Personalized prognosis & treatment using Ledley-Jaynes machines: An example study on conversion from Mild Cognitive Impairment to Alzheimer's Disease

P.G.L. Porta Mana [1,2,*], **I. Rye** [3], **A. Vik** [2], **M. Kociński** [2,4], **A. Lundervold** [2,4], **A. J. Lundervold** [3], **A. S. Lundervold** [1,2]

[1] *Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway*

[2] *Mohn Medical Imaging and Visualization Centre (MMIV), Department of Radiology, Haukeland University Hospital, Bergen, Norway*

[3] *Department of Biological and Medical Psychology, University of Bergen, Norway*

[4] *Department of Biomedicine, University of Bergen, Norway*

Correspondence*:
P.G.L Porta Mana, HVL, Inndalsveien 28, 5063 Bergen
pgl@portamana.org

ABSTRACT

🔧 TO BE REWRITTEN

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 battery of clinical examinations. Today they received the results. Based on these results, on available clinical statistical data, and on other relevant information, their clinician will assess their risk of developing Alzheimer. Then, together with the patients and their relatives, the clinician will make a decision among four distinct preventive-treatment options.[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 clinical results 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, 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 clinical 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 losses from the treatment options. Note that one clinical result for Curtis's (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, 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*.

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 of such computational help is clear from the vast amount of clinical statistical data and the large

---

[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".

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 and background information" (Olivia and Ariel), "difference in benefit and availability of treatments" (Olivia and Bianca), "difference in clinical predictors" (Olivia and 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, 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 and 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 predict 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;

---

[3] literally "quantity to be predicted" (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", or similar.

   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 patient.

8. It can quantify the uncertainty of its own outputs and therefore allow 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 had been available.

9. It can make various kinds of long-term forecasts, such as frequency of prognoses having 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, it is clear that a simple dichotomy "has disease" vs "doesn't have disease" is an oversimplification ✎ more text and references?. Without the capability of being used with patients having peculiar contexts or clinical situations, feature 3, the algorithm would be of no use in the often occurring case of patients having special 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 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 (as determined by information theory) 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 another set of predictors; it itself attains that maximal accuracy.

Further features relevant for Machine Learning are given in §✎ ***.

We call this algorithm a *Ledley-Jaynes machine*, for reasons explained in the next section. It is at the moment available as a collection of scripts[4] in the R programming language (R Core Team, 2023), which we plan to soon assemble into a clinician-friendly R package.

---

[4] `https://github.com/pglpm/ledley-jaynes_machine`

The next section 2 gives an intuitive understanding of the Ledley-Jaynes machine's underlying principles and workings. It is largely independent of the subsequent sections, so it can be skipped by readers who want to immediately see its application. Application is shown in § 3, using the four-patient fictitious scenario of § 1.0 as a concrete example; the final subsection 3.4 discusses 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[5], which explains more in detail how to use the R scripts.

---

[5] `https://github.com/pglpm/ledley-jaynes_machine/raw/main/ledley-jaynes_machine.pdf`

2   THE LEDLEY-JAYNES MACHINE

2.1   Underlying principles 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.[6]

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

The "Ledley-Jaynes machine" is an algorithmic implementation of the main calculations underlying the clinical decision-making process, as dreamed by Ledley and Lusted: from the comparison of a patient's specific predictors with the statistics offered by a clinical database, to the choice of optimal treatment. The name is a homage to Ledley – who, incidentally, died of Alzheimer's Disease (Shah et al., 2013) – and to Jaynes (2003), who brilliantly explained the inductive logic underlying such a "robot".

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 Ledley-Jaynes 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 its informational output, characteristic of 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 applied at insubstantial computational cost.

More concretely, the Ledley-Jaynes machine is ideal because *it computes the probability distribution over all possible long-run joint frequency distributions from which the learning dataset can originate* (joint distributions for all predictor and predictand variates). 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 Ledley-Jaynes 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 $p(Y \mid X)$ of any set of predictands Y given any set of input predictors X.
- *"Generative" algorithms:* the probability $p(X \mid Y)$ 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 average value of any set of variates Y, given any other set of variates X, including the special case of predictand Y and predictors X. The uncertainty or variability around such average is also automatically computed.

---

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

- *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 Ledley-Jaynes 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 that the statistical properties of learning dataset could be different from those of its original population, owing to sampling fluctuations; and it computes 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 Ledley-Jaynes machine computes the general probability distribution over long-run frequencies; but we can already summarize a couple of special characteristics brought about by such computation. Unlike machine-learning algorithms such as neural networks, random forests, support-vector machines, logistic regression, or generalized linear models, the Ledley-Jaynes 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 another special characteristic: the Ledley-Jaynes machine 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 proved 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 Basic principle and learning from a dataset

The calculations by which the Ledley-Jaynes machine learns and operates are univocally determined by Cox's theorem [7], which determines quantitative inference rules from self-consistency requirements, and by de Finetti's theorem[8], 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 patients – past, present, future. Suppose we knew the joint frequency distribution of all possible combinations of predictor and predictand values in such population. We would then judge the probability for a patient's particular variate values to be equal to the their corresponding population frequency. Pure symmetry considerations lead to this result (de Finetti, 1930; Dawid, 2013; Bernardo and Smith, 2000, §§ 4.2–4.3). The same would be

---

[7] Cox, 1946, 1961; Pólya, 1954, 1968; Tribus, 1969; Rosenkrantz, 1977; Snow, 1998; Halpern, 1999; Arnborg and Sjödin, 2001; Snow, 2001; Clayton and Waddington, 2017; for a review see Van Horn, 2003.

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

true for conditional and marginal probabilities and frequencies.[9] 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 Ledley-Jaynes 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 joint probability of the dataset, assuming it was sampled from a population having the candidate frequency. The second factor is a prior probability expressing how reasonable the candidate is.[10] The only natural requirement is that the candidate should have some degree of smoothness, owing to physical and biological reasons. Figure 6 show samples of what the machine considers to be "reasonable candidates" for the population frequency distribution of a discrete variate. Note that no frequency distributions are excluded; they are only given higher or lower probabilities.

In the learning stage the Ledley-Jaynes machine infers, from a given sample dataset, the statistical relationships of a large population of which our future patients can be considered members, at least in some respects. Such relationships will help us in our prognoses. The basic idea is intuitive. If a patient can be considered a member of some population, and if we knew the joint frequencies of all possible combinations of predictor and predictand values in such population – and knew nothing else – then we would say that the probability for the patient to have particular values is equal to the population frequency. Pure symmetry considerations lead to this intuitive result (de Finetti, 1930; Dawid, 2013; Bernardo and Smith, 2000, §§ 4.2–4.3).

But it must be emphasized, and it is essential for our method, that it is *not* necessary (and is seldom true) that a future patient be considered as a member of such a population *in all respects*. A patient can be considered a member only *conditionally* on particular variate values. We shall discuss this point with an example in § 3.2.

If the full statistics of such a population were known, our task would just be to "enumerate" rather than to "learn". Learning comes into play because the full population is not known: we only have a sample from it.

The Ledley-Jaynes machine assigns a probability to each possible frequency distribution for the full population. It determines the probability of each "candidate" frequency distribution by combining two factors: (a) how well the candidate fits the sample data, (b) how biologically or physically reasonable the candidate is. Figure 2 show a fictitious sample data and various candidate frequency distributions ✏ ....

---

[9] 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 to 0. Therefore a functional relationship is still encompassed by this point of view as a particular case.

[10] It must be emphasized that 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? Such judgement implies that we have a preconceived, reasonable reference distribution in the back of our minds. The fit is either intuitively compared with the preconceived reference; or is compared with a known, ground-truth test for testing owing to its similarity with the preconceived reference.

