PERSONALIZED PROGNOSIS & TREATMENT USING AN OPTIMAL PREDICTOR MACHINE: AN EXAMPLE STUDY ON CONVERSION FROM MILD COGNITIVE IMPAIRMENT TO ALZHEIMER'S DISEASE

P.G.L. Porta Mana

western norway university of applied sciences, norway

I. Rye

university of oslo, norway

A. Vik

haukeland university hospital, norway

M. Kociński, A. Lundervold, A.J. Lundervold

university of bergen, norway

A.S. Lundervold

western norway university of applied sciences, norway

4th November 2023

Correspondence should be sent to

E-Mail: pgl@portamana.org

PERSONALIZED PROGNOSIS & TREATMENT USING AN OPTIMAL PREDICTOR MACHINE: AN EXAMPLE STUDY ON CONVERSION FROM MILD COGNITIVE IMPAIRMENT TO ALZHEIMER'S DISEASE

Abstract

The present work presents a statistically sound, rigorous, and model-free algorithm – called "optimal predictor machine" in homage to these two pioneers – for use in personalized medicine. The algorithm is designed first to learn from a set of clinical data with relevant predictors and predictands, and then to assist a clinician in the assessment of prognosis & treatment for new patients. It allows the clinician to input, for each new patient, additional patient-dependent clinical information, as well as patient-dependent information about benefits and drawbacks of available treatments. We apply the algorithm in a realistic setting for clinical decision-making, incorporating clinical, environmental, imaging, and genetic data, using a data set of subjects suffering from mild cognitive impairment and Alzheimer's Disease. We show how the algorithm is theoretically optimal, and discuss some of its major advantages for decision-making under risk, resource planning, imputation of missing values, assessing the prognostic importance of predictors, and further uses.

Key words: Clinical decision making, Utility theory, Probability theory, Artificial Intelligence, Machine Learning, Base-rate fallacy

1. Introduction: Personalized prognosis, treatment, and computer algorithms

1.0. Prologue: Four unique patients

Meet Olivia, Ariel, Bianca, Curtis.[1] These four persons don't know each other, but they have something in common: they all suffer from a mild form of cognitive impairment, and are afraid that their impairment will turn into Alzheimer's Disease within a couple of years. This is why each of them recently underwent a wide range of clinical examinations and tests, including brain imaging. Today they are receiving the results. Based on their individual results, on available clinical statistical data, and on other relevant information, their clinician will assess their risk of developing Alzheimer's Disease. Then, together with the patients and their relatives, the clinician will make a decision among four distinct preventive-treatment options, available to each patient.[2] In these tasks, the clinician will be helped by a computer algorithm.

Besides a shared diagnosis of Mild Cognitive Impairment and associated worries, these patients have other things in common – but also some differences. Let's take Olivia as reference, and list the similarities and differences between her and the other three patients:

- Olivia and Ariel have identical results on the clinical and laboratory measures and age. They would also incur similar benefits and losses from the four available treatment options. Ariel, however, comes from a different geographical region, which presents a higher rate of conversion from Mild Cognitive Impairment to Alzheimer's Disease. And unlike Olivia, Ariel comes from a family with a heavy history of Alzheimer's Disease. Because of this geographical and family background and some relevant statistics found in some publications, the clinician judges, before seeing the clinical data, that there's a 65% probability that Ariel's cognitive impairment will convert to Alzheimer's Disease.
- Olivia and Bianca have identical clinical results and age; they also come from the same geographical region and have very similar family histories. In fact, we shall see that they have the same probability of developing Alzheimer's Disease. Bianca, however, suffers from several allergies and additional clinical conditions that render some of the treatment options slightly riskier for her.
- Olivia and Curtis have different results on all measures included in the clinical and laboratory examinations; Olivia is also more than 10 years older than Curtis. They otherwise come from the same geographical region, have very similar family histories, and would incur similar benefits or

[1] These are purely fictive characters but with clinically realistic conditions; any reference to real persons is purely coincidental.

[2] In the present paper, we use "prognosis" in a general sense to include also "diagnosis", and "treatment" quite loosely to mean any course of action a clinician might take, including preventive treatment or even "additional tests".

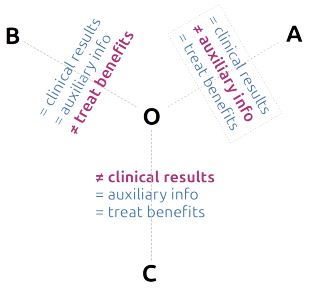
losses from the treatment options. Note that the imaging result for Curtis (hippocampal volume) is missing.

Considering the similarities and differences among these patients, which of the four available treatments will be optimal for each of them? The clinician will find that, despite the many factors in common among our four patients – even despite Olivia's, Ariel's, and Bianca's identical clinical results, and Olivia's and Bianca's identical probability of conversion to Alzheimer's Disease – *the optimal treatment for each patient is different from those for the other three* – how come?

1.1. Assistive computer algorithms: personalized input and output

In the example above, we said "in these tasks, the clinician will be helped by a computer algorithm". The need for such computational help is clear from the vast amount of clinical statistical data and the large number of clinical predictors today available to clinicians. But how should such an assistive computer algorithm be designed in order to take fully into account patient differences?

Although the example above concerns specifically Alzheimer's Disease, the differences among patients described there apply more generally to most, if not all, clinical problems of prognosis and treatment. These differences can be broadly categorized as "difference in auxiliary or supplementary tests and background information" (Olivia vs Ariel), "difference in benefit and availability of treatments" (Olivia vs Bianca), "difference in clinical predictors" (Olivia vs Curtis), as schematized in the side figure.



Each of these difference categories can affect the clinician's final choice of optimal treatment. An assistive algorithm should therefore reflect these differences in its input, its output, or both:

- In principle, there could be three kinds of input "slots", where the clinician can input the current patient's specific values as regards clinical predictors, auxiliary information, and treatment options & benefits.
- If input slots are only available for one or two of the categories above, the output should at least be of such a kind as to allow the clinician to integrate the current patient's specific values of the missing input categories.

To appreciate these requirements, one should contrast the input and output of many kinds of machine-learning classification algorithms. These typically only allow the input of a patient's clinical predictors, with no space for patient-specific auxiliary information or for adjustments of differences in background statistics (think of Olivia vs Ariel). And they typically output only a discrete prognostic label (say, "stable Mild Cognitive Impairment" vs "conversion to Alzheimer's

Disease"), but no measure of the uncertainty about that label. Unfortunately, such output does not allow the clinician to assess treatment benefits and losses for the current patient, for this assessment depends not on the presence (present or future) of a disease, but on the *risk* of its presence. We shall discuss these points at length in §§ 3.2 and 3.3.

The purpose of the present work is to present an assistive algorithm that meets the requirements above. This algorithm is designed to first learn from a dataset of clinical data with relevant predictors and predictand[3], and then assist a clinician in the assessment of prognosis & treatment for new patients. It offers these ten features:

1. It can work with clinical predictors comprising any combination of categorical and one-dimensional (continuous, discrete ordinal, unbounded or bounded, uncensored or censored) variates. The predictand can also be any combination of categorical and one-dimensional variates.
2. It treats predictor and predictand variates on equal footing, in the sense that the clinician can at any moment decide to infer some other variate given the rest.
3. It does not require that the current patient be considered in all respects as a member of the population underlying the learning dataset. The patient can be considered a member only conditionally on particular variate values.
4. It accepts three inputs:
   (a) the clinical-predictor values for the current patient;
   (b) information about which predictand-predictor relationships learned from the dataset can be generalized to the current patient, and a prior prognostic probability representing auxiliary information;
   (c) a set of treatment options and their benefits and losses for the current patient.
5. It yields three basic outputs:
   (a) any prognostic probabilities or likelihoods about predictors and predictand desired by the clinician, given input 4a;
   (b) final prognostic probabilities, given inputs 4a–4b;
   (c) optimal treatment, given inputs 4a–4c;
6. Its input and outputs are modular, in the sense that the clinician can, for instance, give inputs 4a–4b only, get a prognostic probability 5b as output, and then proceed to treatment assessment by other means or algorithms.
7. It works even if predictor data are missing, both in the learning dataset and for the current

[3] literally "quantity to be predicted" or, more generally, inferred (cf. *measurand* in metrology, JCGM 2012, 2.3). We find this term, used in meteorology and climate science, more precise and less obscure or misleading than "dependent variate", "response variate", "outcome variable", or similar.

patient.

8. It can quantify the uncertainty of its own outputs, allowing for sensitivity analyses. For example, it can tell how much a prognostic probability could have been different if the learning dataset had been larger, or whether the optimal treatment could be different if a particular missing predictor for the current patient were available.
9. It can make various kinds of long-term forecasts, such as frequency of prognoses with given probabilities, frequency of prescribed treatments, and similar – provided that the dataset used for its learning can be considered representative of the full population.
10. It is model-free and extracts the maximal amount of information theoretically contained in the learning dataset, and therefore achieves the maximal prognostic power that the predictors can yield. In other words, it is unbeatable.

Let us comment on some of these features. We believe that the capability of working with complex predictands, feature 1, is important for a more realistic and nuanced approach to prognosis. In the case of Alzheimer's Disease, for instance, a simple dichotomy "has disease" vs. "doesn't have disease" is possibly an oversimplification[4]. Without feature 3, the capability of auxiliary contextual information, the algorithm would be of no use in the often occurring case of patients having peculiar clinical contexts. The capability of dealing with missing data, feature 7, is important for a concrete implementation in a clinical setting, typically afflicted by imputation problems. Feature 8 is extremely important for a clinician to assess the reliability of final decisions and honestly inform the patient of the possibility of unwanted outcomes. Finally, features 2 and 10, the fact that this algorithm yields the maximal amount of information jointly contained in all variates, makes it valuable in general clinical research. The algorithm can, for example, forecast the maximal accuracy obtainable by *any* inference algorithm based on the same predictors or a subset of those predictors; and it attains, by construction, that maximal accuracy. Further features of interest in Machine Learning are discussed in the next section.

We call this algorithm a *optimal predictor machine*, for reasons explained in the next section. It is at the moment available as a collection of scripts[5] in the R programming language (R Core Team, 2023), which we plan to assemble into a clinician-friendly R package soon.

The methodology underlying this algorithm has been successfully demonstrated for Alzheimer's Disease with a smaller number of predictors (Antoniano-Villalobos et al., 2014), is used in many applications in astrophysics (Event Horizon Telescope Collaboration, 2019, 2022; Del Pozzo et al., 2018), and its advantages in neurocritical care have recently been emphasized (Jawa and Maslove,

[4] see e.g. Edmonds et al. (2015, 2020), whose methods we find, however, inconclusive.

[5] DOI:10.17605/osf.io/zb26t, https://github.com/pglpm/bayes_nonparametric_inference.

2023).

The next section 2 gives an intuitive understanding of the optimal predictor machine's underlying principles and workings. The machine's concrete application is shown in § 3, using the four-patient fictitious scenario of § 1.0 as a concrete example, and subsection 3.4 discusses further applications to general medical research. A summary and discussion is given in § 4. Mathematical details and proofs on which the present work is grounded are given in a companion technical note[6], which also explains how to use the R scripts.

We apologize to readers who may find some discussions or explanations too obvious, or some mathematical details too scarce. We wanted the present work to be accessible to a wide audience, from clinicians and students of medicine to researchers in machine learning and probability theory.

2. The optimal predictor machine

This section can be especially of interest to readers from Machine Learning and Artificial Intelligence. It is largely independent of the next one, which describes the machine's application. It can be read after § 3 by readers who would like to see the machine in action first.

2.1. Underlying theory and characteristics

The method to solve clinical decision-making problems such as the one of § 1 is none other than Decision Theory: the combination of probability theory and utility theory. It integrates available clinical statistical data with each patient's unique combination of clinical results, auxiliary information, and treatment benefits, in a mathematical framework, completely determined by basic self-consistency requirements.[7]

Medicine has the distinction of having been one of the first fields to adopt Decision Theory, with the pioneering work by Ledley – who, incidentally, died of Alzheimer's Disease (Shah et al., 2013) – and Lusted (Ledley and Lusted, 1959a,b, 1960; Lusted and Ledley, 1960; Lusted, 1967), who also promoted its algorithmic implementation (Lusted, 1968; Ledley, 1959, 1960, § 1-5 p. 21). Clinical decision-making is today explained and exemplified in brilliant textbooks for medical students and clinicians (Weinstein and Fineberg, 1980; Sox et al., 2013; Hunink et al., 2014). An outline is given in § 3.3.

[6] https://github.com/pglpm/bayes_nonparametric_inference/raw/main/omni-predictor_machine.pdf

[7] Jaynes (2003, chs 13–14); von Neumann and Morgenstern (1955); Cox (1946); Savage (1972); Luce and Raiffa (1957); Raiffa and Schlaifer (2000); Raiffa (1970); Lindley (1988); Kreps (1988).

The "optimal predictor machine" is an algorithmic implementation, as dreamed by Lusted and Ledley[8], of the main calculations underlying the clinical decision-making process: from the comparison of a patient's specific predictors with the statistics offered by a clinical database, to the choice of optimal treatment.[9]

Decision theory is also the normative foundation for the construction of an Artificial Intelligence agent capable of rational inference and decision making (Russell and Norvig 2022, part IV; Jaynes 2003, chs 1–2, 13–14). The optimal predictor machine can therefore be seen as an *ideal* machine-learning algorithm. It is "ideal" in the sense of being free from special modelling assumptions (this is why we do not call it a "model") and from limitations of informational output which affect most common machine-learning algorithms; not "ideal" in the sense of being impracticable. Quite the opposite, the present work shows that this ideal machine-learning algorithm can today be used in a wide range of inference problems at insubstantial computational cost.

More concretely, the optimal predictor machine is ideal because *it computes the probability distribution over all possible long-run frequency distributions from which the learning dataset can originate*, these frequency distributions being joint ones for all predictor and predictand variates.[10] This is the maximum possible amount of information that can be extracted from the learning dataset, in a strict information-theoretic sense. From this probability distribution, the optimal predictor machine can indeed calculate any quantity outputted by other machine-learning algorithms. For example (for terminology see e.g. Murphy, 2012, § 8.6):

- *"Discriminative" algorithms:* the probability $\mathrm{p}(Y \mid X)$ of any set of predictands Y given any set of input predictors X.
- *"Generative" algorithms:* the probability $\mathrm{p}(X \mid Y)$ of any set of input predictors X given any set of predictand values Y.

  More generally, the machine can compute any joint, marginal, or conditional probabilities $p(Z', Z'')$, $p(Z')$, $p(Z' \mid Z'')$ for any desired subsets of variates Z', Z''.
- *Regression or classification:* the expected value $\mathrm{E}(Y \mid X)$ of any set of variates Y, given any other set of variates X, including the particular case of Y predictand, and X predictors. The uncertainty or variability around such an average is also automatically computed.

[8] cf. the Appendices in Lusted 1968.

