**	**42.64**	20.98	4.80
long. MMSE	0.86	**0.96**	0.72	0.95	29.23	15.34	6.29	4.53
long. RAVLT	0.84	0.82	0.87	0.92	31.74	15.58	6.40	4.59
long. MRI	0.81	0.84	0.77	0.89	39.87	20.93	6.40	4.07
cross. $A\beta_{1-42}$ CSF	0.72	**0.67**	0.79	0.77	32.19	4.16	1.00	0.08
cross. Biomarkers	0.83	0.80	0.86	0.90	25.61	8.19	4.00	0.09
cross MMSE	**0.67**	0.81	**0.48**	**0.69**	33.68	**1.11**	1.00	0.05
cross. RAVLT	0.73	0.72	0.75	0.81	27.92	1.15	1.00	0.08
cross. MRI	0.75	0.81	0.67	0.82	26.75	2.00	1.00	0.00
greedy seq. markers[7]	0.84	0.86	0.82	0.88	21.67	6.01	2.86	0.53
exhaustive seq. markers[7]	0.86	0.88	0.84	0.91	**20.15**	6.26	2.03	1.60

[1-4]: Accuracy, specificity, sensitivity and pre-conversion sensitivity with decision boundaries depending on the estimated unconditional class probabilities and specified misclassification costs

[5]: Depend on misclassification costs, marker-specific costs of acquisition and costs of waiting

[6]: Depend on marker-specific costs of acquisition and costs of waiting

[7]: Chosen sequence of measurements depend on the specified misclassification costs, marker-specific costs of acquisition and costs of waiting

Tab. 2: Descriptive measures of the greedy sequential strategy for a specified cost structure.

No. measurements	**All** (mean/prop > 1)[1]	**$A\beta_{1-42}$-CSF** (mean/prop >0)[2]	**MMSE** (mean/prop >0)[2]	**RAVLT** (mean/prop >0)[2]	**SPARE_AD** (mean/prop >0)[2]
All subjects	2.86/0.70	0.30/0.30	0.58/0.34	0.8/0.65	1.18/0.99
MCI-stable	2.75/0.68	0.25/0.25	0.53/0.30	0.79/0.65	1.18/1.00
MCI-converter	3.02/0.72	0.38/0.38	0.65/0.38	0.82/0.67	1.17/0.99
Time	(last)[3]	(first /last)[4]	(first/ last)[4]	(first/ last)[4]	(first/ last)[4]
All subjects	0.53	0.11/0.11	0.71/1.45	0.09/0.27	0.00/0.17
MCI-stable	0.53	0.13/0.13	0.79/1.57	0.10/0.27	0.00/0.17
MCI-converter	0.54	0.09/0.09	0.61/1.31	0.09/0.28	0.00/0.18

[1]: Mean number of measurements (from all markers) and relative frequency of cases where more than one observation was chosen

[2]: Mean number of measurements for the specified marker and relative frequency of cases where at least one measurement of the specified marker was chosen

[3]: Mean follow-up time (time of last chosen measurement)

[4]: Mean average time at which the first respectively last measurement of the specified marker was chosen