✏ EITHER HERE OR IN APPENDIX A.2: Add some more intuition and details about the maths, principles, and characteristics (Dunson and Bhattacharya, 2011; Rossi, 2014; Rasmussen, 1999).

The Ledley-Jaynes machine computes the probabilities of all possible frequency distributions for the full population, and from these the joint probability distribution $p(X, Y, Z, \dots)$ for all variates $X, Y, Z, \dots$ available in the dataset.

For readers with an interest in machine learning and artificial intelligence, we briefly discuss the drawbacks of some popular inference algorithms with respect to the Ledley-Jaynes machine. 🔧 Rewrite the paragraphs below to better discuss the limitations. Maybe better to list according to limitations than to algorithms:

- *Special assumptions*, such as functional dependence (neural networks, Gaussian processes, partly random forests), or special shape of underlying distributions (neural networks, support vector machines, linear and logistic regression, generalized linear models).
- *Lack of uncertainty quantification* (neural networks), or non-probabilistic quantification (random forests for classification)
- *One-way-only inference*, from predictors to predictands only (all)
- *Blindness to possible fluctuations affecting the whole dataset* (all algorithms except gaussian processes): the algorithms try to generalize by splitting the dataset into a training set and a test set. This procedure does not protect up from fluctuations that affect the *whole* dataset. Such fluctuations are important for datasets of small size ✏ be precise here about the size. The Ledley-Jaynes machine intrinsically check how the full population could have been different from the dataset.

Figure 1. ✏ Examples probable candidates of frequency distribution for a variate such as `RAVLT-del` or `RAVLT-rec`
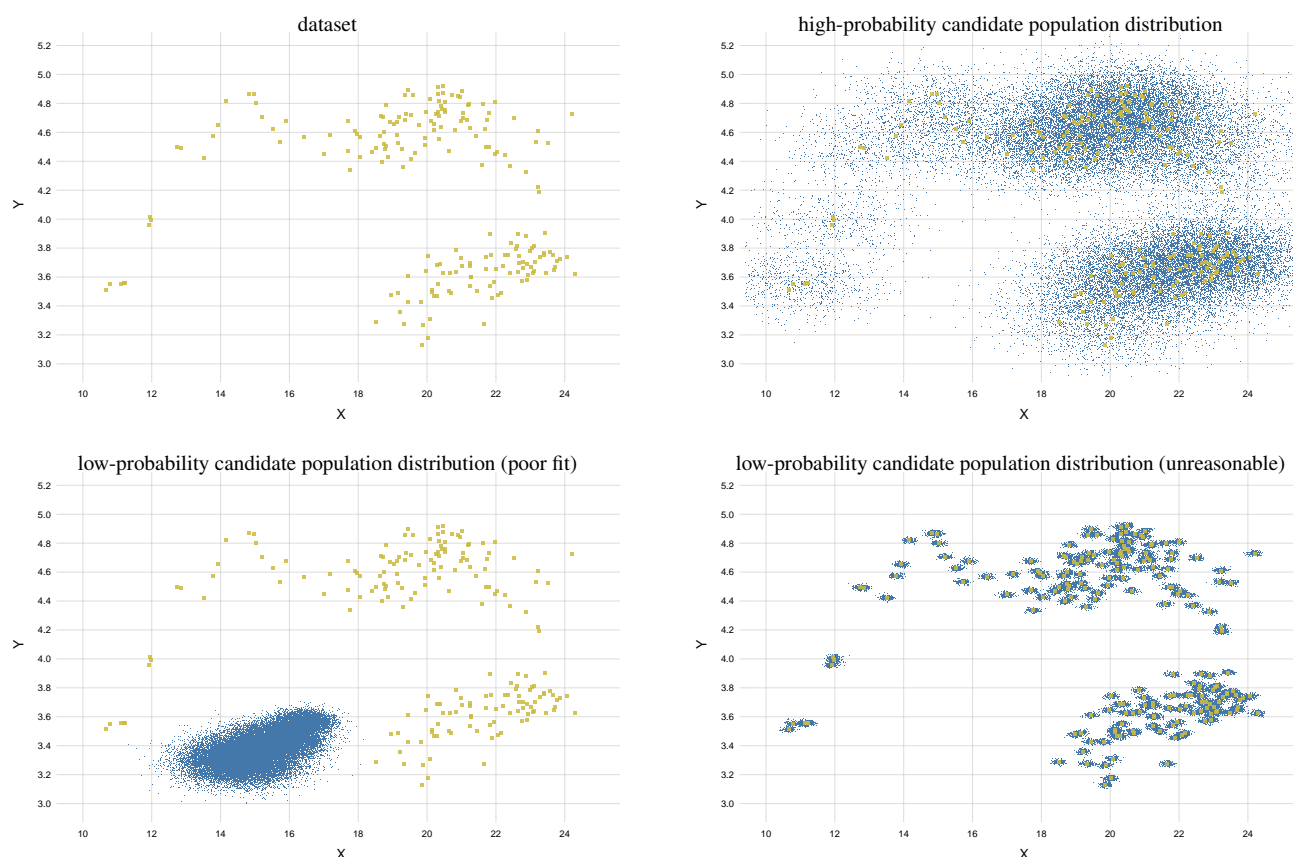
Figure 2. ✎ Upper-left: Sample data. Upper-right: candidate frequency distribution that fits the data and does not look unnatural. Lower-left: candidate distribution that might look natural but doesn't fit the sample data. Lower-right: candidate distribution that fits the data very well but looks unnatural.

Neural networks and gaussian processes are based on the assumption that there is a functional relationship from predictors to predictand, possibly contaminated by a little noise, typically assumed gaussian. This is a very strong assumption, quite unrealistic for many kinds of variables considered in medicine. It can only be justified in the presence of informationally very rich predictors such as images. In the case of our dataset, the mutual information between predictors and the conversion variate is $0.14\,\mathrm{Sh}$, to be compared with $1\,\mathrm{Sh}$, if conversion were a function of the predictors, and with $0\,\mathrm{Sh}$, if conversion were completely unpredictable. See § 3.4 for further details. An additional deficiency of neural networks is that they do not yield any probabilities, even if there are many efforts to render such an output possible (Pearce et al., 2020; Osband et al., 2021; Back and Keith, 2019). Such an advance, however, would still not solve the final deficiency of neural networks and gaussian processes: they try to infer the predictand from the predictors but cannot be used for the reverse inference.

Random forests also assume a functional relationship from predictors to predictand. This assumptions is mitigated when the predictand is a discrete variate; in this case a random forest can output an agreement score from its constituent decision trees. This score can give an idea of the underlying uncertainty, but it is not a probability,[11] and therefore cannot be used in the decision-making stage, see § 3.3. It is possible to transform this score into a proper probability (Dyrland et al., 2022b), but this possibility does not solve the

---

[11] It is sometimes called a "non-calibrated probability", something akin to a "non-round circle".

final deficiency of random forests: like neural networks, they try to infer the predictand from the predictors but cannot be used for the reverse inference.

Parametric models and machine-learning algorithms such as logistic or linear regression, support-vector machines, or generalized linear models make even stronger assumptions than neural networks and random forests. They assume specific functional shapes or frequency distributions. Their use may be justified when we are extremely sure – for instance thanks to underlying physical or biological knowledge – of the validity of their assumptions; or when the computational resources are extremely scarce. But it is otherwise unnecessary to be hampered by their restrictive and often unrealistic assumptions.

✏ Add note on "ensembling": it does not give probabilities (Murphy, 2022, § 18.2)

An important common deficiency of most inference algorithms mentioned above is that their inference only goes from predictors to predictands. In the next section we shall see that this limitation precludes – or makes much riskier – the prognostic use of the learning dataset for patients, such as Ariel, who belong to different populations.