[9] In previous drafts we called this kind of machine "Lusted-Jaynes machine" as a homage to Lusted and to Jaynes (2003), who brilliantly explained the inductive logic underlying such a "robot".

[10] This goes by the Sibylline technical name of "Bayesian nonparametric density regression"; see e.g. Rodríguez et al. (2009); Bhattacharya and Dunson (2010); and Walker's (2010) witty overview.

- *Functional regression:* if the predictand Y or any other variate of interest turns out to be a function f of variates X, then their conditional probability will be a delta distribution: $\mathrm{p}(Y \mid X) = \delta[Y - f(X)]$. Thus the optimal predictor machine always recovers a functional relationship if there is one, as well as its noise distribution.

Furthermore, the machine also quantifies the uncertainty of all outputs above. More precisely, it takes into account how the statistical properties of the learning dataset could be different from those of its original population, owing to sampling fluctuations; and it can compute how much any of the outputs above would probably change if more learning data were collected.

In the next section we explain intuitively how the optimal predictor machine computes the general probability distribution over long-run frequencies. A couple of special characteristics brought about by such computation can already be summarized here. First, in contrast to machine-learning algorithms such as neural networks, random forests, Gaussian processes, support-vector machines, or generalized linear models, the optimal predictor machine does not assume the existence of a function (possibly contaminated by a little noise) from predictors to predictands. This is a very strong assumption, justifiable in the presence of informationally very rich predictors such as images, but otherwise quite unrealistic for many kinds of predictors considered in medicine, especially those that are more readily available and less invasive and, therefore, more desirable. Second, in contrast to algorithms such as neural networks, random forests, support-vector machines, logistic regression, or generalized linear models, the optimal predictor machine does not do an optimization during the learning phase, searching for the minimum of some objective function. It does a full *hypothesis-space survey*. The optimization done by most machine-learning algorithms is an approximate form of this survey, based on the assumption or hope that the most relevant portion of the hypothesis space will be around the extremum (MacKay, 1992a; Murphy, 2012, ch. 16; see also Self and Cheeseman, 1987). The underlying necessity of a more extensive survey, however, becomes manifest in many of the obligatory procedures that go together with the training of most machine-learning algorithms, cross-validation being a prominent example (MacKay, 1992b). This leads to a third special characteristic of the optimal predictor machine: it does not need validation sets, test sets, or other data splits; nor does it need cross-validation procedures. Intuitively this is the case because the underlying hypothesis-space survey realizes a sort of full-fledged cross-validation and data partition. It can indeed be proven that one of the internal computations of the machine is mathematically equivalent to doing k-fold cross-validations for *all possible* data splits and k (Porta Mana, 2019; Fong and Holmes, 2020).

Such flexibility and informationally rich output come, of course, at a computational cost. Until some years ago, the cost would have been prohibitive in all but the simplest inferential problems. But today an inference problem involving 13 variates and 700 datapoints, such as the example considered

in the present work, takes less than six hours of computation on an office computer workstation. We discuss computational limitations further in § 4.2.

2.2. Intuitive understanding of the learning algorithm

The calculations by which the optimal predictor machine learns and operates are univocally determined by Cox's theorem[11], which yields quantitative inference rules from self-consistency requirements, and by de Finetti's theorem[12], which further constrains these rules in the case of "generalization from similar cases". These calculations have a very intuitive interpretation.

We consider a patient to be a member of some population of similar past, present, and future patients. Suppose we knew the joint frequency distribution of all possible combinations of predictor and predictand values in such a population. We would then judge the probability for a patient's variate values to be equal to their corresponding population frequency. Pure symmetry considerations lead to this result (Johnson, 1924, Appendix on eduction; Johnson, 1932; de Finetti, 1930; Dawid, 2013; Bernardo and Smith, 2000, §§ 4.2–4.3). The same would be true for conditional and marginal probabilities and frequencies.[13] This population frequency distribution would bound the maximal prognostic power attainable with the given predictors in the population. A higher prognostic power could only be attainable by using additional or different predictors having sharper conditional frequencies for the predictand in the population. Given knowledge of such frequency distribution, there would be no problem of "generalizing" to new patients, because each new patient would already be counted in the known frequencies. An inference algorithm would only need to enumerate and memorize, rather than to learn and generalize.

Learning and generalization come into play because the frequency distribution for the population is unknown: we only have a sample from it, the "learning dataset". Thus we can, at most, assign a probability to each possible frequency distribution. This is precisely what the optimal predictor machine does.

The way in which the machine assigns a probability to each "candidate" true frequency distribution is also intuitive. It combines two factors: (i) how well the candidate fits the sample dataset, (ii) how biologically or physically reasonable the candidate is. The first factor is easily computed: it is the

[11] Cox, 1946, 1961; Pólya, 1954, 1968; Tribus, 1969; Fine, 1973; Rosenkrantz, 1977; Paris, 2006; Snow, 1998; Halpern, 1999; Arnborg and Sjödin, 2001; Snow, 2001; Clayton and Waddington, 2017; see also Hailperin, 1996; and Van Horn, 2003 for a review.

[12] de Finetti, 1930, 1937; Bernardo and Smith, 2000, §§ 4.2–4.3; for a review see Dawid, 2013.

[13] If there were a functional relationship from predictors to predictand, then the predictand value corresponding to the function output would have conditional frequency and probability equal to 1, and all other values having 0. Therefore, this point of view still encompasses a functional relationship as a particular case.

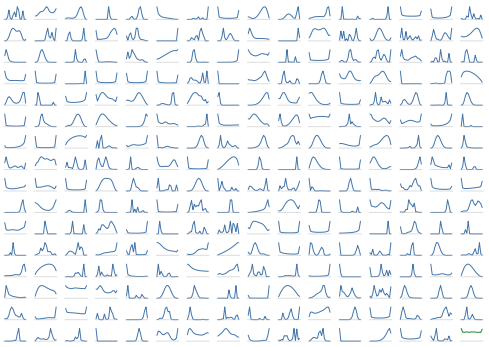


Figure 1: Samples of initially probable candidates of the true population frequency distribution of an integer variate (for example `RAVLT-del` or `RAVLT-rec`, to be introduced in § 3.0).

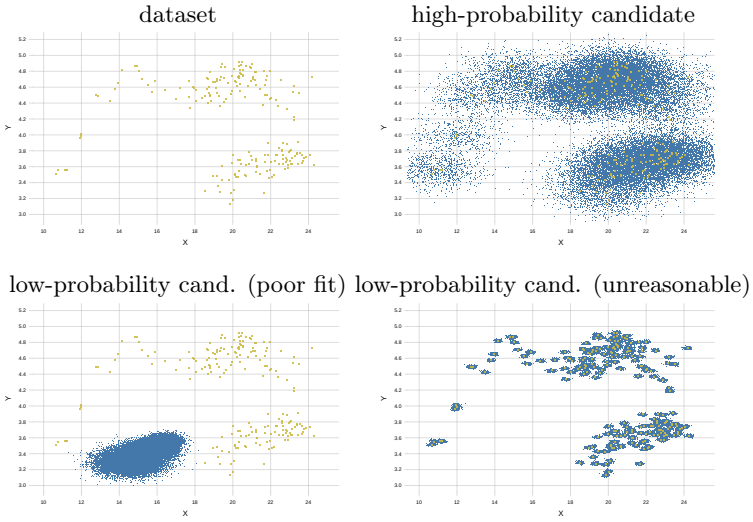


Figure 2: Illustration of the two factors determining the final probability of a candidate population-frequency distribution represented as a blue scatterplot. *Upper-left*: an example dataset (yellow points) with two variates. *Upper-right*: candidate frequency distribution with high final probability; it fits the data and is reasonable. *Lower-left*: candidate distribution with low final probability; it is reasonable but does not fit the data. *Lower-right*: candidate distribution with low final probability; it fits the data but is not reasonable.

joint probability of the dataset if it were sampled from a population having that candidate frequency. The second factor is a prior probability expressing how reasonable that candidate is.[14] The most general natural requirement for "reasonableness" is that the candidate should have some degree of smoothness, owing to physical and biological reasons. This prior probability prevents overfitting and underfitting; in fact, it actually *defines* mathematically what can be considered "overfitting" and "underfitting". Figure 1 shows samples of what the machine has been programmed to consider "reasonable candidates" for the population frequency distribution of a discrete variate. This choice can be altered by the clinician. Note that no frequency distributions are excluded; they are only given higher or lower probabilities.

[14]Some notion of "reasonable candidate" is unavoidable and clearly present in the construction or testing of any inference algorithm. How can we otherwise judge that an algorithm is over- or under-fitting, given that we do not know the ground truth? (if we knew the latter we would not be making inferences.) Such judgement reveals that we have some *preconceived* reasonable distributions in the back of our minds. The fit is either qualitatively compared with this reasonable reference; or it is compared with a known ground-truth, which was in turn chosen because of its similarity with the reasonable reference.

The product of the two factors (i), (ii), normalized, yields the probability of each possible frequency distribution. An illustration of factors (i), (ii) at work is given in fig. 2 for an example problem with two variates. The optimal predictor machine outputs the distribution of these final probabilities in the form of a large sample (the amount is decided by the user) drawn from it. In this form, all other marginal or conditional probabilities and averages of interest are calculated via Monte Carlo integration. This methodology has been successfully demonstrated for Alzheimer's Disease with a smaller number of predictors (Antoniano-Villalobos et al., 2014), and it is the same, but in nonparametric form, used in various inferences about the black holes M87 and Sagittarius A* (Event Horizon Telescope Collaboration, 2019, 2022). Further mathematical and computational details are given in appendix A.

We close this section emphasizing that the inferential steps of the machine, from input to output, consist of, *literally*, no more than the reiterated application of just four inductive-logic rules:

$$\mathrm{p}(A \mid A \text{ and } I) = 1 \qquad \mathrm{p}(\text{not-}A \mid I) = 1 - \mathrm{p}(A \mid I)$$
$$\mathrm{p}(A \text{ and } B \mid I) = \mathrm{p}(A \mid B \text{ and } I)\,\mathrm{p}(B \mid I) \qquad \mathrm{p}(A \text{ or } B \mid I) = \mathrm{p}(A \mid I) + \mathrm{p}(B \mid I) - \mathrm{p}(A \text{ and } B \mid I)\ .$$

3. Example application

In this section we illustrate how the optimal predictor machine is applied in the example case outlined in § 1.0. Although the patients are fictitious, the dataset is real and briefly discussed in the next subsection. The main inferential and decision-making steps are summarized in table 1. Steps 1.–3. are modular: the clinician is free to stop after any of them and use their output in other ways or with other algorithms.

These steps are illustrated in the next three subsections, preceded by an explanation of their rationale. They are presented in chronological order as the clinician would apply them. Steps 1.–3. could also be presented in reverse order; which would be more suited to their logical dependence, as the procedure in each step is actually motivated by the one in the next. We suggest that readers familiar with the principles of clinical decision-making read the following subsections in 3.0–3.1–3.2–3.3 order; whereas readers unfamiliar with these principles read them in 3.0–3.3–3.2–3.1 order.

Table 1: **Main inferential and decision-making steps**

0. Find or build an appropriate dataset of clinical cases comprising values of the predictors and predictand of interest. Datapoints with partially missing values are allowed.

   Input the dataset into the optimal predictor machine and let it infer the joint full-population frequencies of predictors and predictand underlying the dataset.

1. Measure the present patient's predictor values and input them in the optimal predictor machine. Partially missing values are allowed.

2. Assess which conditional statistics of the dataset can be applied to the present patient, and any auxiliary clinical information available. Quantify the latter in a prior probability.

   Input the relevant statistics and auxiliary information for the present patient into the optimal predictor machine.

   Upon request, the machine can now output the final probability of the predictand's true value for the patient, as well as any other probabilities and likelihoods of interest.

3. Assess the clinical courses of action (treatments, more tests, and so on) available for the present patient, and the utility (benefit and loss) of each course of action, depending on each possible predictand value for the patient.

   Input the patient's utilities into the optimal predictor machine. The machine outputs the course of action having maximal expected utility.

   Upon request, the machine can output the probability of gaining different utilities, perform sensitivity analyses for missing data, and do other similar tasks.

3.0. Predictors, predictand, and learning dataset

The dataset used in our example comes from the study by the Alzheimer's Disease Neuroimaging Initiative (ADNI).[15] This longitudinal multicentre study is designed to develop and validate neuroimaging and biochemical biomarkers for the early detection, monitoring, and treatment of Alzheimer's Disease (Petersen et al., 2010). The present dataset consists of 704 ADNI subjects constrained, according to ADNI criteria, to be between 55 and 90 years old. These subjects were chosen to meet the criteria for Mild Cognitive Impairment at their first, baseline assessment, and to have a minimum of two additional following study visits and three MRI examinations. Each subject's diagnostic status was reevaluated at each study visit. This longitudinal diagnostic label is used as the predictand variate `cAD` in our study; it categorizes each subject as either converting to Alzheimer's Disease after the first study visit: `cAD = Y`, or remaining stable with Mild Cognitive Impairment: `cAD = N`. The dataset has 325 subjects (46.2%) with `cAD = Y` and 379 (53.8%) with `cAD = N`. Criteria used for classifying subjects as having Mild Cognitive Impairment or Alzheimer's Disease, as well as ADNI's general criteria for subject inclusion, are described in McKhann et al. (1984); Petersen et al. (2010).

The 12 predictor variates consist of the results from seven cognitive-test measures: a reading test (`ANART`), a word category fluency test (`CFT`), trail-making tests of executive function (`TMTA`, `TMTB`), the immediate-memory, delayed-recall and recognition-subsets of memory function (`RAVLT-imm`, `RAVLT-del`, `RAVLT-rec`); a geriatric depression scale (`GDS`); the presence of the `APOE`-e4 risk allele (Liu et al., 2013); a normalized measure of the sum of left and right hippocampal volume (`HV`); `Age`; `Sex`. Further details about these variates and their selection can be found in Rye et al. (2022). The cognitive and `GDS` variates are integer-valued, hippocampal volume and `Age` are continuous, and `APOE4` and `Sex` are binary. The values of one or two of these predictors were missing for 30 subjects in the dataset.

The optimal predictor machine took less than five hours (on a 16-core Intel Core i9-12900K CPU) to calculate the probability distribution for the possible joint population-frequency distributions of the 13 variates.

Some results can already be visualized after this inference. Figure 3 shows, on the left, the inferred distributions of `RAVLT-del`, `RAVLT-imm`, `GDS`, and hippocampal volume for the subpopulation of patients that will convert to Alzheimer's Disease (red) and the subpopulation that will remain with stable Mild Cognitive Impairment (blue). On the right, the inferred frequency of conversion in the full population is plotted (grey), conditional on the same predictors. The thin curves are 100 samples of highly probable population-frequency distributions; the thicker lines are their means, which are

[15] http://adni.loni.usc.edu. 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.

also the predictive conditional probabilities. The two subpopulations of patients are clearly distinct in the `RAVLT-del`, `RAVLT-imm`, `HV` variates. These predictors can yield probabilities of conversion as high as 70% or as low as 10%. The two subpopulations are practically indistinguishable in the `GDS` variate, which, therefore, always gives very uncertain predictions.

The learning dataset comprises enough data to greatly reduce our uncertainty about the population distributions, as evident from the very narrow spread of the curves. In fact it leads to identical answers, within numerical-computation error, even if we drastically change the prior illustrated in fig. 1, for example favouring more unimodal distributions or more multimodal distributions.

These simple results show the great usefulness of the optimal predictor machine for general medical research.

3.1. Patient's clinical information

The 12 predictor values for our four patients are reported in table 2, top. Note that Curtis's value for the Hippocampal Volume is missing; this is not a problem for the optimal predictor machine. Given these predictor values the optimal predictor machine can output any probabilities of interest to the clinician. Table 2, bottom, reports three probabilities that are important for the step of the next subsection:[16]

- $\mathrm{p}(\texttt{cAD} = \texttt{Y} \mid \texttt{predictors})$: the probability that the patient will convert to Alzheimer's Disease, given the patient's specific predictors and that the patient comes from the same population as the learning dataset.
- $\mathrm{p}(\texttt{predictors} \mid \texttt{cAD} = \texttt{Y})$: the probability that a patient who will convert to Alzheimer's Disease would have these specific predictor values. In other words, the *likelihood*[17] of conversion to Alzheimer's Disease, given the predictors.
- $\mathrm{p}(\texttt{predictors} \mid \texttt{cAD} = \texttt{N})$: the probability that a patient who will remain with stable Mild Cognitive Impairment would have these specific predictor values. In other words, the *likelihood* of stable Mild Cognitive Impairment, given the predictors.

The optimal predictor machine can also answers other questions of interest to the clinician. For instance, what could be the value of Curtis's Hippocampal Volume? The answer is given in fig. 4, which also shows the full-population distribution as comparison (dashed grey); with 95% probability

[16] All relative uncertainties of the results caused by numerical computation error are below 0.8%, Curtis's two likelihoods being an exception at 2%.

[17] $\mathrm{p}(A \mid B)$ is the probability of A given B, as well as the likelihood of B given A (Good, 1950, § 6.1).

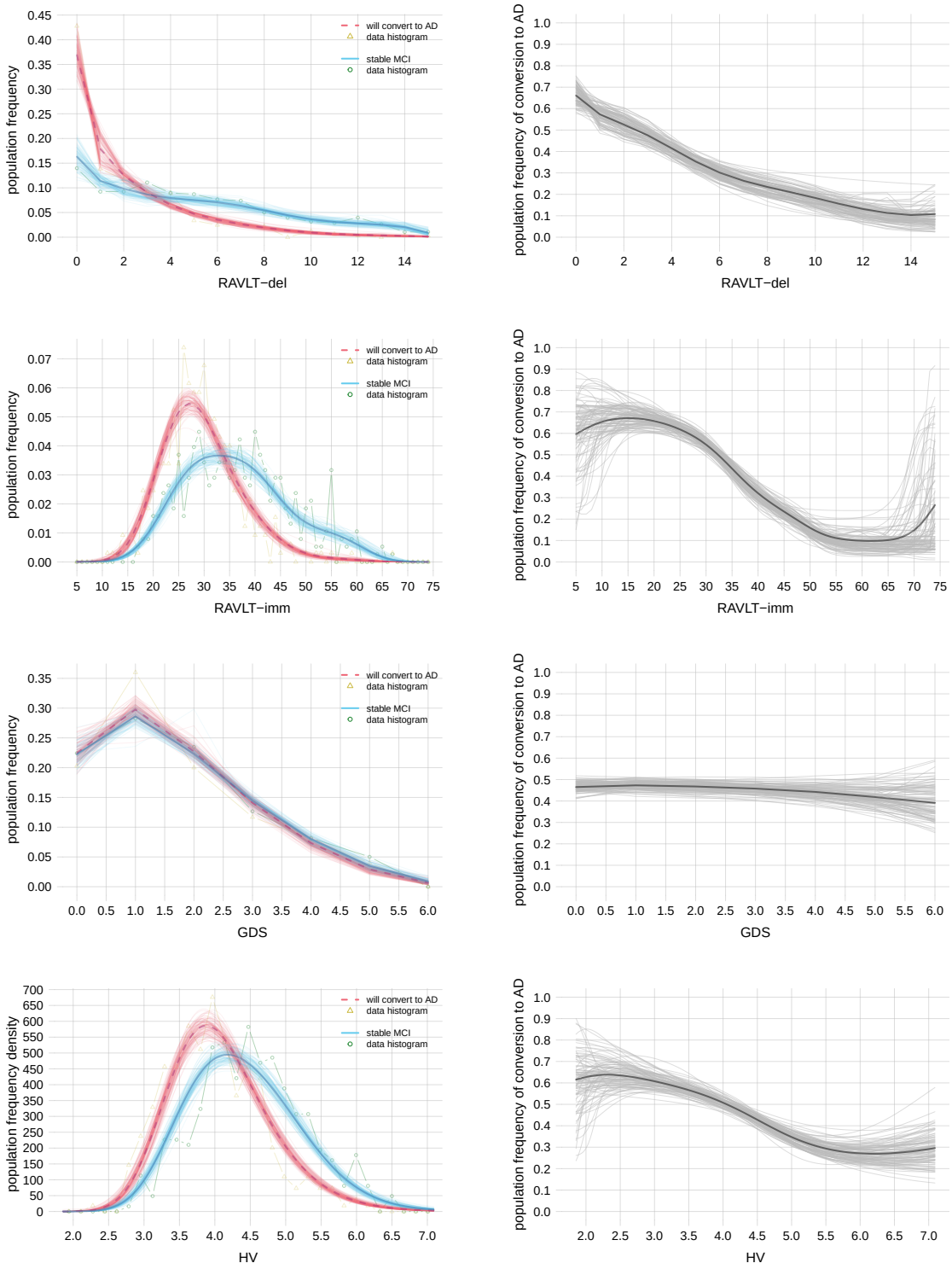


Figure 3: Inferred distributions of some predictor variates, for the subpopulation of patients that will convert to Alzheimer's Disease (red dashed) and the subpopulation with stable Mild Cognitive Impairment (solid blue).

	Olivia	Ariel	Bianca	Curtis
`Age`	75.4	75.4	75.4	63.8
`Sex`	F	F	F	M
`HV`/10^{-3}	4.26	4.26	4.26	[missing]
`APOE4`	`N`	`N`	`N`	`Y`
`ANART`	18	18	18	15
`CFT`	21	21	21	14
`GDS`	3	3	3	2
`RAVLT-imm`	36	36	36	20
`RAVLT-del`	5	5	5	0
`RAVLT-rec`	10	10	10	3
`TMTA`	21	21	21	36
`TMTB`	114	114	114	126
p(`cAD` = `Y` | `predictors`)	0.302	0.302	0.302	0.703
p(`predictors` | `cAD` = `Y`)/10^{-12}	8.97	8.97	8.97	1.14
p(`predictors` | `cAD` = `N`)/10^{-12}	18.6	18.6	18.6	0.343

Table 2: Predictor values for the four patients (see § 3.0), and resulting conditional probabilities.

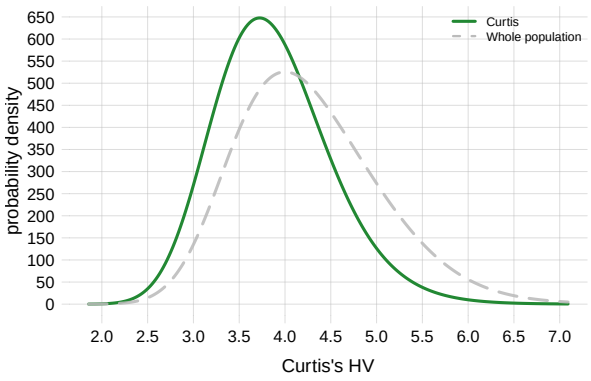


Figure 4: Probability distribution for Curtis's Hippocampal Volume (green). The full-population distribution (dashed grey) is also plotted for reference.

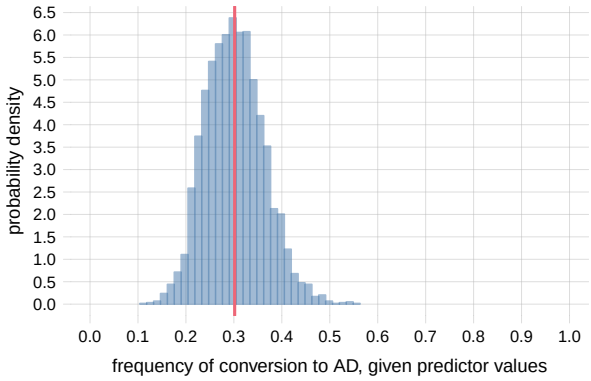


Figure 5: Probability distribution for the frequency of conversion to AD in the subpopulation having Olivia's predictors. The red vertical line is the value of the probability p(`cAD` = `Y`|`predictors`).

Curtis's value is between 2.8 and 5.3, with a median of 3.8. And what is the frequency of conversion to Alzheimer's Disease among the subpopulation having Olivia's, Ariel's, or Bianca's predictors? The answer is given in the histogram of fig. 5: with 95% probability, the fraction of this subpopulation that eventually converts to Alzheimer's Disease is between 0.19 and 0.43; this uncertainty range is due to the limited size of the learning dataset. The probability $\mathrm{p}(\texttt{cAD} = \texttt{Y} \mid \texttt{predictors})$ is equal to the average of such a distribution (e.g. Bernardo and Smith, 2000, §§ 4.2–4.3), provided the patient and dataset can be considered as belonging to the same population.

3.2. Assessment of relevant subpopulation and auxiliary information

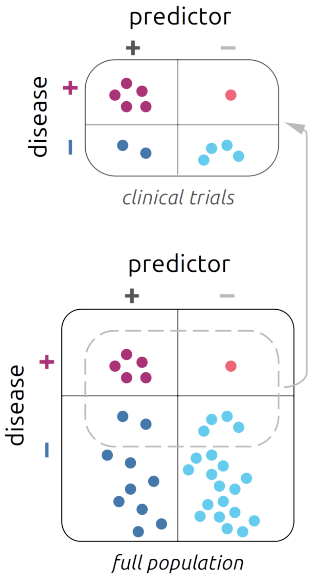
Rationale

As already mentioned, and as will be argued more concretely in the next section, the clinician needs a probability in order to choose a treatment or other course of action for the current patient. This probability is computed by generalizing associations between predictors and predictand hidden in a dataset of similar patients, as discussed in § 2. The way this generalization is made, however, can differ from patient to patient in two respects:

- Only some particular directed associations can be generalized to the current patient, whereas others would be inappropriate to generalize. In some cases, for example when the learning dataset is artificially assembled with balancing or stratification methods, some associations cannot be generalized to any patients at all.
- There can be additional information available for the current patient, for instance some clinical predictors not included in the learning dataset, or other "softer" information such as family history or geographic background.

There is no sharp separation between these two items. The presence of additional information often automatically implies that some associations cannot be generalized from the learning dataset to the current patient.

Let us explain with a familiar example why particular associations cannot be generalized. Most students of medicine learn about the *base-rate fallacy* (Bar-Hillel, 1980; Jenny et al., 2018; Sprenger and Weinberger, 2021; Matthews, 1996). Consider a large set of clinical trials, illustrated in the upper table on the side, where each dot represents, say, 10 000 patients. In this sample dataset it is found that, among patients having a particular value "+" of some predictors (left column), 71.4% of them (or 5/7, upper square) eventually developed a disease. The fallacy lies in judging that a new real patient from the full population, who has predictor value "+", also has a 71.4% probability of developing that disease. In fact, *this probability will in general be different*. In our example, it is 33.3% (5/15), as can be seen in the lower table illustrating the full population. This difference would be noticed as soon as the inappropriate probability was used to make prognoses in the full population. A similar situation happens for the predictor value "−".



There is a discrepancy in the conditional frequencies of predictand given predictors, between the sample dataset and the full population, because the proportion of positive vs negative disease cases in the latter has some value, 16.7%/83.3% in our example, whereas the samples for the trials (dashed line in the lower table) were hand-chosen so as to have a 50%/50% proportion. This sampling procedure is called "class balancing" in machine learning (Provost, 2000; Drummond and Holte, 2005; Weiss and Provost, 2003). More generally this discrepancy can appear whenever a population and a sample dataset from it do not have the same frequency distribution for the predictand. In this case, we cannot rely on the probabilities of "predictand given predictors" obtained from the sample dataset, which we symbolically write as

$$\mathrm{p}(\texttt{predictand} \mid \texttt{predictors}, \texttt{dataset}) \tag{1}$$

A little counting in the side figure reveals, however, that other frequencies may be relied upon. Consider the full population. Among all patients who developed the disease, 83.3% of them (or 5/6, upper row) had the predictor value "+", while among those who did not develop the disease, 33.3% (or 1/3, lower row) had the predictor value "−". *And these frequencies are the same in the sample dataset*. These frequencies from the clinical trials can therefore be used to make a prognosis using Bayes's theorem. For brevity, denote the predictors by X, the predictand by Y, the dataset or trials

by D, and the full-population base rate by B. Bayes's theorem yields

$$\mathrm{p}(Y \mid X, D, B) = \frac{\mathrm{p}(X \mid Y, D) \cdot \mathrm{p}(Y \mid B)}{\sum_{Y} \mathrm{p}(X \mid Y, D) \cdot \mathrm{p}(Y \mid B)} \tag{2}$$

In our example we find

$$\begin{aligned} \mathrm{p}(Y = + \mid X = +, D, B) &= \frac{\mathrm{p}(X = + \mid Y = +, D) \cdot \mathrm{p}(Y = + \mid B)}{\mathrm{p}(X = + \mid Y = +, D) \cdot \mathrm{p}(Y = + \mid B)\mathrm{p}(X = + \mid Y = -, D) \cdot \mathrm{p}(Y = - \mid B)} \\ &\approx \frac{0.833 \cdot 0.167}{0.833 \cdot 0.167 + 0.333 \cdot 0.833} = 0.33 \end{aligned} \tag{3}$$

which is indeed the correct full-population frequency.

If the samples of the clinical trials had been chosen with the same frequencies as the full population (no "class balancing"), then the probability $\mathrm{p}(\texttt{predictand} \mid \texttt{predictors}, \texttt{dataset})$ from the dataset would be the appropriate one to use. But the probabilities $\mathrm{p}(\texttt{predictors} \mid \texttt{predictand}, \texttt{dataset})$ together with Bayes's theorem as in eq. (2) would also lead to exactly the same probability. We thus see that *using the probabilities*

$$\mathrm{p}(\texttt{predictors} \mid \texttt{predictand}, \texttt{dataset})$$

from the dataset is preferable to using $\mathrm{p}(\texttt{predictand} \mid \texttt{predictors}, \texttt{dataset})$. The former yield the same results as the latter when use of the latter is appropriate, and allow us to apply corrections when use of the latter is inappropriate. The superiority of using $\mathrm{p}(\texttt{predictors} \mid \texttt{predictand}, \texttt{dataset})$ probabilities (called "generative" in machine learning, see e.g. Murphy, 2012, § 8.6) is illustrated with a toy example in table 3.