<div style="border:1px solid black">

Inset: Inferential and decision-making steps

1. 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 Ledley-Jaynes machine and let it infer the joint full-population frequencies of predictors and predictand behind the dataset.

2. Measure the present patient's predictor values and input them in the Ledley-Jaynes machine. Partially missing values are allowed.

3. Assess which conditional statistics of the dataset can be applied to the present patient, and any auxiliary clinical information available. Quantify the latter it in a prior probability.
Input the relevant statistics and auxiliary information for the present patient into the Ledley-Jaynes 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.

4. 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 Ledley-Jaynes 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.

</div>

3 EXAMPLE APPLICATION

The inferential and decision-making steps based on the Ledley-Jaynes machine are summarized in the inset at the bottom of this page.

How does the clinician use the Ledley-Jaynes machine in a specific application to a patient, after the machine has learned from the dataset? The clinician inputs:

1. the patient's clinical data

2. the choice of statistical relation that can be generalized from dataset to the patient (see § 3.2)

3. prior probabilities coming from auxiliary information

4. set of available treatments and their utilities

Then the machine will output (in a fraction of a second) the optimal treatment for the patient. Alternatively the clinician can stop at step 2 above, and do the rest of the calculation by hand.

These steps taken by the clinician and use of the Ledley-Jaynes machine will be explained and illustrated in the next three sections. They are presented in chronological order as the clinician would apply them. In each section, a general overview and discussion of the theory and method behind the specific step is first given, followed by the concrete application to our example case.

The three steps and sections could also be presented in reverse order, more suited to their logical dependence, because the procedure in each one is actually motivated by the next. We suggest that readers familiar with the principles of clinical decision making read them in § **??**-§ 3.2-§ 3.3 order; whereas readers

---

unfamiliar with these principles (who may include readers with a specialized background in machine learning) read them in § 3.3-§ 3.2-§ **??** order.

🔧 Remark that the approach used here is the same as in the research behind the last two black-hole observations (Event Horizon Telescope Collaboration, 2019, 2022).

3.1 Patient's clinical information

Theory and method

Application to the example study

✎ Show some graphs about the full-population distribution inferred by the Ledley-Jaynes machine; for example the population distributions for some predictor, given converted/non-converted.

The predictor values for our four patients are reported in table 1. Note that Curtis's value for the Hippocampal Volume is missing; the Ledley-Jaynes machine has no difficulty with this.

Given these predictor values the Ledley-Jaynes machine yields the probabilities of main importance for the application in the next sections reported in table 2.

As an illustration of the additional information provided by the Ledley-Jaynes machine, fig. 3 shows the probability distributions for the full-population frequency of conversion to Alzheimer's Disease for patients with predictor values equal to those of Olivia, Ariel, Bianca (left), and to those of Curtis (right). The probability $p(\text{conversion to AD} \mid \text{predictors}, \text{dataset})$ is equal to the average of such a distribution, as required by probability theory (e.g. Bernardo and Smith, 2000, §§ 4.2–4.3). We emphasize that the distributions shown in the figure are *not* the uncertainties or "errors" on these probabilities. There is a numerical uncertainty over the probabilities, caused by finite computational precision, but it affects at most the third significant digit of the reported probabilities. All relative uncertainties are below 0.8%, Curtis's two likelihoods being an exception at 2% (owing to the illustrative character of this example, we do not fully follow the standards for the expression of measurement uncertainty (JCGM, 2008)).

	Olivia	Ariel	Bianca	Curtis
Age	75.4	75.4	75.4	63.8
Sex	F	F	F	M
Hippocampal volume (HV)$/10^{-3}$	4.26	4.26	4.26	[missing]
APOE4 status	N	N	N	Y
Reading test (ANART)	18	18	18	15
Category Fluency Test (CFT)	21	21	21	14
Geriatric Depression Scale (GDS)	3	3	3	2
RAVLT-immediate memory	36	36	36	20
RAVLT-delayed recall	5	5	5	0
RAVLT-recognition	10	10	10	3
Trail Making Test A (TMTA)	21	21	21	36
Trail Making Test B (TMTB)	114	114	114	126

Table 1. ✎ Predictor values for the four patients

	Olivia	Ariel	Bianca	Curtis
$p(\text{conversion to AD} \mid \text{predictors})$	0.302	0.302	0.302	0.703
$p(\text{predictors} \mid \text{conversion to AD})/10^{-12}$	8.97	8.97	8.97	1.14
$p(\text{predictors} \mid \text{no conversion to AD})/10^{-12}$	18.6	18.6	18.6	0.343

Table 2. ✎ Probabilities computed by the Ledley-Jaynes machine
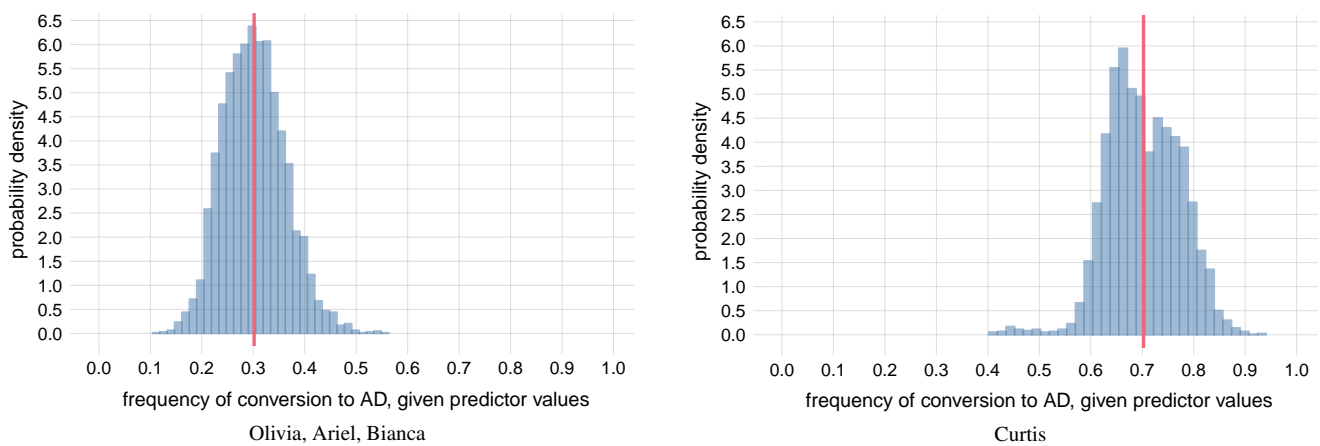
Figure 3. Probability distributions for full-population frequency of conversion to AD, given Olivia, Ariel, Bianca's and Curtis's predictor. Red lines are the values of the probabilities $p(\text{conversion to AD} \mid \text{predictors}, \text{dataset})$

🧩 Use "likelihood" instead of "predictors | predictand"? It's a possibility. I personally prefer the second notation, much more clear and explicit.