The use of dataset probabilities different from $\mathrm{p}(\texttt{predictand} \mid \texttt{predictors}, \texttt{dataset})$ can be necessary even when the dataset has statistics identical with the population it is sampled from. Typical cases are the prognosis of a patient that comes from a peculiar subpopulation or even from a different population (Lindley and Novick 1981; Quintana et al. 2017; Sox et al. 2013, ch. 4; Hunink et al. 2014, ch. 5). For instance, the first case happens when the clinician has additional information not included among the predictor variates, such as the result of an additional clinical test, or family history; the second case happens when the patient comes from a different geographical region. There is of course no sharp distinction between these two cases.

What is important is that, in either case, it can still be possible to use statistical information from the sample dataset to make prognoses. It is sufficient that some *conditional* statistics may be applicable to the specific patient. For a patient coming from a different region, for example, it may be

Table 3: **superiority of the "predictors | predictand" (or "generative") approach**

We split our learning dataset into two subsets:

- One with 361 subjects and a ratio of 29.9%/70.1% of subjects with `cAD = Y` vs `cAD = N`.
- One with 343 subjects and a ratio of 63.3%/36.7% of subjects with `cAD = Y` vs `cAD = N`. This subset is used as a fictive full population.

This partition was made with no systematic sampling of any variates except the predictand `cAD`.

After training on the learning dataset, we make a prognosis for each of the 343 "new" patients, through four separate approaches: (a) using the probabilities p(`predictand` | `predictors`, `dataset`), as typical of machine-learning algorithms; (b) using p(`predictors` | `predictand`, `dataset`) together with the base rate, as explained above; (c) tossing a coin; (d) always prognosing "`cAD = Y`", which guarantees 63.3% correct prognoses owing to the base rate of the full population. Finally, the accuracies (number of prognoses giving more than 50% probability to the correct outcome) of these four approaches are calculated. Here are the results from lowest to highest:

predictand | predictors	coin toss	always predict conversion	predictors | predictand & base rate
37.3%	50%	63.3%	73.2%

The "predictand | predictors" approach ("discriminative" in machine-learning parlance) leads to worse results than a coin toss because of its underlying base-rate fallacy. The "predictors | predictand" approach ("generative" in machine-learning parlance) leads to better results than simply always prognosing the most common base-rate outcome; this shows that the dataset can still provide useful statistical information despite its mismatched base rate. Inference algorithms that only yield "predictand | predictors" outputs, unlike the optimal predictor machine, are incapable of extracting this useful information.

judged that the conditional probabilities $\mathrm{p}(\texttt{predictand} \mid \texttt{predictors}, \texttt{dataset})$ still apply. In other words, the patient may still be considered a member of the subpopulation having those specific predictor values. Using more technical language we say that a new patient can be considered *exchangeable* with the patients constituting the dataset, but only conditional on particular variates. See Lindley (2014, especially around §§ 7.3, 8.6; 1981) for a clear and logically impeccable presentation not obscured by technical language (more technical references are de Finetti 1930, 1937; Dawid 2013; Bernardo and Smith 2000, §§ 4.2–4.3, 4.6; see also Malinas and Bigelow 2016, Sprenger and Weinberger 2021 about confounding and Simpson's paradox, to which this topic is tightly related).

This topic is complex and of extreme importance for inference, but its detailed study is not the goal of the present work. Our main point here is that population variability and auxiliary clinical information are important factors that differentiate patients, and a personalized approach ought to take them into account. The method here presented does this naturally, allowing a great flexibility in selecting which statistical features of the sample dataset should be used for each new patient, and the integration of auxiliary clinical information in the form of a prior probability. As discussed in § 3.1, the optimal predictor machine allows us to quickly calculate conditional probabilities $\mathrm{p}(Y \mid X, \texttt{dataset})$ for any desired variate subsets Y and X required by the patient's relevant population.

Application to the example study

In our example of § 1.0, all statistics of the dataset are considered relevant for Olivia, Bianca, and Curtis. For these patients the clinician can therefore use Bayes's theorem with the likelihoods of table 2 and the dataset conversion rate of 0.463 – or equivalently directly the probabilities $\mathrm{p}(\texttt{cAD} = \texttt{Y} \mid \texttt{predictors}, \texttt{dataset})$ provided in the same table.

For Ariel, however, the clinician judges that a different base rate or prior probability of conversion should be used, equal to 65%, because of her different geographical origin and family history. In her case the clinician uses Bayes's theorem with the likelihoods of table 2 and the prior probability of 0.65.

The final probabilities of conversion to Alzheimer's Disease for our four patients are reported in table 4. Note how the final probability for Ariel is higher than that for Olivia and Bianca, even if the predictor data are the same for these three patients.

	Olivia	Ariel	Bianca	Curtis
initial probability $\mathrm{p}(\texttt{cAD} = \texttt{Y} \mid \texttt{aux info})$	0.463	0.65	0.463	0.463
final probability $\mathrm{p}(\texttt{cAD} = \texttt{Y} \mid \texttt{predictors}, \texttt{dataset}, \texttt{aux info})$	0.302	0.47	0.302	0.703

Table 4: Final probabilities of conversion computed from dataset and auxiliary information

3.3. Assessments of treatments and benefits; final decision

Rationale

A crucial point in clinical decision-making is this: the clinician needs to assess, not the presence (present or future) of a disease, but the *risk* of its presence. Is there a difference between these two problems? and why is the difference important?

In clinical practice, we can rarely diagnose or prognose a medical condition with full certainty. Perfect classification is therefore impossible. But also a "most probable" classification, which may be enough in other contexts, is inadequate in clinical ones. The problem is that the clinician has to decide among different courses of action, such as different treatments, more tests, and so on, and the optimal one depends on *how probable* the medical condition is, not just on whether it is more probable than not.

Two examples illustrate this point. Suppose there is a dangerous treatment that extends the patient's lifetime by 1 year if the disease is on its course, but shortens the patient's lifetime by 5 years if the disease is not present. Also suppose that some algorithm tells the clinician whether the disease's presence is "more probable than not", given some predictor values; in which case the clinician administers the dangerous treatment. It turns out that 60 out of 100 treated patients having these same predictor values eventually develop the disease, so "more probable than not" is correct. However, the final result is that the clinician has added $1 \times 60 = 60$ years but also *subtracted* $5 \times 40 = 240$ *years* from the combined lifespans of the treated patients! The conclusion is that the treatment cannot be prescribed just because the disease is "more probably present than not". As an opposite example, suppose that a less dangerous treatment extends the patient's lifespan by five years if the disease is on its course, but shortens it by one month if the disease is not present. In this case, it may be advisable to undergo the treatment even if the disease is *less* probably present than not. If the clinician administer the treatment to 100 similar patients, and 20 of them develop the disease, then the clinician has added $5 \times 20 = 100$ and subtracted $\frac{1}{12} \times 60 = 5$ years to their combined lifespans.

In both examples, it is clearly important to assess the *probability* – having precise connections with the population frequency – that the patient will develop the disease. In the first example, the

treatment should only be administered if the probability is higher than 83.3%; in the second, it can be administered if the probability is at least 1.6%. The optimal predictor machine, as explained in the previous sections, tells the clinician the specific probability for the current patient.

But the choice between treatments depends not only on the probability of the medical condition. Here is where differences between patients vary and matter the most. Consider again the second example above, about the less dangerous treatment. Let us add that the treatment would extend the lifespan by five years, but would also somewhat worsen the quality of life of the patient and of the patient's family. Suppose our patient is quite old and tired, has had a happy life, and is now looking with a peaceful mind towards death as a natural part of life. Such a patient may prefer to forego the bother of the treatment and the additional five years, even if the probability for the disease is quite high.

The benefits of the different treatments, and the probability thresholds at which one treatment becomes preferable to another, must therefore be judged and quantified primarily by the patient. Utility theory and maximization of expected utility allow clinician and patient to make such judgements and decisions in a coherent way (Sox et al. 2013; Hunink et al. 2014; Lusted 1968; see also the clear and charming exposition by Lindley 1988, and O'Hagan et al. 2006).

We summarize the main, patient-dependent procedure for decision-making, and show how our computations so far fit perfectly with it.

The clinician first assesses and lists the mutually exclusive courses of action available for the specific patient. These could be preventive or curative treatments, more tests, doing nothing, and so on. Often there are *sequences* of decisions available, but the utility framework can be applied to them as well (see references above and Raiffa, 1970). In the present work we are calling these heterogeneous alternatives simply "treatments" for simplicity (see footnote 2, p. 3). The list treatments is already patient-dependent: some alternatives may not be medically suitable (say, owing to allergies or other clinical conditions), some may be economically too costly, and so on.

Each treatment will have different consequences, which additionally depend on the patient's unknown clinical condition of interest. A treatment may have some consequences if the patient has or will develop the disease, and different consequences otherwise. The patient quantifies, with the clinician's guidance, the benefits and costs – technically called "utilities" – of such possible consequences. The quantification of utilities is not within the scope of the present work. The references cited above offer guidelines and rules for numerically translating factors such as quality of life and expected lifespan into utilities.

The treatments, uncertain clinical conditions, and the quantified utilities U of their consequences can be organized into a table of this form:

	clinical condition a	clinical condition b	...
treatment α	$U_{\alpha a}$	$U_{\alpha b}$	...
treatment β	$U_{\beta a}$	$U_{\beta b}$	...
...	...	...	...

which can be compactly represented by a so-called *utility matrix* $\big(U_{ij}\big)$, the row index i enumerating the treatments, and the column index j the clinical conditions. Note that the number of possible treatments and clinical conditions do not need to be equal; generally, they are not.

The *expected utility* $\bar{U}_i$ of a treatment i is calculated as the expectation of its utilities $U_{ia}, U_{ib}, \ldots$ with respect to the probabilities $\mathrm{p}(a), \mathrm{p}(b), \ldots$ of the clinical conditions $a, b, \ldots$:

$$\bar{U}_i \coloneqq U_{ia}\,\mathrm{p}(a) + U_{ib}\,\mathrm{p}(b) + \cdots \tag{4}$$

Note that this corresponds to a matrix multiplication between the matrix of utilities and the vector of probabilities.

Finally, the recommended treatment is the one having *maximal expected utility*.

Application to the example study

At present there are no cures for Alzheimer's Disease, although some recent pharmacological agents are shown to delay onset of pathology related to Alzheimer's Disease[18]. But for the sake of our case study let us imagine that in the near future there are three mutually exclusive treatment options for prevention or retardation of the disease; call them β, γ, δ, the simple option of "no treatment" being denoted by α. The clinical conditions to be considered are just two: the patient will convert to Alzheimer's Disease ($\texttt{cAD} = \texttt{Y}$), or will remain with stable Mild Cognitive Impairment ($\texttt{cAD} = \texttt{N}$).

We have therefore $4 \times 2 = 8$ possible consequences of the four treatments depending on the two clinical conditions. Our four patients and the clinician quantify the utilities, arriving at the utility matrices shown in table 5, top. Olivia, Ariel, and Curtis quantify the benefits of the treatments in exactly the same way, but Bianca's quantification differs slightly, because of the interaction of the treatments with several allergies and additional clinical conditions, as explained in § 1.0.

The probabilities for the two medical conditions are those found in the previous subsection, table 4. For brevity, we denote just by $\mathrm{p}(\texttt{cAD})$ the probability of conversion given a patient's predictor values,

[18] e.g. lecanemab, a monoclonal antibody infusion given every two weeks, targeting amyloid beta plaques; see https://www.fda.gov/news-events/press-announcements/fda-grants-accelerated-approval-alzheimers-disease-treatment.

Utility matrices

	Olivia	Ariel	Bianca	Curtis
	cAD	cAD	cAD	cAD
	N Y	N Y	N Y	N Y
treatment α	10 0	10 0	10 0	10 0
treatment β	9 3	9 3	8 3	9 3
treatment γ	8 5	8 5	7 5	8 5
treatment δ	0 10	0 10	0 10	0 10

Expected utilities and optimal treatments

	Olivia	Ariel	Bianca	Curtis
treatment α	6.98	5.27	**6.98**	2.97
treatment β	**7.19**	6.16	6.49	4.78
treatment γ	7.09	**6.58**	6.40	5.89
treatment δ	3.02	4.73	3.02	**7.03**
optimal	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$	$\boldsymbol{\alpha}$	$\boldsymbol{\delta}$

Table 5: Utility matrices, expected utilities, and optimal treatments for our four patients

and by $\mathrm{p}(\texttt{sMCI}) \equiv 1 - \mathrm{p}(\texttt{cAD})$ the complementary probability of stable Mild Cognitive Impairment, given the same predictor values. The expected utilities of each treatment for each patient can then be easily computed. For example, for Olivia the expected utility of treatment β is

$$\begin{aligned} \bar{U}_\beta &= 9 \cdot \mathrm{p}(\texttt{cAD} = \texttt{N} \mid \texttt{predictors}, \texttt{dataset}, \texttt{aux info}) + 3 \cdot \mathrm{p}(\texttt{cAD} = \texttt{Y} \mid \texttt{predictors}, \texttt{dataset}, \texttt{aux info}) \\ &= 9 \cdot (1 - 0.463) + 3 \cdot 0.463 = 7.19 \end{aligned} \tag{5}$$

The results for all patients are reported in table 5, bottom, with the maximal expected utilities in **boldface**.

A summary of the clinician's inputs, the optimal predictor machine's outputs, and the final decisions is given in table 6 on page 34.

3.4. Additional information provided by the optimal predictor machine

As discussed in § 2, the output of the optimal predictor machine concerns the full population of past, present, and future patients, with all its statistics. This output can therefore be used for additional purposes such as resource planning, imputation of missing data, sensitivity checks, and the investigation of each predictor's importance in the prognosis. We briefly discuss these possible uses.

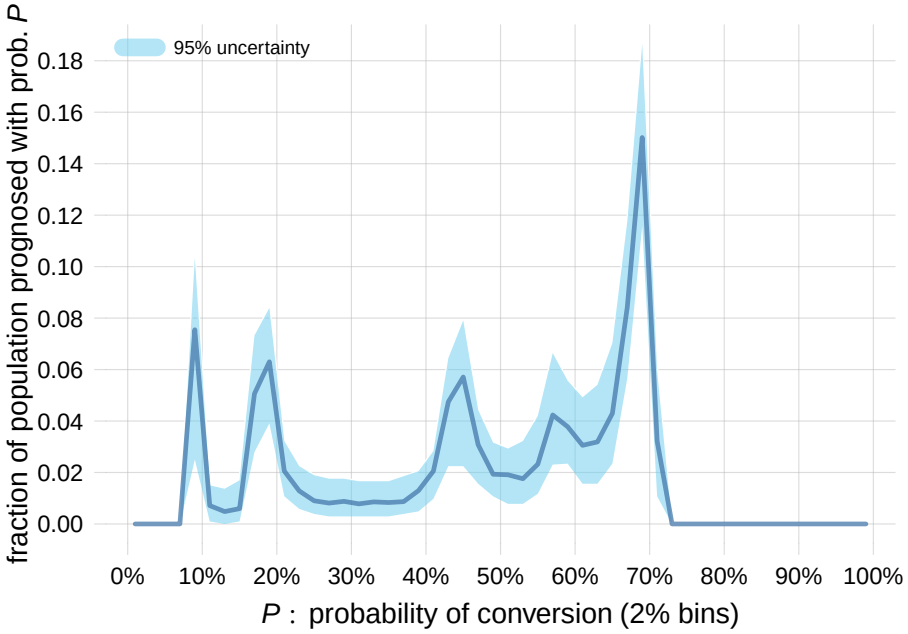


Figure 6: Possible distribution of prognostic probabilities of conversion in the full patient population

Resource planning

Let us ask the following question: if the learning dataset were representative of the full population, then how often, in the long run, would a clinician prognose a conversion to Alzheimer's Disease with probability between 0%–2%, or 2%–4%, and so on, with 50 bins up to 98%–100%?

The optimal predictor machine can answer this question probabilistically; the answer is plotted in fig. 6. Note that the calculation assumes that the optimal predictor machine will not be regularly updated with new patients' data (the calculation could also be made with the opposite assumption). The light-blue bands are 95% uncertainty coverage intervals [19]; this uncertainty comes from the fact that we are not certain about the full-population frequencies. We see that it is very improbable that many patients will be prognosed with probabilities around 30%, smaller than 5%, or larger than 75%. The full population is likely to be grouped into three to five "conversion-probability clusters", as evident from the peaks.

This kind of information allows us to make other forecasts of various kinds. For instance, it would be useful to forecast in which proportions the treatments $\alpha, \beta, \gamma, \delta$ (see § 3.3) will be prescribed,

[19] for terminology see JCGM (2008, C.2.30). A p% coverage interval or credible interval is an interval containing the true value with p% probability. Note that it is different from a "confidence interval", which cannot be interpreted in such a simple way (Pratt, 1961, pp. 165–166; Jaynes, 1976; MacKay, 2005, § 37.3)

assuming that the full population has on average the utility matrix of Olivia, table 5. We find these coverage intervals with a 90% probability that the true future proportion will be within each:

$$\alpha\text{: } 21\%\text{–}28\% \qquad \beta\text{: } 2\%\text{–}6\% \qquad \gamma\text{: } 31\%\text{–}42\% \qquad \delta\text{: } 31\%\text{–}40\% \tag{6}$$

Again, despite the obvious uncertainties, we can be quite sure that treatments γ, δ will be prescribed more often than α, and that β will be only prescribed in 5% of cases. Semi-quantitative forecasts such as this can be very useful for resource planning.

This kind of analysis, as recommended by Smith and Winkler (2006), can also be made for a single patient to avoid the "optimizer's curse". It tells the clinician how much and with which probability the final utility that a patient will gain could deviate from optimality, allowing the clinician to honestly inform the patient about the possibility and extent of unwanted outcomes.

Imputation of missing data

As mentioned in §§ 1.1 and 2, the optimal predictor machine treats all variates of the learning dataset on the same footing. This is why it can output probabilities "predictand | predictors", "predictors | predictand", and other combinations with equal ease, exploiting them to correct subpopulation mismatches as discussed in § 3.2. This feature allows us to impute missing data for one or more patients, giving a probability distribution of what those data could be. We saw and example of this possibility in § 3.1 for Curtis's Hippocampal Volume, fig. 4. Such imputation can be done at prognostic time for sensitivity checks, as discussed in more detail in the next section. The imputation can also be done a posteriori, possibly years later, when the actual predictand value becomes known. This can be useful for many purposes, for instance, for the comparison of biological hypotheses.

Sensitivity checks

The imputation of missing data at prognostic time is useful for various kinds of sensitivity analysis regarding the current patient. Let us consider for instance the case of Curtis, whose value of Hippocampal Volume is missing (table 2). His clinician thus wonders if the acquisition of this value would lead to a different and more beneficial treatment choice. The optimal predictor machine can answer this question by outputting the probability distribution of Curtis's expected utilities (table 5) *if the Hippocampal Volume had been known.* The result is that the expected utilities for the four treatments in Curtis's case must be within the following coverage intervals with 90% probability:

$$\alpha\text{: } 2.97\%\text{–}2.98\% \qquad \beta\text{: } 4.78\%\text{–}4.79\% \qquad \gamma\text{: } 5.89\%\text{–}5.90\% \qquad \boldsymbol{\delta}\textbf{: } \mathbf{7.02\%\text{–}7.03\%} \tag{7}$$

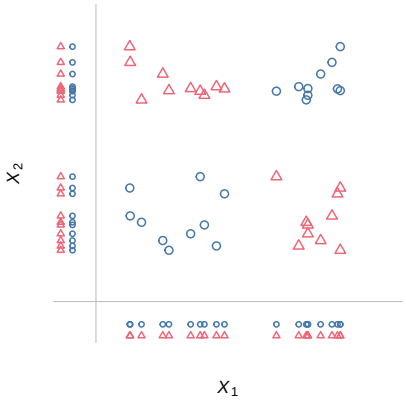
(the four corresponding probability histograms, if plotted jointly, would look like distinct vertical lines). It is clear that knowledge of the Hippocampal Volume is extremely unlikely to change Curtis's optimal treatment from δ. Considering that the negligible information gained would not outweigh the economic costs (involving an MRI-scan) for obtaining this predictor, the clinician decides to proceed without it.

Predictor importance

The question about Curtis in the previous subsection can be generalized to a whole population. Predictors that are too invasive or too expensive to obtain, but that are uninformative for the prognosis, could be dropped altogether. So how important, in general, is each predictor in prognosing the conversion to Alzheimer's Disease?

As posed, this question is too vague (ill-posed) because it does not exactly specify how a predictor is used, and what "important" means. Let us see why these details matter.

The schematic picture on the side illustrates the necessity of specifying a predictor's context. Individuals in this population can be either blue circles ○ or red triangles △, and have two predictors X_1 and X_2. Predictor X_1, if used by itself, is worthless in distinguishing the two subpopulations, because these have identical marginal distributions (depicted underneath the grey horizontal line). If used in conjunction with X_2, however, predictor X_1 allows us to identify an individual's subpopulation with full certainty, as is clear from the two-dimensional view. It is therefore an essential predictor in this case: dropping it would lead to a complete loss of predictive power. An analogous discussion holds for X_2 in the present case. The converse can also happen (not illustrated): a predictor might be "good" if used by itself, and yet it might be discarded without any loss if used in combination with others.



In our question about a predictor's importance, we want to know what happens if the predictor is dropped from the set of all predictors.

Regarding the meaning of "importance" or "prognostic power", we must specify a relevant metric, and predictors could be ranked differently by different metrics. From our discussion so far it is clear that in clinical decision-making the canonical metric is the final *expected utility* – and therefore the choice of optimal treatment – which a predictor's presence or absence leads to (see § 3.3). This point was illustrated with Curtis's example in the previous subsection. What if we want to make a similar assessment, not for a single patient, but for the full population? which utility matrix should we use? It can be proved, again from decision-theoretic principles, that the population average

of all utility matrices should be used in this case (cf. Dyrland et al., 2022a, § 4.1). This seems a quantity very difficult to assess, but it can also be shown (Dyrland et al., 2022a, § 4.2) that even a semi-quantitative assessment leads to better results than using some other general-purpose metric.

The optimal predictor machine allows us to compute the expected value of virtually any prognostic-importance metric, and for any subset of predictors available in the dataset. This computation has moreover two properties of paramount importance: (a) *the prognostic power of a set of predictors found with the optimal predictor machine is the maximum possible obtainable by* ***any*** *inference algorithm*, or in other words it is an intrinsic property of that set of predictors; (b) *the optimal predictor machine achieves this maximum power*. Thus, if the optimal predictor machine says that the accuracy obtainable with a given set of predictors is 70%, then we know that no other inference algorithm can reach a higher accuracy than 70%; inference algorithms that reach lower accuracy can in principle be improved upon. The optimal predictor machine, by construction, will reach this accuracy. Note that we mean accuracy in the long run, over the full population; an inference algorithm could reach higher accuracies in some test dataset thanks to sampling fluctuations; in fact this is bound to happen from time to time.[20]

Let us illustrate this kind of "predictor importance" assessment for our dataset. We use (a) two metrics: the accuracy and the mutual information (Shannon, 1948; Cover and Thomas, 2006) between a set of predictors and the `cAD` predictand; (b) 27 different sets of predictors:

- every predictor, used individually (12 sets);
- all cognitive-test predictors used together, jointly with information about depression (`GDS`) and demographics (`Age` and `Sex`).
- `APOE4` and Hippocampal Volume, jointly with demographic information;
- all predictors jointly excluding one, each single predictor being excluded in turn (12 sets);
- all predictors jointly.

Use of the accuracy assumes that the population of patients has only two available treatments having average utility matrix $\left[\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right]$. Mutual information is a model-free measure of the relation between two sets of variates, with diverse operational interpretations (MacKay, 2005; Woodward, 1964; Minka, 2003; Good, 1961; Good and Toulmin, 1968; Kelly, 1956; Kullback, 1978) and international standards (ISO, 2008). A set of predictors and a binary variate (such as our conversion to Alzheimer's Disease) have a mutual information of 1 Sh if and only if there is a non-constant deterministic function from the former to the latter.

[20]The optimal predictor machine can also calculate, with a somewhat expensive computation, the size of such fluctuations, given the size of the test dataset.

Our specific questions are the following: "What is the expected value of the accuracy for the next new patient, if we use the given set of predictors?" and "What is the mutual information between the given set of predictors and the predictand, given the presently available data?".

The answers to these questions are reported in fig. 7, ordered from bottom to top according to increasing metric. The ordering of mutual information and accuracy agree within the uncertainty of the numerical computation (Monte Carlo integration). The latter is reported as coverage intervals of $\pm$ two standard deviations.

The plots reveal several findings, *valid within the population selected for the dataset*, which can be compared with the analysis in Rye et al. (2022, see especially Fig. 3 and Table 3):

- The set of 12 predictors considered in the present work and in Rye et al. (2022) can at most yield a prognostic accuracy of around $67.7\% \pm 0.7\%$ over the full population, for any inference algorithm. This fact agrees with the (completely independent) findings in Rye et al. (2022), where a maximal accuracy of 68.3% *on a test dataset* was found using an ensemble model. The present analysis also shows that the ensemble model managed to achieve the maximal accuracy possible with these predictors (but see § 4 for limitations of that model).
- The mutual information using all 12 predictors is quite low at (0.140 ± 0.008) Sh, indicating that we cannot reasonably consider the predictand to be an approximate function of the predictors (0 Sh corresponds to a coin toss, 1 Sh to a perfect function). Machine-learning algorithms based on functional regression, such as neural networks, are therefore not appropriate for this prognostic problem.
- `APOE4`, `GDS`, `Age`, `Sex`, and to some degree `ANART` are poor predictors (within this population) when used alone and when used in combination with all other predictors. The latter point is evident from the fact that the mutual information and accuracy of the combined predictors barely decreases if any one of these four predictors is omitted.
- The combined cognitive and demographic variates are better predictors than the joint use of Hippocampal Volume, `APOE4`, and demographic variates.
- `RAVLT-imm`, `RAVLT-del`, and to a lesser degree `RAVLT-rec` are good predictors, both when used alone and when used jointly with all other predictors. Hippocampal Volume is a poorer predictor than any of the `RAVLT` when used alone, and likely also when used in combination with all others (contrast this with Rye et al., 2022). This last finding is also clear in Curtis's case: fig. 8 shows that his probability of conversion to Alzheimer's Disease, given his current predictors, would practically be the same for all values of Hippocampal Volume; and it would probably be the same even if the learning dataset contained more points.

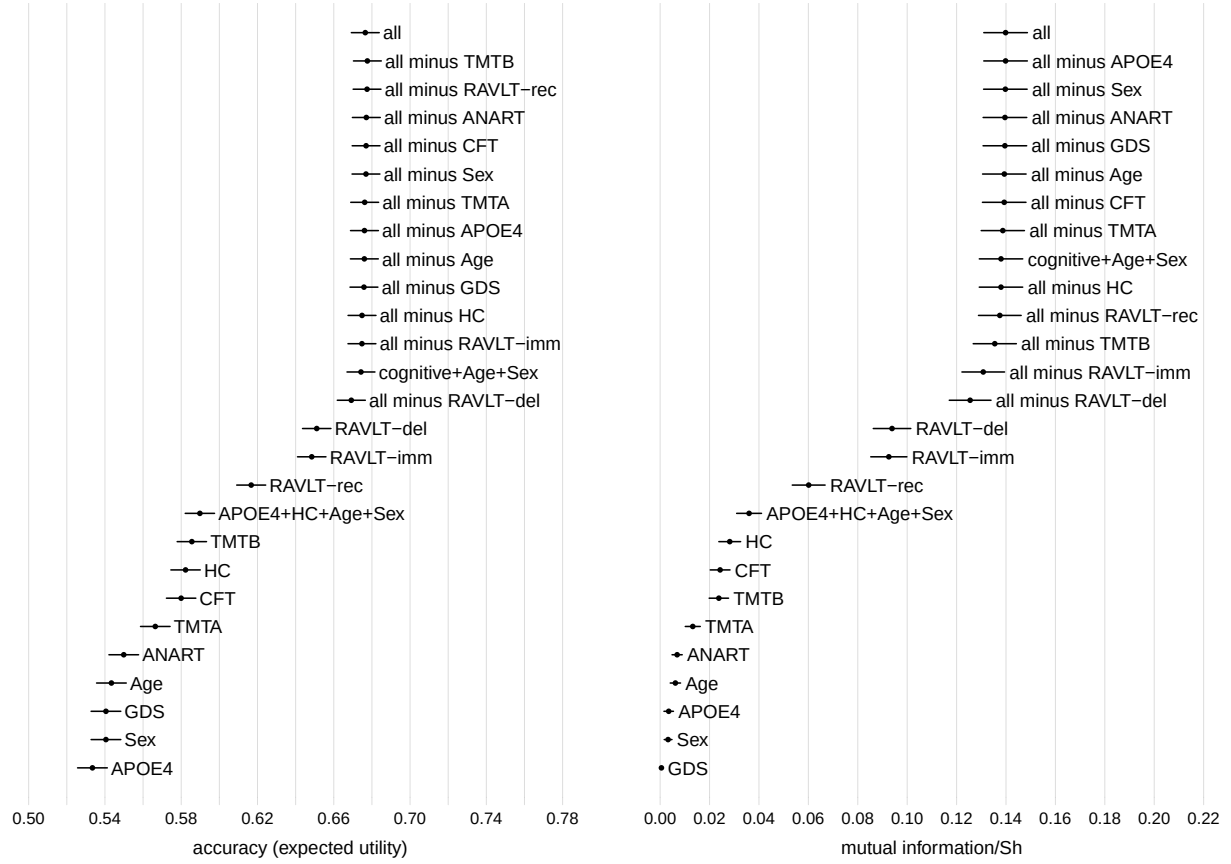


Figure 7: Expected accuracy for the next new patient (left), and mutual information (right), of several sets of predictors for the prognosis of conversion to Alzheimer's Disease. Each graph has been vertically ordered according to increasing values; the two rankings agree within the respective uncertainties. The `all` predictor set is mathematically guaranteed to be optimal according to both metrics and has therefore been ranked first. Bars show the uncertainty interval (± two standard deviations).