3.2   Assessment of population and auxiliary information

As already mentioned and as it will be discussed more in detail in the next section, the clinician needs a probability in order to choose a treatment or other course of action about the current patient. This probability is computed by generalizing associations between predictors and predictand hidden in a dataset of similar patients. 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 medicine students 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% (or 5/7, upper square) of them eventually developed a disease. The fallacy lies in judging that a new real patient from the full population, who has that particular 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 were used to make prognoses in the full population.

clinical trials

full population

There is a discrepancy in the frequencies of predictand given predictors for the sample dataset and for 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 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

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

🔧 Maybe make clear at the beginning that we denote predictand with Y, predictors with X, dataset with D, and just write $\mathrm{p}(Y \mid X\,D)$.
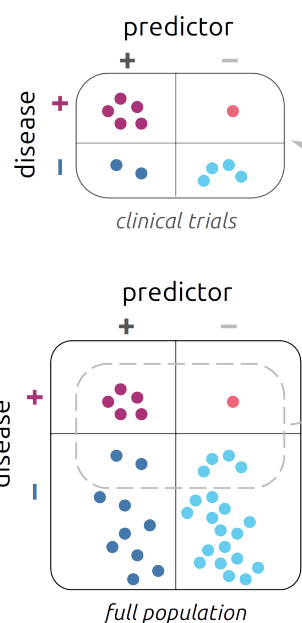
A little counting in the side figure reveals, however, that other frequencies may be relied upon. Consider the full population. Among all patients who developed the disease, 83.3% (or 5/6, upper row) of them had the particular predictor value, while among those who did not develop the disease, 33.3% (or 1/3, lower row) had the particular 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:

$$\mathrm{p}(\mathtt{predictand|predictors}) = \frac{\mathrm{p}(\mathtt{predictors\,|\,predictand, dataset}) \cdot \mathrm{p}(\mathtt{predictand\,|\,population})}{\sum\limits_{\mathtt{predictand}} \mathrm{p}(\mathtt{predictors\,|\,predictand, dataset}) \cdot \mathrm{p}(\mathtt{predictand\,|\,populat})}$$

(2)

In our example we find

$$\mathrm{p}(\mathtt{disease+\,|\,predictor+}) = \frac{\mathrm{p}(\mathtt{predictor+\,|\,disease+, trials}) \cdot \mathrm{p}(\mathtt{disease+\,|\,population})}{\left[\begin{matrix}\mathrm{p}(\mathtt{predictor+\,|\,disease+, trials}) \cdot \mathrm{p}(\mathtt{disease+\,|\,population}) + \\ \mathrm{p}(\mathtt{predictor+\,|\,disease-, trials}) \cdot \mathrm{p}(\mathtt{disease-\,|\,populatio}}\end{matrix}\right.}$$

$$\approx \frac{0.833 \cdot 0.167}{0.833 \cdot 0.167 + 0.333 \cdot 0.833} = 0.33$$

(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 $\mathrm{p}(\mathtt{predictand\,|\,predictors, dataset})$ from the dataset would be the appropriate one to use. But the probabilities $\mathrm{p}(\mathtt{predictors\,|\,predictand, dataset})$ together with Bayes's theorem as in eq. (2) would also lead to exactly the same probability. We thus see that *using the probabilities*

$$\mathrm{p}(\mathtt{predictors\,|\,predictand, dataset})$$

from the dataset is preferable to using $\mathrm{p}(\mathtt{predictand\,|\,predictors, dataset})$. 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 latter is inappropriate. The superiority of using $\mathrm{p}(\mathtt{predictors\,|\,predictand, dataset})$ probabilities is illustrated with a toy example in the inset at the bottom of this page.

The use of dataset probabilities different from $\mathrm{p}(\mathtt{predictand\,|\,predictors, dataset})$ 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). The first case happens for instance 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 for instance 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 judged that the conditional probabilities $\mathrm{p}(\mathtt{predictand\,|\,predictors, dataset})$ still apply. In other words, the patient may still be considered of a member of the subpopulation having those specific predictor values.

Inset: superiority of the "predictors | predictand" approach

We split our learning dataset in two subsets:

- One with 361 datapoints and a ratio of 29.9%/70.1% of conversions to Alzheimer's Disease vs stable Mild Cognitive Impairment. This is used as learning set.
- One with 343 datapoints and a ratio of 63.3%/36.7% of conversions to Alzheimer's Disease vs stable Mild Cognitive Impairment. This is used as a fictive full population.

This partition was made with no systematic sampling of any variates except the conversion variate.

We then 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 "conversion to Alzheimer's Disease", which guarantees 63.3% correct prognoses owing to the base rate. The accuracies (number of prognoses giving more than 50% probability to the correct course) of these four approaches are finally calculate. Here are the results from lowest to highest:

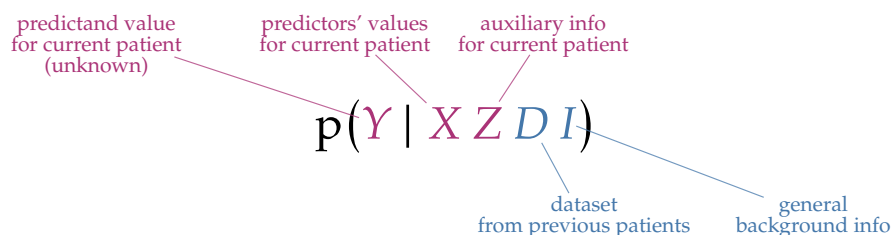| predictand | predictors | coin toss | always predict conversion | predictors | predictand & base rate |
|---|---|---|---|
| 37.3% | 50% | 63.3% | 73.2% |

The "predictand | predictors" approach leads to worse results than a coin toss because of its underlying base-rate fallacy. The "predictors | predictand" approach 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 Ledley-Jaynes machine, are incapable of extracting this useful information.

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; 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 § **??**, the Ledley-Jaynes 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.

🔧 [Luca] testing a slightly different presentation:

The probability necessary for the clinician's decision problem is the following:

---

predictand value
for current patient
(unknown)

predictors' values
for current patient

auxiliary info
for current patient

$$\mathrm{p}(Y \mid X\,Z\,D\,I)$$

dataset
from previous patients

general
background info

It can be rewritten as follows, by the rules for conditional and marginal probabilities:

$$\mathrm{p}(Y \mid X\,Z\,D\,I) \equiv \frac{\mathrm{p}(Y\,X \mid Z\,D\,I)}{\sum_Y \mathrm{p}(Y\,X \mid Z\,D\,I)} \tag{4}$$

The Ledley-Jaynes machine calculates and outputs this probability in several different ways, as demanded by the clinician according to current patient's situation:

1. The patient and the datasets are mutually representative with regard to predictand *and* all predictors; there is no auxiliary information:

$$\mathrm{p}(Y \mid X\,Z\,D\,I) \equiv \frac{\mathrm{p}(Y\,X \mid Z\,D\,I)}{\sum_Y \mathrm{p}(Y\,X \mid Z\,D\,I)} = \frac{\mathrm{p}(Y\,X \mid D\,I)}{\sum_Y \mathrm{p}(Y\,X \mid D\,I)} \tag{5}$$

2. The patient and the datasets are mutually representative with regards to predictand *given* the predictors, but not with regard to the predictors alone; there is no auxiliary information:

$$\mathrm{p}(Y \mid X\,Z\,D\,I) \equiv \frac{\mathrm{p}(Y\,X \mid Z\,D\,I)}{\sum_Y \mathrm{p}(Y\,X \mid Z\,D\,I)} = \mathrm{p}(Y \mid X\,D\,I) \tag{6}$$

3. The patient and the datasets are mutually representative with regards to predictors *given* the predictand, but not with regard to the predictand alone; there is auxiliary information:

$$\mathrm{p}(Y \mid X\,Z\,D\,I) \equiv \frac{\mathrm{p}(Y\,X \mid Z\,D\,I)}{\sum_Y \mathrm{p}(Y\,X \mid Z\,D\,I)} = \frac{\mathrm{p}(X \mid Y\,D\,I)\,\mathrm{p}(Y \mid Z\,I)}{\sum_Y \mathrm{p}(X \mid Y\,D\,I)\,\mathrm{p}(Y \mid Z\,I)} \tag{7}$$

4. Other intermediate situations.

The difference between situations 1 and 2 is that in the former the Ledley-Jaynes machine first updates what it learned from the dataset by including the patient's predictor values, which can rightly be considered as a new datapoint for that set; whereas such update is not done in the latter case. If the dataset has a large number of datapoints, the results from these two situations are typically numerically identical.