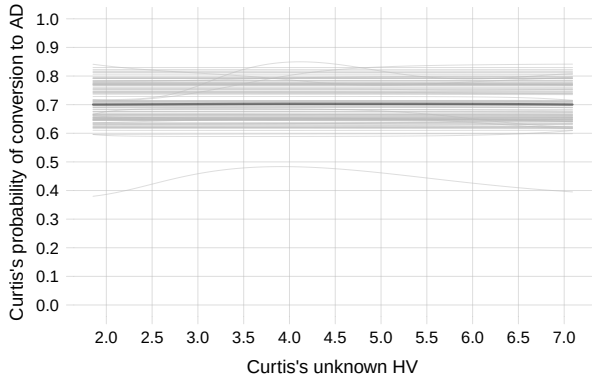


Figure 8: Probability of `cAD = Y` for Curtis, given Curtis's known predictors and different possible values of his unknown Hippocampal Volume. The thinner curves are 100 probable samples of how this probability would change with a larger learning dataset. Compare this figure with fig. 3, p. 16, bottom-right.

The optimal predictor machine shows that the omission of any one of the 12 predictors, except `RAVLT-del` and possibly `RAVLT-imm`, does not lead to an appreciable decrease in accuracy (relative decrease of 0.3% or less) or in mutual information (relative decrease of less than 3%). This puts the prognostic-importance analysis of Rye et al. (2022) into perspective. The exact quantification of these subtle differences is computationally quite expensive, and we did not carry it out further.

4. Discussion

Which requirements does a personalized approach to prognosis and treatment impose on assistive computational technology? This is an important question, because with the increasing amount of statistical clinical data and clinical predictors available for medical care, assistive computational technology is today not merely a useful option, but a necessity in clinical practice.

In the present work we started from the perspective of the clinician's ultimate task, *decision-making under risk*, and saw that patients' differences relevant to prognosis and treatment can be approximately divided into three categories:

- differences in the values – and availability – of a core set of clinical predictors, for which we have population-wide statistical information;
- differences in the availability and values of auxiliary and usually semi-quantitative clinical information, such as geographical or family background;
- differences in the availability and values or "utilities" of clinical courses of action, such as preventive treatments or further tests; such values can have a highly variable, patient-dependent subjective component.

Luckily there is a theory that takes into account and integrates these differences towards the final goal: Decision Theory, which is the subject of several good textbooks on clinical decision-making (Weinstein and Fineberg, 1980; Sox et al., 2013; Hunink et al., 2014) after the pioneering work of Lesley & Lusted (1959a; 1959b; 1960; 1960; 1968) (a summary and references were given in § 3.3).

Decision-making under risk requires any assistive algorithm to work, explicitly or implicitly, in terms of probabilities, having precise connections with population statistics (§ 3.3). Without this condition the integration of patient-dependent treatment utilities would be impossible. The handling of these probabilities should moreover be enough flexible to take into account peculiar but common subpopulations of patients having special contexts or auxiliary information (§ 3.2), and the common possibility of missing values for some clinical predictors (§ 3.1). Most, if not all, popular machine-learning algorithms either do not meet these requirements, or they do so at the

Table 6: Summary of the clinician's patient-dependent inputs and the optimal predictor machine's outputs. Input data and final results that distinguish Ariel, Bianca, Curtis from Olivia are in red.

	Olivia	Ariel	Bianca	Curtis
		Clinician's patient-dependent inputs		
Predictor values				
`Age`	75.4	75.4	75.4	63.8
`Sex`	F	F	F	M
`HV`/10^{-3}	4.26	4.26	4.26	[missing]
`APOE4`	N	N	N	Y
`ANART`	18	18	18	15
`CFT`	21	21	21	14
`GDS`	3	3	3	2
`RAVLT-imm`	36	36	36	20
`RAVLT-del`	5	5	5	0
`RAVLT-rec`	10	10	10	3
`TMTA`	21	21	21	36
`TMTB`	114	114	114	126
Additional information				
auxiliary info	none	family history, base rate	none	none
applicable dataset subpopulation	all	predictor | predictand	all	all
prior probability of conversion	0.463	0.65	0.463	0.463
Available treatments and utilities				
treatment α	$\begin{matrix}\texttt{cAD}\\\texttt{N}\\10\\9\\8\\0\end{matrix}\ \begin{matrix}\\\texttt{Y}\\0\\3\\5\\10\end{matrix}$	$\begin{matrix}\texttt{cAD}\\\texttt{N}\\10\\9\\8\\0\end{matrix}\ \begin{matrix}\\\texttt{Y}\\0\\3\\5\\10\end{matrix}$	$\begin{matrix}\texttt{cAD}\\\texttt{N}\\10\\8\\7\\0\end{matrix}\ \begin{matrix}\\\texttt{Y}\\0\\3\\5\\10\end{matrix}$	$\begin{matrix}\texttt{cAD}\\\texttt{N}\\10\\9\\8\\0\end{matrix}\ \begin{matrix}\\\texttt{Y}\\0\\3\\5\\10\end{matrix}$
treatment β				
treatment γ				
treatment δ				
		Outputs of optimal predictor machine		
$\mathrm{p}(\texttt{cAD}=\texttt{Y} \mid \texttt{predictors}, \texttt{dataset})$	0.302	0.302	0.302	0.703
$\mathrm{p}(\texttt{predictors} \mid \texttt{cAD}=\texttt{Y}, \texttt{dataset})/10^{-12}$	8.97	8.97	8.97	1.14
$\mathrm{p}(\texttt{predictors} \mid \texttt{cAD}=\texttt{N}, \texttt{dataset})/10^{-12}$	18.6	18.6	18.6	0.343
final probability of conversion $\mathrm{p}(\texttt{cAD}=\texttt{Y}\|\texttt{predictors}, \texttt{dataset}, \texttt{aux info})$	0.302	0.47	0.302	0.703
exp. utility treatment α	6.98	5.27	**6.98**	2.97
exp. utility treatment β	**7.19**	6.16	6.49	4.78
exp. utility treatment γ	7.09	**6.58**	6.40	5.89
exp. utility treatment δ	3.02	4.73	3.02	**7.03**
Optimal treatment	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$	$\boldsymbol{\alpha}$	$\boldsymbol{\delta}$

cost of unrealistic modelling assumptions. Unfortunately they tend to overly simplify the problem of decision-making under risk, as if it were a simple classification or regression task.

We presented an assistive algorithm, the "optimal predictor machine", that meets all these requirements (§ 2) and carries out the calculations required by decision theory. This algorithm is moreover model-free, not making a-priori assumptions about functional dependencies or particular distributions in the variates. The inference principles on which it is based have recently been recommended for the study of Alzheimer's Disease (Temp et al., 2021; see also ASA, 2016, 2019), and have been successfully demonstrated in a simpler predictor setting (Antoniano-Villalobos et al., 2014). We showed its application in an example of prognosis and treatment of conversion from Mild Cognitive Impairment to Alzheimer's Disease for four different patients, where all three categories of differences listed above appeared. The patients were fictitious but the underlying learning database, originating from ADNI, is real, and was explored in a previous work (Rye et al., 2022).

The optimal predictor machine was also shown to have uses that go beyond individual clinical decision-making but are still of importance to personalized medicine. For instance, it can assess the maximum possible prognostic power of particular sets of predictors, potentially allowing us to discard clinical predictors that are too invasive or expensive and yet prognostically unimportant.

In actual deployment, we would recommend the hospital, medical centre, or clinician using a optimal predictor machine to keep a database of incoming patients, with their predictor values, adding the true values of their predictand later in time, once they become known. The optimal predictor machine can then be retrained on such local database when the latter reaches a size comparable to the original one's, and periodically retrained afterwards. All inferences would thus become increasingly more reliable, because the machine would base them on updated population statistics that are characteristic to the specific hospital.

4.1. Counters to possible critiques

Any inference or decision-making algorithm aspiring to take into account patient differences must perforce have some open "input slots" for such differences. We saw that the optimal predictor machine requires inputs about a patient's specific predictors, relevant statistical relations and auxiliary data, and treatment utilities.

The most difficult input to quantify is probably the third: translating benefits and drawbacks of different treatments into numbers. On this complex topic we refer the reader to specially dedicated textbooks on clinical decision making, for example Sox et al.'s (2013) and Hunink et al.'s (2014).

But some readers may wonder: "can all these additional inputs be avoided?", fearing that errors could sneak in through them.

This question is answered by a mathematical theorem at the very core of decision theory [21], which is too seldom emphasized: Any decision we make, either (A) comes explicitly or implicitly through some set of utilities and maximization of their expectations, or (B) is logically inconsistent. There is no third alternative. Thus the choice is not between using utilities or not using utilities, but between choosing them explicitly or letting them be chosen in a way we do not know. If we use a decision-making algorithm that does not ask us for utilities, then the algorithm is internally supplying utilities not chosen by us (and probably divorced from our specific problem), or, worse, is committing logical inconsistencies.

The first advantage of explicitly operating through utilities, probabilities, decision theory, is that we are, at the very least, sure of not acting in a self-contradictory way. The second advantage is that the utilities used to arrive at a decision appear openly in front of us. We can analyse and change them if we find them inappropriate to a specific problem. If they are hidden, it is more difficult to analyse which are inappropriate and how they should be changed.

The fact that an algorithm works according to decision theory is also an assurance of striving towards theoretical optimality. This point has very subtle consequences. Consider a non-optimal algorithm that leads to saving 85 000 patients out of 100 000. Given these numbers it might be deemed a success. But what if a theoretically optimal algorithm leading to 95 000 saved patients is feasible? What shall we say to the families of the 10 000 patients who could have been saved but weren't?

The most difficult input to quantify is probably the third: translating benefits and drawbacks of different treatments into numbers. On this complex topic we refer the reader to specially dedicated textbooks on clinical decision making, for example Sox et al.'s (2013) and Hunink et al.'s (2014).

4.2. Range of application of optimal predictor machines

First let us emphasize, even if it is obvious, that the quality of the results obtained with the optimal predictor machine depends on the quality of the learning dataset. Any peculiar sampling biases (or numerical errors) in the dataset that are unknown to the clinician will affect the final results. This is of course true for any inference algorithm. But we saw that the optimal predictor machine allows the clinician to correct for particular sampling biases present in the dataset, if they are known.

The range of application of the optimal predictor machine has two kinds of bounds: computational and theoretical.

[21] Savage (1972); Luce and Raiffa (1957); Raiffa and Schlaifer (2000); Atkinson et al. (1964); Ferguson (1967); Lindley (1988, 1977); Kreps (1988); Bernardo and Smith (2000); Pratt et al. (1996); Lindley (2014); Pettigrew (2019)

The fact that the optimal predictor machine extracts all available information from the dataset makes it computationally expensive (see § 2). At present it cannot be used with high-dimensional predictors: if our dataset had included a predictor such as a $128 \times 128 \times 128$ greyscale MRI image, the learning stage would have taken around 100 years. Approximate but much faster algorithms such as neural networks and random forests are thus, at present, still the only options with such predictors. There is, however, the interesting possibility of combining these fast algorithms together with a optimal predictor machine, as a post-processor of their raw output. The machine extracts useful information usually hidden in their output at a low computational cost (Dyrland et al., 2022b); this information can then be used for clinical decision-making as illustrated in the present work.

The sole assumption underlying the optimal predictor machine's inference and its practical use with new patients, is that the latter can be assumed to come, at least in some respects, from the same population as the learning dataset (in probability-theory jargon, partial or conditional exchangeability applies; see § 3.2). This precludes using the optimal predictor machine to forecast how the statistics of the full population could change in the future. However, the machine can be used for time-dependent (longitudinal) inferences within a stable population, such as forecasts of the future time of disease onset, expected lifelength, and similar. For instance, if data about the time of conversion to Alzheimer's Disease were available in the dataset, the optimal predictor machine could forecast not only *whether*, but also *when* the conversion could take place (cf. e.g. De la Cruz-Mesía et al., 2007).

Finally, the machine is not meant to handle sequences of decisions in a clinical decision tree (Sox et al., 2013, ch. 6; Hunink et al., 2014, ch. 1) – it would be impossible in a personalized approach, because such a tree is fully patient-dependent – but it could be used in individual decision branches.

A. Further mathematical and computational details about the example application of the optimal predictor machine

The optimal predictor machine surveys the space of possible distributions of frequencies of all 13 variates discussed in § 3.0, for the full population of patients from which the dataset originates (see § 3.2). In the present study, it does so by mathematically representing a generic joint frequency distribution $F(Y, X_1, X_2, \dots)$ as a convex mixture of appropriate kernels products:

$$F(Y, X_1, X_2, \dots) = \sum_i w^i \, K(Y \mid \boldsymbol{\upsilon}^i) \, K(X_1 \mid \boldsymbol{\xi}_1^i) \, K(X_2 \mid \boldsymbol{\xi}_2^i) \cdots \ ,$$

along the ideas in Dunson and Bhattacharya (2011) and Ishwaran and Zarepour (2002); see also Rossi (2014); Rasmussen (1999). This representation uses a total of 1535 independent parameters $(w^i, \boldsymbol{\upsilon}^i, \boldsymbol{\xi}_1^i, \dots)$, with roughly 190 parameters for each continuous or integer variate. As a crude

intuition, it is as if we divided the range of each variate into 190 bins, and considered all possible frequency histograms over these. The actual parametrization is smarter, incrementally using parameters to represent less and less smooth traits of the distribution. We indeed expect the distribution for a full population to have some degree of smoothness, owing to physical and biological reasons. Actually, the number of parameters used is in principle infinite, because the machine gives a warning if the data indicate that more parameters are needed. In the present study, the data indicate, on the contrary, that fewer than 250 parameters would be enough. Note that the machine constructs the kernels $K(\,|\,)$ and their product automatically, depending on how many and what kinds of variates the dataset comprises.