Application to the example study

In our present example, all statistics of the dataset are considered relevant for Olivia, Bianca, and Curtis. For these patients we 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 $\mathrm{p}(\texttt{conversion} \mid \texttt{predictors}, \texttt{dataset})$ 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%, owing to her different geographical origin and family history. In her case we must use 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 3. 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
prior probability p(conversion to AD $\mid$ aux info)	0.463	0.65	0.463	0.463
final probability p(conversion to AD $\mid$ predictors, dataset, aux info)	0.302	0.47	0.302	0.703

Table 3. Final probabilities of conversion computed from dataset and auxiliary information

3.3 Assessments of treatments and benefits; final decision

Theory and method

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? and why is it 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 actions, 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. Say there is a dangerous treatment that extends the patient's lifetime by one year if the disease is actually on its course, but shortens the patient's lifetime by five years if the disease is actually not present. Obviously the clinician cannot prescribe the treatment just because the disease is "more probably present than not". If 60 out of 100 treated similar patients actually develop the disease (so "more probable than not" is correct), the clinician has added $1 \times 60 = 60$ years *but subtracted $5 \times 40 = 240$ years* from their combined lifespans. As an opposite example, say 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 20 out of 100 treated similar patients develop the disease, 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* that the patient will develop the disease. And the Ledley-Jaynes machine, as explained in the previous sections, gives us such probabilities.

But the choice between treatments does not only rely 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 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; 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 list the mutually exclusive courses of actions available for the specific patient. These could be treatments, more tests, do nothing, and so on. Often there are *sequences* of decision available, but the utility framework can be applied to them as well (see references above and Raiffa, 1970). The list of courses of action is already patient-dependent: some alternatives may not be suitable (say, owing to allergies), some may be economically too costly, and so on.

Each course of action 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 several guidelines and rules for numerically translating factors such as quality of life and expected lifespan into utilities.

The courses of actions, 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	…
action α	$U_{\alpha a}$	$U_{\alpha b}$	…
action β	$U_{\beta a}$	$U_{\beta b}$	…
…	…	…	…

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

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

$$\bar{U}_i := U_{ia}\,\mathrm{p}(a) + U_{ib}\,\mathrm{p}(b) + \cdots \tag{8}$$

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

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

✏️ Add a couple of comments about the inevitability of the rules of decision theory (Lindley, 1988)

Application to the example study

At present there are no cure for Alzheimer's Disease, although some recent pharmacological agents are shown to extend the time before a patient is cognitively severely impaired 🔧 ref. But for the sake of our case study let us imagine that there, in the near future, are three mutually exclusive treatment options for prevention or retardation of the disease; call them β, γ, δ. And denote the simple option of "no treatment" by α. The clinical conditions to be considered are just two: the patient will have stable Mild Cognitive Impairment, or will convert to Alzheimer's Disease. Denote them by $\neg\mathrm{AD}$ and AD.

We have therefore $4 \times 2 = 8$ possible consequences of the four treatments depending on the two clinical conditions. Our four patients and clinician quantify the utilities, arriving at the utility matrices shown in table 4. Note that Olivia, Ariel, Curtis quantify the benefits of the treatments in exactly the same way, but Bianca's quantification differs slightly 🔧 add an example of why.

The probabilities for the two medical conditions are those found in the previous section, reported in table 3. For brevity we denote just by $\mathrm{p}(\mathrm{AD})$ the probability of conversion given a patient's predictor values, and by $\mathrm{p}(\neg\mathrm{AD}) \equiv 1 - \mathrm{p}(\mathrm{AD})$ the complementary probability of stable Mild Cognitive Impairment, given

	Olivia		Ariel		Bianca		Curtis	
	¬AD	AD	¬AD	AD	¬AD	AD	¬AD	AD
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

Table 4. Utility matrices for the four patients

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

$$\bar{U}_\beta = 9 \cdot (1 - 0.463) + 3 \cdot 0.463 = 7.19 \tag{9}$$

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

	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	β	γ	α	δ

Table 5. Expected utilities and optimal treatment for our four patients

3.4   Additional information provided by the Ledley-Jaynes machine

As already said, the output of the Ledley-Jaynes machine is about the full population of 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 the predictors' importance in the prognosis.

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 Ledley-Jaynes machine can answer this question probabilistically; the answer is plotted in fig. 4. Note that the calculation assumes that the Ledley-Jaynes 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 [12]; 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 four or five "conversion-probability clusters", as evident from the peaks.
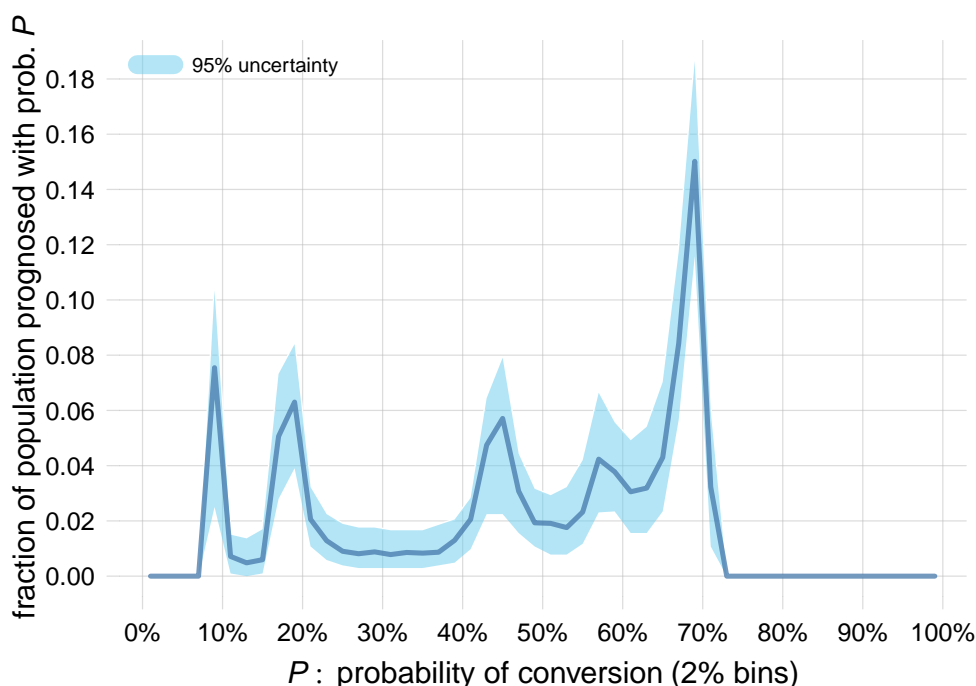
Figure 4. Possible distribution of prognostic probabilities of conversion in the full patient population

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, assuming that the full population has on average the utility matrix of Olivia, table 4. We find these coverage intervals

---

[12] 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 (Jaynes, 1976; MacKay, 2005, § 37.3).

with a 90% probability that the true future proportion will be within each:

$$\alpha\text{: } 21\%\text{–}28\% \qquad \beta\text{: } 2\%\text{–}6\% \qquad \gamma\text{: } 31\%\text{–}42\% \qquad \delta\text{: } 31\%\text{–}40\% \qquad (10)$$

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. This kind of semi-quantitative forecast can be very useful for resource planning.

Imputation of missing data

As mentioned in § **??**, the Ledley-Jaynes 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 illustrated 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. Such imputation can be done at prognostic time, for sensitivity checks; we discuss this more in 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 the comparison of biological hypotheses.

Sensitivity checks

The imputation of missing data at prognostic time is useful for various kinds of sensitivity analysis.

Let us consider for instance the case of Curtis, whose value of Hippocampal Volume is missing (table 1). His clinician thus wonders if the acquisition of this value could lead to a different treatment. The Ledley-Jaynes machine can give a probabilistic answer to this question.

The clinician uses the Ledley-Jaynes machine to find the probability distribution of Curtis's expected utilities (see 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 in the following coverage intervals with 90% probability:

$$\alpha\text{: } 2.97\%\text{–}2.98\% \qquad \beta\text{: } 4.78\%\text{–}4.79\% \qquad \gamma\text{: } 5.89\%\text{–}5.90\% \qquad \boldsymbol{\delta\text{: } 7.02\%\text{–}7.03\%} \qquad (11)$$

(the four corresponding probability histograms, if plotted jointly, would look like distinct vertical lines). It is clear that knowledge of the Hippocampal Volume would be overly unlikely to change Curtis's optimal treatment from δ. The clinician therefore proceeds without this predictor.
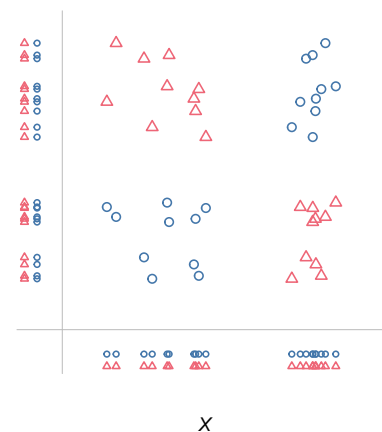
Predictor importance

🔧 [Luca] I think the initial presentation of the problem, here below, still slightly misses the target. The point is that the vague and ill-posed question "which predictor is most important?" hides a decision problem: "if I had to drop one feature for all patients, which one should I drop last?" or "if I had to use only one feature for all patients, which one should I prioritize?". This formulation uniquely determines the relevant metric. I'll rewrite this later.

The question about Curtis in the previous subsection can be generalized to a whole population. How important, in general, is each predictor in prognosing the conversion to Alzheimer's Disease? Predictors that are too invasive or too expensive to obtain and do not make a real difference in the prognosis could be dropped altogether.

To answer this general question, which is too vague (ill-posed), we must first specify which other predictors the predictor of interest is used with, and what we mean by "important".

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, Y. Predictor X, if used by itself, is worthless in distinguishing the two subpopulations, because these have identical marginal distributions (depicted beside the grey lines). If used in conjunction with Y, however, predictor X allows us to identify an individual's subpopulation with full certainty. 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 Y 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.

A predictor's "importance" or "prognostic power" is undefined until we specify a relevant metric, for predictors can be ranked differently by different metrics. From our discussion so far it should be clear that in clinical decision-making the canonical metric is the *utility* which a predictor's presence or absence leads to. This point was partially illustrated with Curtis's example in the previous subsection. What if we want to make an assessment, not for a single patient, but for full population? 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 proved that even just a semi-quantitative assessment leads to better results than using some other "general-purpose" metric (Dyrland et al., 2022a, § 4.2).

The Ledley-Jaynes machine allows us to compute the expected value of virtually any prognostic-importance metric of any subset of predictors available in the dataset. Two facts of paramount advantage are that (a) *the prognostic power of a set of predictors found with the Ledley-Jaynes machine is the maximal one obtainable by **any** inference algorithm*, or in other words it is an intrinsic property of that set of predictors; (b) *the Ledley-Jaynes machine achieves this maximal power*. Thus, if the Ledley-Jaynes machine says that the accuracy obtainable with a given set of predictors is 70%, then we know that no other inference algorithm of any kind can reach a higher accuracy than 70%; inference algorithms that reach lower accuracy can in principle be improved upon. The Ledley-Jaynes machine, by construction, will reach this accuracy.

We illustrate this kind of computation in our example case, using: (a) two metrics, the accuracy and the mutual information (Shannon, 1948; Cover and Thomas, 2006) between a set of predictors and the event of conversion to Alzheimer's Disease; and (b) 27 different sets of predictors:

- every predictor, used individually (12 sets);
- all cognitive-test predictors used together, jointly with demographic information (`Age` and `Sex`);
- `APOE4`, Hippocampal Volume, and demographic information, used together;
- all predictors together minus one, excluding each single predictor 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;
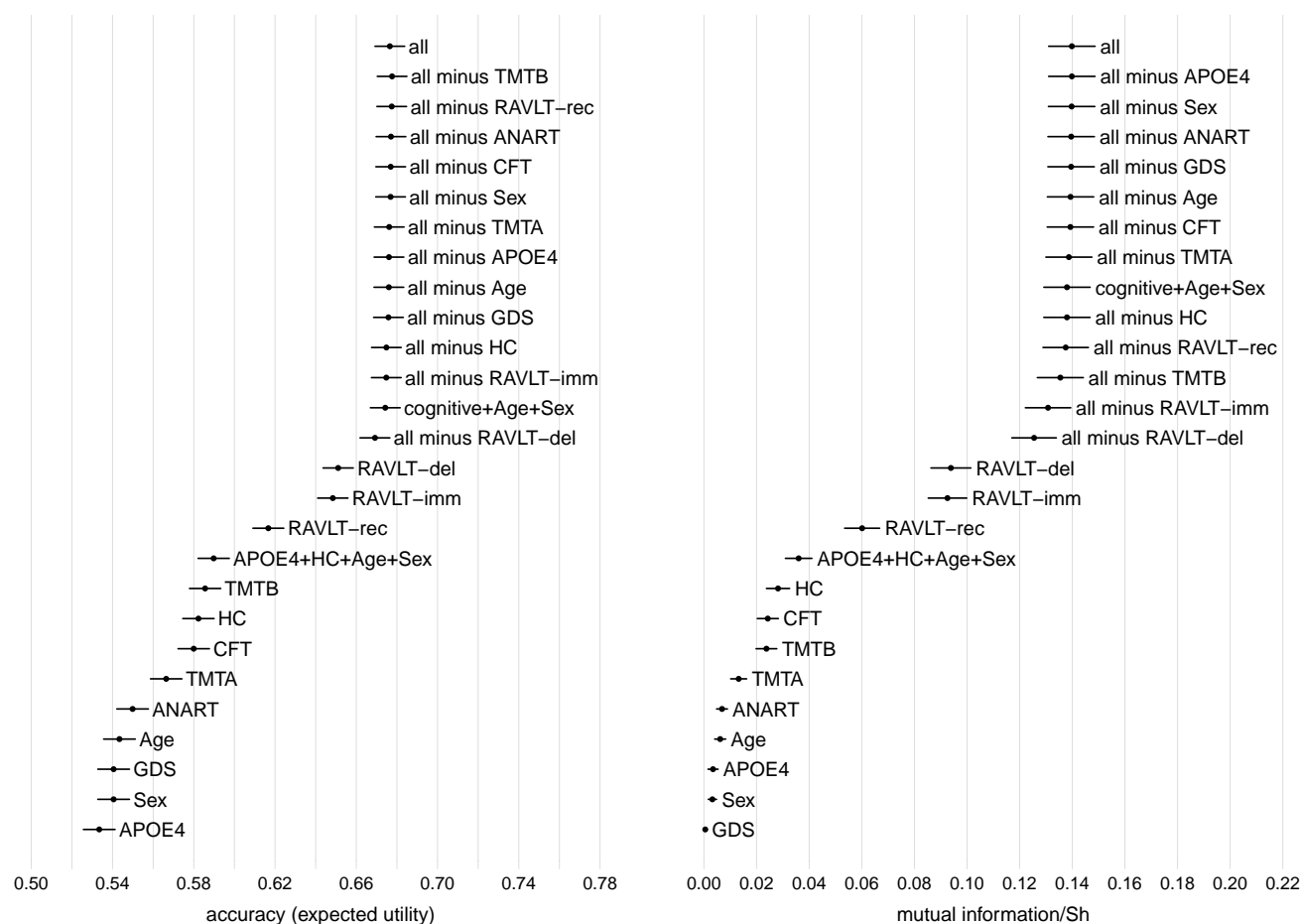
Figure 5. Expected accuracy (for the next new patient) and mutual information of several sets of predictors for the prognosis of conversion to Alzheimer's Disease. Both graphs have been vertically ordered according to increasing values; the two rankings agree within the respective uncertainties. The `all` predictor set is mathematically guaranteed to be optimal according to both metrics, and has therefore be ranked first. Bars show the uncertainty interval ($\pm$ two standard deviations).