The probability of a candidate frequency distribution F is determined by its "fit" $F(D)$ of the data D and a prior-expectation factor $\mathrm{p}(F)$, as explained in § 2.2:

$$\mathrm{p}(F \mid D) \propto F(D)\,\mathrm{p}(F)\ .$$

Finally, the predictive conditional probability for any two sets of variates Z', Z'', given the dataset, is given by the expectation over the unknown F, as required by the probability calculus and de Finetti's theorem (Bernardo and Smith, 2000, § 4.6):

$$\mathrm{p}(Z' \mid Z'', D) = \int F(Z' \mid Z'')\,\mathrm{p}(F \mid D)\,\mathrm{d}F$$

where the conditional frequencies $F(Z' \mid Z'') \coloneqq F(Z', Z'')/F(Z'')$.

The frequency-space survey and the calculation of the probabilities $\mathrm{p}(F \mid D)$ for the population-frequency distributions was done via Gibbs sampling (Neal, 1993, ch. 4; MacKay, 2005, § 29.5; Casella and George, 1992) with the R package Nimble (de Valpine et al., 2021), using 1024 independent Markov chains. Stationarity was assessed by common diagnostic measures (Gilks et al., 1998), especially integrated autocorrelation time (Christen and Fox, 2010) and Hellinger distance (Boone et al., 2014), as well as visual inspection. An automated method for stationarity check was developed, to be discussed in future publications.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

The authors were too immersed in the development of the present work to keep a detailed record of who did what.

Funding

The study was supported by grants from the Trond Mohn Research Foundation, grant number BFS2018TMT0, and from The Research Council of Norway, project number 294594.

Acknowledgements

PGLPM thanks Soledad Gonzalo Cogno and Iván Davidovich for inspiring discussions; Maja, Mari, Miri, Emma for continuous encouragement and affection; Buster Keaton and Saitama for filling life with awe and inspiration; and the developers and maintainers of Nimble, LaTeX, Emacs, AUCTeX, Open Science Framework, R, Inkscape, LibreOffice, Sci-Hub for making a free and impartial scientific exchange possible.

Computations underlying the optimal predictor machine were initially performed on resources provided by Sigma2 – the National Infrastructure for High Performance Computing and Data Storage in Norway (project NN8050K).

Clinical-data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is 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.

Software-availability Statement

The R scripts used for this study can be found in the Open Science Framework project DOI: 10.17605/osf.io/zb26t (see also the repository https://github.com/pglpm/bayes_nonparametr

`ic_inference`). We hope to assemble them into an R package soon.

References

Alvarez-Melis, D. and Broderick, T. (2015). A translation of "The characteristic function of a random phenomenon" by Bruno de Finetti. arXiv DOI:10.48550/arXiv.1512.01229. Transl. of de Finetti (1929)

Antoniano-Villalobos, I., Wade, S., and Walker, S. G. (2014). A Bayesian nonparametric regression model with normalized weights: A study of hippocampal atrophy in Alzheimer's disease. *J. Am. Stat. Assoc.* 109, 477–490. DOI:10.1080/01621459.2013.879061

Arnborg, S. and Sjödin, G. (2001). On the foundations of Bayesianism. *Am. Inst. Phys. Conf. Proc.* 568, 61–71. DOI:10.1063/1.1381871, http://www.stats.org.uk/bayesian/ArnborgSjodin2001.pdf

ASA (2016). ASA statement on statistical significance and p-values. *Am. Stat.* 70, 131–133. Ed. by R. L. Wasserstein. DOI:10.1080/00031305.2016.1154108. See also introductory editorial in Wasserstein and Lazar (2016) and discussion in Greenland et al. (2016)

ASA (2019). Moving to a world beyond "$p < 0.05$". *Am. Stat.* 73, 1–19. Ed. by R. L. Wasserstein, A. L. Schirm, N. A. Lazar. DOI:10.1080/00031305.2019.1583913

Atkinson, F. V., Church, J. D., and Harris, B. (1964). Decision procedures for finite decision problems under complete ignorance. *Ann. Math. Stat.* 35, 1644–1655

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychol.* 44, 211–233. DOI:10.1016/0001-6918(80)90046-3

Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., et al. (eds.) (2011). *Bayesian Statistics 9* (Oxford: Oxford University Press). DOI:10.1093/acprof:oso/9780199694587.001.0001

Bernardo, J.-M. and Smith, A. F. (2000). *Bayesian Theory* (New York: Wiley), repr. edn. DOI:10.1002/9780470316870. First publ. 1994

Bhattacharya, A. and Dunson, D. B. (2010). Nonparametric Bayesian density estimation on manifolds with applications to planar shapes. *Biometrika* 97, 851–865. DOI:10.1093/biomet/asq044

Boone, E. L., Merrick, J. R., and Krachey, M. J. (2014). A Hellinger distance approach to MCMC diagnostics. *J. Statist. Comput. Simul.* 8, 833–849

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *Am. Stat.* 46, 167–174. DOI: 10.1080/00031305.1992.10475878

Cherry, C. (ed.) (1961). *Information Theory* (London: Butterworths)

Christen, J. A. and Fox, C. (2010). A general purpose sampling algorithm for continuous distributions (the t-walk). *Bayesian Anal.* 5, 263–281

Clayton, A. and Waddington, T. (2017). Bridging the intuition gap in Cox's theorem: A Jaynesian argument for universality. *Int. J. Approximate Reasoning* 80, 36–51. DOI:10.1016/j.ijar.2016.08.002

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory* (Hoboken, USA: Wiley), 2 edn. DOI:10.1002/0471200611. First publ. 1991

Cox, R. T. (1946). Probability, frequency, and reasonable expectation. *Am. J. Phys.* 14, 1–13. DOI: 10.1119/1.1990764

Cox, R. T. (1961). *The Algebra of Probable Inference* (Baltimore: The Johns Hopkins Press)

Damien, P., Dellaportas, P., Polson, N. G., and Stephens, D. A. (eds.) (2013). *Bayesian Theory and Applications* (Oxford: Oxford University Press). DOI:10.1093/acprof:oso/9780199695607.001.0001

Dawid, A. P. (2013). Exchangeability and its ramifications. In Damien et al. (2013), chap. ch. 2. 19–29. DOI:10.1093/acprof:oso/9780199695607.003.0002

de Finetti, B. (1929). Funzione caratteristica di un fenomeno aleatorio. In *Atti del Congresso Internazionale dei Matematici:*, ed. S. Pincherle (Bologna: Zanichelli), vol. 6. 179–190. https://www.mathunion.org/icm/proceedings, http://www.brunodefinetti.it/Opere.htm. Transl. in Alvarez-Melis and Broderick (2015). See also de Finetti (1930)

de Finetti, B. (1930). Funzione caratteristica di un fenomeno aleatorio. *Atti Accad. Lincei: Sc. Fis. Mat. Nat.* IV, 86–133. http://www.brunodefinetti.it/Opere.htm. Summary in de Finetti (1929)

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. Henri Poincaré* 7, 1–68. http://www.numdam.org/item/AIHP_1937__7_1_1_0. Transl. in Kyburg and Smokler (1980), pp. 53–118, by Henry E. Kyburg, Jr.

De la Cruz-Mesía, R., Quintana, F. A., and Müller, P. (2007). Semiparametric Bayesian classification with longitudinal markers. *J. R. Stat. Soc. C* 56, 119–137. DOI:10.1111/j.1467-9876.2007.00569.x

de Valpine, P., Paciorek, C., Turek, D., Michaud, N., Anderson-Bergman, C., Obermeyer, F., et al. (2021). NIMBLE: MCMC, particle filtering, and programmable hierarchical modeling. https://cran.r-project.org/package=nimble, DOI:10.5281/zenodo.1211190, https://r-nimble.org. First publ. 2016

Del Pozzo, W., Berry, C. P. L., Ghosh, A., Haines, T. S. F., Singer, L. P., and Vecchio, A. (2018). Dirichlet process Gaussian-mixture model: An application to localizing coalescing binary neutron stars with gravitational-wave observations. *Mon. Notices Royal Astron. Soc.* 479, 601–614. DOI:10.1093/mnras/sty1485

Drummond, C. and Holte, R. C. (2005). Severe class imbalance: Why better algorithms aren't the answer. *Eur. Conf. Mach. Learn.* 2005, 539–546. DOI:10.1007/11564096_52, https://webdocs.cs.ualberta.ca/~holte/Publications

Dunson, D. B. and Bhattacharya, A. (2011). Nonparametric Bayes regression and classification through mixtures of product kernels. In Bernardo et al. (2011). 145–158. DOI:10.1093/acprof:oso/9780199694587.003.0005, older version at https://www.researchgate.net/publication/228447342_Nonparametric_Bayes_Regression_and_Classification_Through_Mixtures_of_Product_Kernels

Dyrland, K., Lundervold, A. S., and Porta Mana, P. G. L. (2022a). Does the evaluation stand up to evaluation?: A first-principle approach to the evaluation of classifiers. Open Science Framework DOI:10.31219/osf.io/7rz8t

Dyrland, K., Lundervold, A. S., and Porta Mana, P. G. L. (2022b). A probability transducer and decision-theoretic augmentation for machine-learning classifiers. Open Science Framework DOI:10.31219/osf.io/vct9y

Edmonds, E. C., Delano-Wood, L., Clark, L. R., Jak, A. J., Nation, D. A., McDonald, C. R., et al. (2015). Susceptibility of the conventional criteria for mild cognitive impairment to false-positive diagnostic errors. *Alzheimer's Dement.* 11, 415–424. DOI:10.1016/j.jalz.2014.03.005

Edmonds, E. C., Weigand, A. J., Hatton, S. N., Marshall, A. J., Thomas, K. R., Ayala, D. A., et al. (2020). Patterns of longitudinal cortical atrophy over 3 years in empirically derived MCI subtypes. *Neurology* 9, e2532–e2544. DOI:10.1212/WNL.0000000000009462

Event Horizon Telescope Collaboration (2019). First M87 Event Horizon Telescope results. I. The shadow of the supermassive black hole. II. Array and instrumentation. III. Data processing and calibration. IV. Imaging the central supermassive black hole. V. Physical origin of the asymmetric ring. VI. The shadow and mass of the central black hole. *Astrophys. J. Lett.* 875, L1–L6. DOI:10.3847/2041-8213/ab0ec7, DOI:10.3847/2041-8213/ab0c96, DOI:10.3847/2041-8213/ab0c57, DOI:10.3847/2041-8213/ab0e85, DOI:10.3847/2041-8213/ab0f43, DOI:10.3847/2041-8213/ab1141

Event Horizon Telescope Collaboration (2022). First Sagittarius A* Event Horizon Telescope results. I. The shadow of the supermassive black hole in the center of the Milky Way. II EHT and multiwavelength observations, data processing, and calibration. III. Imaging of the galactic center supermassive black hole. IV. Variability, morphology, and black hole mass. V. Testing astrophysical models of the galactic center black hole. VI. Testing the black hole metric. *Astrophys. J. Lett.* 930, L12–L17. DOI:10.3847/2041-8213/ac6674, DOI:10.3847/2041-8213/ac6675, DOI:10.3847/2041-8213/ac6429, DOI:10.3847/2041-8213/ac6736, DOI:10.3847/2041-8213/ac6672, DOI:10.3847/2041-8213/ac6756

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach* (New York: Academic Press). DOI:10.1016/C2013-0-07705-5

Fine, T. L. (1973). *Theories of Probability: An Examination of Foundations* (New York: Academic Press). DOI:10.1016/C2013-0-10655-1

Fong, E. and Holmes, C. C. (2020). On the marginal likelihood and cross-validation. *Biometrika* 107, 489–496. DOI:10.1093/biomet/asz077

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (eds.) (1998). *Markov Chain Monte Carlo in Practice* (Boca Raton, USA: Chapman & Hall/CRC), repr. edn. DOI:10.1201/b14835. First publ. 1996

Good, I. J. (1950). *Probability and the Weighing of Evidence* (London: Griffin)

Good, I. J. (1961). Weight of evidence, causality and false-alarm probabilities. In Cherry (1961), chap. 11. 125–136. With Discussion by A. J. Mayne and D. M. MacKay and reply

Good, I. J. and Toulmin, G. H. (1968). Coding theorems and weight of evidence. *IMA J. Appl. Math.* 4, 94–105. DOI:10.1093/imamat/4.1.94

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., et al. (2016). Online supplement and discussion: ASA statement on statistical significance and p-values. *Am.*

Stat. 70, 129. DOI:10.1080/00031305.2016.1154108 supplemental material. See ASA (2016) and Wasserstein and Lazar (2016)

Hailperin, T. (1996). *Sentential Probability Logic: Origins, Development, Current Status, and Technical Applications* (London: Associated University Presses)

Halpern, J. Y. (1999). Cox's theorem revisited. *J. Artif. Intell. Res.* 11, 429–435. DOI:10.1613/jair.644. See also Snow (1998)

Harper, W. L. and Hooker, C. A. (eds.) (1976). *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science. Vol. II: Foundations and Philosophy of Statistical Inference* (Dordrecht: Reidel)

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (eds.) (2010). *Bayesian Nonparametrics* (Cambridge: Cambridge University Press). DOI:10.1017/CBO9780511802478

Hunink, M. G. M., Weinstein, M. C., Wittenberg, E., Drummond, M. F., Pliskin, J. S., Wong, J. B., et al. (2014). *Decision Making in Health and Medicine: Integrating Evidence and Values* (Cambridge: Cambridge University Press), 2 edn. DOI:10.1017/CBO9781139506779. First publ. 2001

Ishwaran, H. and Zarepour, M. (2002). Dirichlet prior sieves in finite normal mixtures. *Stat. Sinica* 12, 941–963. http://www3.stat.sinica.edu.tw/statistica/J12n3/j12n316/j12n316.htm

ISO (2008). *ISO 80000-13:2008: Quantities and units 13: Information science and technology.* International Organization for Standardization, Geneva

Jawa, N. A. and Maslove, D. M. (2023). Bayes' theorem in neurocritical care: Principles and practice. *Neurocrit. Care* 2023, 1. DOI:10.1007/s12028-022-01665-2. Note that some statements in this paper are incorrect or misleading. Table 5 states as advantage of frequentist statistics "not susceptible to prior beliefs"; this is not true. Frequentist statistics *does* use a prior, but keeps it hidden: see e.g. Pratt (1961) p. 167, Savage et al. (1962) p. 49 item (ii); so in frequentist statistics it becomes difficult to check whether this hidden prior is appropriate to the given problem. On the other hand, if the frequentist statistics did not implicitly use a prior, it would lead to contradictions – and in some cases indeed it does – see references above and Lindley (1977). Table 5 also states as disadvantage of Bayesian statistics "Priors are subjective and may be biased". This statement is unfair: a prior is no less subjective than choosing "0.05" as a threshold of "statistical significance": why this particular value? what is its rationale? If one answers that it is a value generally agreed upon by the community, then such general community agreement exists also for priors; thus

"subjective" is untrue. But the truth is that there are debates about the appropriate "significance" level, which is therefore as "subjective" as a prior.

Jaynes, E. T. (1976). Confidence intervals vs Bayesian intervals. In Harper and Hooker (1976). 175–257. With discussion, comments by M. Maxfield and O. Kempthorne, and reply. Repr. with an introduction in Jaynes (1989, 149–209); http://bayes.wustl.edu/etj/node1.html

Jaynes, E. T. (1989). *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics* (Dordrecht: Kluwer), repr. edn. Edited by R. D. Rosenkrantz. First publ. 1983

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science* (Cambridge: Cambridge University Press). Ed. by G. Larry Bretthorst. First publ. 1994. DOI:10.1017/CBO9780511790423, https://archive.org/details/XQUHIUXHIQUHIQXUIHX2, http://www-biba.inrialpes.fr/Jaynes/prob.html

JCGM (2008). *JCGM 100:2008: Evaluation of measurement data – Guide to the Expression of Uncertainty in Measurement*. Joint Committee for Guides in Metrology (JCGM): BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, OIML, corr. version edn. http://www.bipm.org/en/publications/guides/gum.html. Includes various supplements. First publ. 1993

JCGM (2012). *JCGM 200:2012: International vocabulary of metrology – Basic and general concepts and associated terms (VIM)*. Joint Committee for Guides in Metrology (JCGM): BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, OIML, 3 edn. https://www.bipm.org/en/publications/guides/vim.html. First publ. 1997

Jenny, M. A., Keller, N., and Gigerenzer, G. (2018). Assessing minimal medical statistical literacy using the Quick Risk Test: a prospective observational study in Germany. *BMJ Open* 8, e020847, e020847corr2. DOI:10.1136/bmjopen-2017-020847, DOI:10.1136/bmjopen-2017-020847corr2

Johnson, W. E. (1924). *Logic. Part III: The Logical Foundations of Science* (Cambridge: Cambridge University Press). https://archive.org/details/logic03john

Johnson, W. E. (1932). Probability: The deductive and inductive problems. *Mind* 41, 409–423. With some notes and an appendix by R. B. Braithwaite. DOI:10.1093/mind/XLI.164.409

Kelly, J. L., Jr. (1956). A new interpretation of information rate. *Bell Syst. Tech. J.* 35, 917–926. http://turtletrader.com/kelly.pdf, https://archive.org/details/bstj35-4-917

Kreps, D. (1988). *Notes On The Theory Of Choice* (New York: Routledge). DOI:10.4324/9780429498619

Kullback, S. (1978). *Information Theory and Statistics* (New York: Dover). Republ. with a new preface and corrections and additions by the author. First publ. 1959

Kyburg, H. E., Jr. and Smokler, H. E. (eds.) (1980). *Studies in Subjective Probability* (Huntington, USA: Robert E. Krieger), 2 edn. First publ. 1964

Ledley, R. S. (1959). Digital electronic computers in biomedical science: Computers make solutions to complex biomedical problems feasible, but obstacles curb widespread use. *Science* 130, 1225–1234. DOI:10.1126/science.130.3384.1225

Ledley, R. S. (1960). *Digital Computer and Control Engineering* (New York: McGraw-Hill). Written with the assistance of Louis S. Rotolo and James Bruce Wilson. https://archive.org/details/bitsavers_columbiaUnuterandControlEngineering1960_40752710

Ledley, R. S. and Lusted, L. B. (1959a). Reasoning foundations of medical diagnosis: Symbolic logic, probability, and value theory aid our understanding of how physicians reason. *Science* 130, 9—21. DOI:10.1126/science.130.3366.9

Ledley, R. S. and Lusted, L. B. (1959b). The use of electronic computers to aid in medical diagnosis. *Proc. IRE* 47, 1970–1977. DOI:10.1109/JRPROC.1959.287213

Ledley, R. S. and Lusted, L. B. (1960). Computers in medical data processing. *Oper. Res.* 8, 299–310. DOI:10.1287/opre.8.3.299

Lindley, D. V. (1977). The distinction between inference and decision. *Synthese* 36, 51–58. DOI:10.1007/BF00485691

Lindley, D. V. (1988). *Making Decisions* (London: Wiley), 2 edn. First publ. 1971

Lindley, D. V. (2014). *Understanding Uncertainty* (Hoboken, USA: Wiley), rev. ed. edn. First publ. 2006

Lindley, D. V. and Novick, M. R. (1981). The role of exchangeability in inference. *Ann. Stat.* 9, 45–58. DOI:10.1214/aos/1176345331

Liu, C.-C., Kanekiyo, T., Xu, H., and Bu, G. (2013). Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nat. Rev. Neurol.* 9, 106–118, 184. DOI:10.1038/nrneurol.2012.263, DOI:10.1038/nrneurol.2013.32

Luce, R. D. and Raiffa, H. (1957). *Games and Decisions: introduction and critical survey* (New York: Wiley)

Lusted, L. B. (1967). Logical analysis in medical diagnosis. *Berkeley Symp. Math. Stat. Probab.* 5/IV, 903–923. https://projecteuclid.org/proceedings/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/Chapter/Logical-analysis-in-medical-diagnosis/bsmsp/1200513835

Lusted, L. B. (1968). *Introduction to Medical Decision Making* (Springfield, USA: Thomas)

Lusted, L. B. and Ledley, R. S. (1960). Mathematical models in medical diagnosis. *J. Med. Educ.* 35, 214–222. https://journals.lww.com/academicmedicine/Citation/1960/03000/Mathematical_Models_in_Medical_Diagnosis.2.aspx

MacKay, D. J. C. (1992a). Bayesian interpolation. *Neural Comput.* 4, 415–447. http://www.inference.phy.cam.ac.uk/mackay/PhD.html, DOI:10.1162/neco.1992.4.3.415

MacKay, D. J. C. (1992b). A practical Bayesian framework for backpropagation networks. *Neural Comput.* 4, 448–472. http://www.inference.phy.cam.ac.uk/mackay/PhD.html, DOI:10.1162/neco.1992.4.3.448

MacKay, D. J. C. (2005). *Information Theory, Inference, and Learning Algorithms* (Cambridge: Cambridge University Press), version 7.2 (4th pr.) edn. https://www.inference.org.uk/itila/book.html. First publ. 1995

Malinas, G. and Bigelow, J. (2016). Simpson's paradox. In *Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta (Stanford: The Metaphysics Research Lab). https://plato.stanford.edu/archives/fall2016/entries/paradox-simpson. First publ. 2004

Matthews, R. A. J. (1996). Base-rate errors and rain forecasts. *Nature* 382, 766. DOI:10.1038/382766a0

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's disease. *Neurology* 34, 939–944. DOI:10.1212/WNL.34.7.939

Minka, T. P. (2003). *Bayesian inference, entropy, and the multinomial distribution.* Tech. rep., MIT media Lab, Cambridge, USA. https://tminka.github.io/papers/multinomial.html. First publ. 1998

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective* (Cambridge, USA: MIT Press). https://probml.github.io/pml-book/book0.html

Neal, R. M. (1993). *Probabilistic Inference using Markov Chain Monte Carlo Methods*. Tech. Rep. CRG-TR-93-1, University of Toronto, Toronto. http://www.cs.utoronto.ca/~radford/review.abstract.html, https://omega0.xyz/omega8008/neal.pdf

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., et al. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities* (Chichester: Wiley). DOI:10.1002/0470033312

Paris, J. B. (2006). *The Uncertain Reasoner's Companion: A Mathematical Perspective* (Cambridge: Cambridge University Press), repr. edn. DOI:10.1017/CBO9780511526596. See also Snow (1998)

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., et al. (2010). Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization. *Neurology* 74, 201–209. DOI:10.1212/WNL.0b013e3181cb3e25

Pettigrew, R. (2019). Epistemic utility arguments for probabilism. In *Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta (Stanford: The Metaphysics Research Lab). https://plato.stanford.edu/archives/win2019/entries/epistemic-utility. First publ. 2011

Pólya, G. (1954). *Mathematics and Plausible Reasoning: Vol. I: Induction and Analogy in Mathematics* (Princeton: Princeton University Press). https://archive.org/details/Induction_And_Analogy_In_Mathematics_1_, DOI:10.1515/9780691218304

Pólya, G. (1968). *Mathematics and Plausible Reasoning: Vol. II: Patterns of Plausible Inference* (Princeton: Princeton University Press), 2 edn. First publ. 1954

Porta Mana, P. G. L. (2019). A relation between log-likelihood and cross-validation log-scores. Open Science Framework DOI:10.31219/osf.io/k8mj3, HAL:hal-02267943, arXiv DOI:10.48550/arXiv.1908.08741

Pratt, J. W. (1961). Book review: *Testing Statistical Hypotheses*, E. L. Lehmann. *J. Am. Stat. Assoc.* 56, 163–167

Pratt, J. W., Raiffa, H., and Schlaifer, R. (1996). *Introduction to Statistical Decision Theory* (Cambridge, USA: MIT Press), 2nd pr. edn. First publ. 1995

Provost, F. (2000). *Machine Learning from Imbalanced Data Sets 101*. Tech. Rep. WS-00-05-001, AAAI, Menlo Park, USA. https://aaai.org/Library/Workshops/2000/ws00-05-001.php

Quintana, M., Viele, K., and Lewis, R. J. (2017). Bayesian analysis: Using prior information to interpret the results of clinical trials. *J. Am. Med. Assoc.* 318, 1605–1606. DOI:10.1001/jama.2017.15574

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org. First released 1995

Raiffa, H. (1970). *Decision Analysis: Introductory Lectures on Choices under Uncertainty* (Reading, USA: Addison-Wesley), 2nd pr. edn. First publ. 1968

Raiffa, H. and Schlaifer, R. (2000). *Applied Statistical Decision Theory* (New York: Wiley), repr. edn. First publ. 1961

Rasmussen, C. E. (1999). The infinite Gaussian mixture model. *Adv. Neural Inf. Process. Syst. (NIPS)* 12, 554–560. https://www.seas.harvard.edu/courses/cs281/papers/rasmussen-1999a.pdf

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2009). Bayesian nonparametric functional data analysis through density estimation. *Biometrika* 96, 149–162. DOI:10.1093/biomet/asn054

Rosenkrantz, R. D. (1977). *Inference, Method and Decision: Towards a Bayesian Philosophy of Science* (Dordrecht: Reidel)

Rossi, P. E. (2014). *Bayesian Non- and Semi-parametric Methods and Applications* (Princeton: Princeton University Press). DOI:10.1515/9781400850303

Russell, S. J. and Norvig, P. (2022). *Artificial Intelligence: A Modern Approach* (Harlow, UK: Pearson), fourth global ed. edn. http://aima.cs.berkeley.edu/global-index.html, https://archive.org/details/artificial-intelligence-a-modern-approach-4th-edition. First publ. 1995

Rye, I., Vik, A., Kocinski, M., Lundervold, A. S., and Lundervold, A. J. (2022). Predicting conversion to Alzheimer's disease in individuals with Mild Cognitive Impairment using clinically transferable features. *Sci. Rep.* 12, 15566. DOI:10.1038/s41598-022-18805-5

Savage, L. J. (1972). *The Foundations of Statistics* (New York: Dover), 2nd rev. and enl. ed. edn. First publ. 1954

Savage, L. J., Bartlett, M. S., Barnard, G. A., Cox, D. R., Pearson, E. S., and Smith, C. A. B. (1962). *The Foundations of Statistical Inference: A Discussion* (London: Methuen). With a discussion including H. Ruben, I. J. Good, D. V. Lindley, P. Armitage, C. B. Winsten, R. Syski, E. D. Van Rest, G. M. Jenkins

Self, M. and Cheeseman, P. C. (1987). Bayesian prediction for artificial intelligence. In *Proceedings of the Third Conference on Uncertainty in Artificial Intelligence (UAI'87)*, eds. J. Lemmer, T. Levitt, and L. Kanal (Arlington, USA: AUAI Press). 61–69. Repr. in arXiv DOI:10.48550/arXiv.1304.2717

Shah, S., Beck, J. R., and Pauker, S. G. (2013). In memoriam: Robert Steven Ledley, DDS, MS (physics), 1926–2012. *Med. Decis. Making* 33, 731–733. DOI:10.1177/0272989X1348794

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423, 623–656. https://archive.org/details/bstj27-3-379, https://archive.org/details/bstj27-4-623, http://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf

Smith, J. E. and Winkler, R. L. (2006). The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Manag. Sci.* 52. DOI:10.1287/mnsc.1050.0451

Snow, P. (1998). On the correctness and reasonableness of Cox's theorem for finite domains. *Comput. Intell.* 14, 452–459. DOI:10.1111/0824-7935.00070

Snow, P. (2001). The reasonableness of possibility from the perspective of Cox. *Comput. Intell.* 17, 178–192. DOI:10.1111/0824-7935.00138

Sox, H. C., Higgins, M. C., and Owens, D. K. (2013). *Medical Decision Making* (New York: Wiley), 2 edn. DOI:10.1002/9781118341544. First publ. 1988

Sprenger, J. and Weinberger, N. (2021). Simpson's paradox. In *Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta (Stanford: The Metaphysics Research Lab). https://plato.stanford.edu/archives/sum2021/entries/paradox-simpson

Temp, A. G. M., Lutz, M. W., Trepel, D., Tang, Y., Wagenmakers, E.-J., Khachaturian, A. S., et al. (2021). How Bayesian statistics may help answer some of the controversial questions in clinical research on Alzheimer's disease. *Alzheimer's Dement.* 17, 917–919. DOI:10.1002/alz.12374

Tribus, M. (1969). *Rational Descriptions, Decisions and Designs* (New York: Pergamon). DOI:10.1016/C2013-0-01558-7

Van Horn, K. S. (2003). Constructing a logic of plausible inference: a guide to Cox's theorem. *Int. J. Approximate Reasoning* 34, 3–24. DOI:10.1016/S0888-613X(03)00051-3

von Neumann, J. and Morgenstern, O. (1955). *Theory of Games and Economic Behavior* (Princeton: Princeton University Press), 3rd ed., 6th pr. edn. https://archive.org/details/in.ernet.dli.2015.215284. First publ. 1944

Walker, S. G. (2010). Bayesian nonparametric methods: motivation and ideas. In Hjort et al. (2010), chap. ch. 1. 22–34

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *Am. Stat.* 70, 129–133. DOI:10.1080/00031305.2016.1154108. See ASA (2016) and discussion in Greenland et al. (2016)

Weinstein, M. C. and Fineberg, H. V. (1980). *Clinical Decision Analysis* (Philadelphia: Saunders)

Weiss, G. M. and Provost, F. (2003). Learning when training data are costly: The effect of class distribution on tree induction. *J. Artif. Intell. Res.* 19, 315–354. DOI:10.1613/jair.1199

Woodward, P. M. (1964). *Probability and Information Theory, with Applications to Radar* (Oxford: Pergamon), 2 edn. DOI:10.1016/C2013-0-05390-X. First publ. 1953