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 $1\,\mathrm{Sh}$ if and only if there is a non-constant deterministic function from the former to the latter. 🔧 Consider using conditional entropy instead of mutual info; maybe both. Show the distribution of cAD/sMCI across the main diagonal o f the 12D predictor space

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 present available data?".

The answers to these questions are reported in fig. 5, 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 figure reveals several interesting 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 lead to a prognostic accuracy of around 68%, for any inference algorithm. This fact agrees with the (completely independent) findings in Rye et al. (2022), where a maximal accuracy of 68% was found using an ensemble model; the present analysis also shows that that model managed to achieve the maximal accuracy possible with these predictors (but see § 4 for limitations of that model.)

- `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 of the combined predictors barely decrease if any one of these four predictors is omitted.

- The combined cognitive and demographic variates are better predictors than the combined hippocampal, `APOE4`, and demographic variates.

- `RAVLT-imm`, `RAVLT-del`, and to a lesser degree are good predictors, both when used alone and when used in combination 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).

The Ledley-Jaynes 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 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.

	Olivia	Ariel	Bianca	Curtis
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, final probability				
auxiliary info	none	family history, base rate	none	none
applicable dataset statistics	all	predictor | predictand	all	all
prior probability of conversion	0.463	0.65	0.463	0.463
Available actions and utilities				

Available actions and utilities:

$$
\begin{array}{l}
 & \begin{array}{cc} \neg\text{AD} & \text{AD} \end{array} \\
\begin{array}{l}\text{treatment }\alpha\\ \text{treatment }\beta\\ \text{treatment }\gamma\\ \text{treatment }\delta\end{array} & \left[\begin{array}{cc} & \\ & \\ & \\ & \end{array}\right]
\end{array}
$$

Olivia: $\begin{bmatrix}10 & 0\\ 9 & 3\\ 8 & 5\\ 0 & 10\end{bmatrix}$ Ariel: $\begin{bmatrix}10 & 0\\ 9 & 3\\ 8 & 5\\ 0 & 10\end{bmatrix}$ Bianca: $\begin{bmatrix}10 & 0\\ 8 & 3\\ 7 & 5\\ 0 & 10\end{bmatrix}$ Curtis: $\begin{bmatrix}10 & 0\\ 9 & 3\\ 8 & 5\\ 0 & 10\end{bmatrix}$

	Olivia	Ariel	Bianca	Curtis
Outputs of Ledley-Jaynes machine				
$\mathrm{p}(\text{AD} \mid \text{predictors})$	0.302	0.302	0.302	0.703
$\mathrm{p}(\text{predictors} \mid \text{AD})/10^{-12}$	8.97	8.97	8.97	1.14
$\mathrm{p}(\text{predictors} \mid \neg\text{AD})/10^{-12}$	18.6	18.6	18.6	0.343
final probability of conversion	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	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$	$\boldsymbol{\alpha}$	$\boldsymbol{\delta}$

Table 6. ✎ Summary. The basic data that distinguish Ariel, Bianca, Curtis from Olivia are in red.

3.5 Summary of the case study

Summary in table 6.

4 DISCUSSION

🔧 [Luca] There are a couple different ways to introduce the discussion. I'll write some variants as soon as possible

The differences among patients which are relevant to clinical diagnosis/prognosis and decision-making can be approximately divided into three categories:

1. differences in the availability and values of a core set of clinical predictors, for which we have population-wide statistical information;

2. differences in the availability and values of auxiliary and usually "softer" predictors, such as geographical or family background; or more generally of predictors not belonging to the core set;

3. differences in the availability and benefit of clinical courses of actions, such as treatments and further tests.

These differences are intimately connected with, and partially stem from, several hallmarks of a clinician's tasks:

First, there is often an irreducible element of uncertainty in inferring a medical or biological condition from a set of predictors. In mathematical terms, there is no function from predictors to predictands ✏ refer to a figure in the previous sections. This is especially often the case for predictors that are more readily available and less invasive, and therefore more desirable. The irreducible uncertainty originates from the natural variability within a population. Yet this variability can be harnessed, and the way to harness it can crucially vary from patient to patient. We illustrated this with an example in the inset of § 3.2, p. 18, which showed that any relation from predictors to predictands lead to poor prognoses, whereas taking into account the variability in the predictors *given* the predictand improved our prognoses.

Second, owing to the irreducible prognostic uncertainty, the clinician's ultimate task is not one of classification or regression, but one of *decision-making under risk*. This is clear if we consider that a clinician may have *three or more* courses of action to choose from, even if the unknown medical condition only has *two* possible "class labels". A pure class label is therefore of no help to the clinician, who instead needs to know the uncertainty or probability of the label, in order to evaluate risks and benefits, and from these make a final decision. The choice of a course of action is moreover the moment were patient differences can be very dramatic.

✏ to be finished

✏ Possible additional point of view Let's assume that current technology allows for a particular maximal rate of extraction of information per unit time, and if we assume that different inference algorithms reach this maximum rate (or do not reach it by a common amount). Information and communication theory (MacKay, 2005, I–II) then suggests that if one inference algorithm is, say 100 times faster than a Ledley-Jaynes machine, its output has also 100 times less information than a Ledley-Jaynes machine's. Do we need this extra information? The example case illustrated in the present work shows that part of this extra information is necessary for personalized medicine; and the remaining amount can be very useful or at times necessary for sensitivity analyses or resource planning.

4.1 Counters to 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 Ledley-Jaynes machine

requires inputs about a patient's specific predictors, relevant statistical relations and auxiliary data, and treatment utilities.

Some researchers may wonder: can such additional input be avoided? They fear that errors could sneak in through these extra inputs.

This question is answered by a mathematical theorem at the very core of decision theory [13], 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 don't know. If you use a decision-making algorithm that does not ask you for utilities, then the algorithm is internally supplying utilities beyond your control (and probably divorced from your 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 lie 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's more difficult to analyse which are inappropriate and how they should be changed.

🖊 Subtly hidden disastrous consequences of not following normative decision theory: An algorithm can lead to saving 85 000 patients out of 100 000 and be deemed a success. But if the ideal algorithm had been used, 95 000 patients would actually have been saved. What shall we say to the families of the 10 000 patients who could have been saved but weren't?

4.2 Range of application of Ledley-Jaynes machines

The range of application of the Ledley-Jaynes machine used in the present study has two kinds of bound: computational and theoretical.

As mentioned in § **??** and explained in appendix A.2, the fact that the Ledley-Jaynes machine extracts all available information from the dataset also makes it computationally expensive. It is at present impossible to use with high-dimensional predictors (if our dataset had included a predictor such as a $128 \times 128 \times 128$ grayscale 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 such fast algorithms together with a Ledley-Jaynes machine, as a post-processor of their raw output. This allows us to extract useful information usually hidden in such output at a low computational cost (Dyrland et al., 2022b). Such information can be then used for clinical decision making as illustrated in the present work.

As explained in § 3.2, the essentially sole assumption underlying the Ledley-Jaynes machine's inference and its practical used 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 technical jargon, partial or conditional exchangeability applies). This precludes using the Ledley-Jaynes machine to forecast how the statistics of the full population could change in the future. However, the machine can be used for time-dependent (transversal 🧩 correct?) inferences within a stable population, such as forecasts of the future time of disease onset, expected lifelength, and similar. For example, if data about the time of conversion to Alzheimer's Disease were

---

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

available in the dataset, the Ledley-Jaynes machine could forecast not only *whether*, but also *when* the conversion could take place (cf. e.g. De la Cruz-Mesía et al., 2007).

A   APPENDICES

A.1   Predictors and predictand

For the present study, data used in the learning stage of § 3.1 comes from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. This longitudinal multicenter study is designed to develop and validate neuroimaging and biochemical biomarkers for early detection, monitoring, and treatment of Alzheimer's disease (Petersen et al., 2010). The present sample of 704 subjects consisted of all ADNI participants meeting the criteria for Mild Cognitive Impairment at their first, baseline assessment who additionally had a minimum of three study visits (that is, baseline visit and at least two additional visits), and three MRI examinations. A reevaluation of each subject's diagnostic status was conducted at each study visit, and this longitudinal diagnostic label was used to partition our sample into 325 subjects converting to Alzheimer's Disease following the first study visit, and 379 subjects remaining stable with Mild Cognitive Impairment. Criteria used for classifying subjects as having Mild Cognitive Impairment and Alzheimer's Disease, as well as ADNI's general criteria for subject inclusion, are described in McKhann et al. (1984); Petersen et al. (2010).

A.2   Mathematical details about the Ledley-Jaynes machine

As discussed in § **??**, the Ledley-Jaynes machine explores the space of possible distributions of frequencies of all 13 variates listed in table 1, for the full population of patients from which the dataset originates. In the present study it does so by using a total of 1535 independent parameters to represent the distributions, 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, 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 indicates that more parameters are needed. In the present study the data indicates, on the contrary, that fewer than 250 parameters would be enough. More details on the mathematical representation can be found in Dunson and Bhattacharya (2011); see also Rossi (2014); Rasmussen (1999).

✎ Should I add a more precise description of the mathematical representation?

✎ Is the part below superfluous? I think it could be interesting for readers from machine learning

There is a fundamental difference in how the Ledley-Jaynes machine and most popular machine-learning algorithms (including neural networks, random forests, support-vector machines, excluding gaussian processes) work. The latter do, at bottom, an optimization, looking for the minimum of some error function. The Ledley-Jaynes machine does a full *space exploration* and *averaging*, as explained in § **??**. Inference and generalization in fact essentially rely on averaging operations in problems such as the present one (de Finetti 1930, 1937; Dawid 2013; Bernardo and Smith 2000, §§ 4.2–4.3; see also Self and Cheeseman 1987). The optimization done by most machine-learning algorithms is an approximate form of averaging – assuming or hoping that most of the mass to be averaged is around the extremum (MacKay, 1992a; Murphy, 2012, ch. 16). But the underlying necessity of a proper averaging becomes manifest in many of the obligatory procedures that go together with training a machine-learning algorithm; cross-validation for instance (MacKay, 1992b). 🔧 Add note: "ensembling" in ML is not this kind of averaging (Murphy, 2022, § 18.2)

Figure 6. ✏ Examples of a-priori probable candidates of frequency distribution for a variate such as `RAVLT-del` or `RAVLT-rec`

This difference explains why the Ledley-Jaynes machine is computationally much more expensive than other algorithms, but also why its output is informationally so rich, and why it does not need any validation datasets, test datasets, other data splits, or cross-validation procedures (it can be proved 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; see e.g. Porta Mana 2019; Fong and Holmes 2020).

✏ one more remark about extremum-search being equivalent to making a choice, but the utilities are not controlled by the patient and not flexible.

✏ possibly add a plot showing the distributions considered reasonable by the machine

A.3 Computational details

The learning stage, with 13 variates and 704 datapoints, took less than 5 hours. The computation was done with 16 parallel 3.0–4.80 GHz cores. After that, calculation of probabilities and expected utilities for any single patient is immediate. The mutual information and accuracy analysis of § 3.4 took roughly 1 h.

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

Details of all funding sources should be provided, including grant numbers if applicable. Please ensure to add all necessary funding information, as after publication this is no longer possible.

ACKNOWLEDGMENTS

PGLPM thanks Soledad Gonzalo Cogno and Iván Davidovich for inspiring discussions; 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.

SUPPLEMENTAL DATA

Supplementary Material should be uploaded separately on submission, if there are Supplementary Figures, please include the caption in the same file as the figure. LaTeX Supplementary Material templates can be found in the Frontiers LaTeX folder.

DATA AVAILABILITY STATEMENT

The datasets [GENERATED/ANALYZED] for this study can be found in the [NAME OF REPOSITORY] [LINK].

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)

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`

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

Back, A. and Keith, W. (2019). *Bayesian Neural Networks for Financial Asset Forecasting*. Master's thesis, KTH Royal Institute of Technology, Stockholm. `http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-252562`

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

Barnard, G. A., Bernardo, J. M., Dinges, H., Shafer, G., Silverman, B. W., Smith, C. A. B., et al. (1982). Discussion of paper by D. V. Lindley. Reply. *Int. Stat. Rev.* 50, 11–26. DOI:`10.2307/1402449`, DOI: `10.2307/1402450`, DOI:`10.2307/1402451`, DOI:`10.2307/1402452`, DOI:`10.2307/1402453`, DOI:`10.2307/1402454`, DOI:`10.2307/1402455`, DOI:`10.2307/1402456`. See Lindley (1982)

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

Chatfield, C., Lindley, D. V., Ehrenberg, A. S. C., Bound, J. A., Barnard, G. A., Nelder, J. A., et al. (1993). Discussion of the paper by Draper, Hodges, Mallows and Pregibon. *J. R. Stat. Soc. A* 156, 28–37. DOI: `10.2307/2982858`. See Draper et al. (1993)

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`

Draper, D., Hodges, J. S., Mallows, C. L., and Pregibon, D. (1993). Exchangeability and data analysis. *J. R. Stat. Soc. A* 156, 9–28. DOI:`10.2307/2982858`. See also discussion in Chatfield et al. (1993)

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

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

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

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

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)

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

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

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`

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. (1982). Scoring rules and the inevitability of probability. *Int. Stat. Rev.* 50, 1–11. DOI:`10.2307/1402448`. See also discussion and reply in Barnard et al. (1982)

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`

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

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/1960/03 000/Mathematical_Models_in_Medical_Diagnosis.2.aspx`

MacKay, D. J. C. (1992a). Bayesian interpolation. *Neural Comput.* 4, 415–447. `http://www.infere nce.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/archive s/fall2016/entries/paradox-simpson`. First publ. 2004

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

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`

Murphy, K. P. (2022). *Probabilistic Machine Learning: An Introduction* (Cambridge, USA: MIT Press). `https://probml.github.io/pml-book/book1.html`

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:`10.1002/047003 3312`

Osband, I., Wen, Z., Asghari, M., Ibrahimi, M., Lu, X., and Van Roy, B. (2021). Epistemic neural networks. arXiv DOI:`10.48550/arXiv.2107.08924`

Pearce, T., Leibfried, F., Brintrup, A., Zaki, M., and Neely, A. (2020). Uncertainty in neural networks: Approximately Bayesian ensembling. *Proc. Mach. Learn. Res.* 108, 234–244

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/a rchives/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:`10.485 50/arXiv.1908.08741`

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-1 999a.pdf`

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.htm l`, `https://archive.org/details/artificial-intelligence-a-modern-appro ach-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

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/deta ils/bstj27-4-623`, `http://math.harvard.edu/~ctm/home/text/others/shan non/entropy/entropy.pdf`

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/arc` `hives/sum2021/entries/paradox-simpson`

Tribus, M. (1969). *Rational Descriptions, Decisions and Designs* (New York: Pergamon). DOI:`10.101` `6/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

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