Personalized prognosis & treatment using Turing-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** [1,2], **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

Patients with Mild Cognitive Impairment have an increased risk of a trajectory toward Alzheimer's Disease, and early identification of these patients is essential to provide the most efficient treatment. Studies identifying predictors of the prognosis of MCI patients are thus called for.

Although results from a routine clinical examination are essential, several situational factors must be taken into account to ogtain a personalized approach to prognosis, prevention, and treatment.

- the *kinds* of clinical data and evidence available for prognosis;
- the *outcomes* of the same kind of clinical data and evidence;
- the kinds of treatment or prevention strategies available, owing to different additional medical factors such as physical disabilities, different attitudes toward life, different family networks and possibilities of familial support, different economic means;
- the advantages and disadvantages, benefits and costs of the same kinds of treatment or prevention strategies; the patient has a major role in the quantification of such benefits and costs;
- finally, the initial evaluation by the clinician – which often relies on too subtle clues (family history, regional history, previous case experience) to be considered as measurable data.

Statistical decision theory is the normative quantification framework that takes into account these fundamental differences. Medicine has the distinction of having been one of the first fields to adopt this framework, exemplified in brilliant old and new textbooks on clinical decision-making.

Clinical decision-making makes allowance for these differences among patients through two requirements. First, the quantification of prognostic evidence on one side, and of benefits and costs of treatments and prevention strategies on the other, must be clearly separated and handled in a modular way. Two patients can have the same prognostic evidence and yet very different prevention options. Second, the quantification of independent prognostic evidence ought to be in the form of *likelihoods about the health condition* (or equivalently of likelihood ratios, in a binary case), that is, of the probabilities of the observed test outcomes given the hypothesized health conditions. Likelihoods from independent clinical tests and predictors can then be combined with a simple multiplication; for one patient, we could have three kinds of predictor available; for another, we could have five. The clinician's pre-test assessment is included in the form of a probability. These patient-dependent probabilities are combined with the patient-dependent costs and benefits of treatment or prevention to arrive at the best course of action for that patient. To be optimal and logically consistent, this clinical decision making *must* take a statistical decision theory into account.

The present study uses different statistical approached to investigate the prognostic power of results from neuropsychological and Magnetic Resonance Imaging examinations, demographic data, and genetic information about Apolipoprotein-E4The to predict the onset of Alzheimer's Disease in patients defined as mildly cognitively impaired at a baseline examination. The longitudinal data used come from the ADNI database.

The prognostic power of these predictors is quantified in the form of a combined likelihood for the onset of Alzheimer's disease. As a hypothetical example application of personalized clinical decision making, three patient cases are considered where a clinician starts with prognostic uncertainties, possibly coming from other tests, of 50%/50%, 25%/75%, 75%/25%. It is shown how these pre-test probabilities are changed by the predictors. 🔧 update this

🔧 rewrite following This quantification also allows us to rank the relative prognostic power of the predictors. It is found that several neuropsychological examinations have the highest prognostic power, much higher than the genetic and imaging-derived predictors included in the present set.

Several additional advantages of this quantification framework are also exemplified and discussed in the present work:

- missing data are automatically handled, and results having partial data are not discarded; this quantification, therefore, also accounts for patient-dependent availability of *non-independent* predictors;
- no modelling assumptions (e.g., linearity, gaussianity, functional dependence) are made;
- the prognostic power obtained is intrinsic to the predictors, that is, it is a bound for *any* prognostic algorithm;
- variability ranges of the results owing to the finite size of the sample data are automatically quantified.
- the values obtained, being probabilities, are more easily interpretable than scores of various kinds.

0 EACH PATIENT IS UNIQUE

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. In fact, this is why they recently underwent a clinical examination, including a battery of cognitive tests. Today they received the results from this examination. Based on these results, available clinical statistical data, and other relevant information, their clinician will assess their risk of developing Alzheimer, and then make a decision among a set of mutually exclusive preventive-treatment options, together with the patients and their relatives.

Besides a shared MCI diagnosis 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 difference between her and the other three patients:

- Olivia and Ariel turn out to have exactly identical clinical results and age. They would also get similar benefits from the available preventive-treatment options. Ariel, however, comes from a different geographical region, which presents a higher rate of conversion from MCI to Alzheimer's Disease. Moreover, Ariel comes, unlike Olivia, from a family with a heavy history of Alzheimer's Disease. Because of this geographical and family background, the clinician judges a priori a 65% probability that Ariel's cognitive impairment will convert to Alzheimer's Disease

- Olivia and Bianca also have exactly the same clinical results and age. They 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 preventive options slightly riskier for her.

- Olivia and Curtis have different results on all neuropsychological tests and Hippocampal-Volume measure, in disfavour of Curtis. Olivia furthermore does not have the risky Apolipoprotein-E4 (APOE4) allele (Liu et al., 2013) whereas Curtis has. 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 get similar benefits from the preventive options. Note that one clinical result of Curtis's (hippocampal volume) is missing.

We can categorize these differences, schematized in the side figure, as "difference in auxiliary information" (Olivia and Ariel), "difference in treatment benefits" (Olivia and Bianca), "difference in clinical predictors" (Olivia and Curtis).

Considering the similarities and differences among these patients, which treatments are optimal and should prescribed to them?

Our main purpose in the present work is to illustrate, using the four fictitious patients above as example, how this clinical decision-making problem can today be solved methodically, exactly, and at low computational cost, when the available prognostic clinical information involves one-dimensional or categorical variates. The solution method integrates available clinical statistical data with each new patient's unique combination of clinical results, auxiliary information, and treatment benefits.

---

[1] Fictive characters; any reference to real persons is purely coincidental

In our example we shall find that – despite the many factors in common among our four patients, even despite the identical clinical results for Olivia, Ariel, Bianca, and despite the identical probability of conversion for Olivia and Bianca – *the optimal treatment option for each patient is different from those for the other three*. This result exemplifies the importance of differences among patients with regard to clinical results, auxiliary information, or preventive benefits.

The method used is none other than decision theory, the combination of probability theory and utility theory (von Neumann and Morgenstern, 1955; Raiffa and Schlaifer, 2000; Raiffa, 1970; Lindley, 1988; Kreps, 1988; Jaynes, 2003, chs 13–14). Medicine has the distinction of having been one of the first fields to adopt it (Ledley and Lusted, 1959), with old and new brilliant textbooks (Weinstein and Fineberg, 1980; Sox et al., 2013; Hunink et al., 2014) that explain and exemplify its application.

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, ch. IV; Jaynes, 2003, chs 1–2). The present method can therefore be seen as the application of an *ideal* machine-learning algorithm. We call it the *Turing-Jaynes machine* in homage to Turing, who put this method and its underlying algorithms into practice with automated devices (Good, 1979), and Jaynes, who clearly explained the inductive logic behind such a machine (Jaynes, 2003). The Turing-Jaynes machine is "ideal" in the sense of being free from approximations, from special modelling assumptions, and from limitations in its informational output; not "ideal" in the sense of being impracticable. In fact, in the present work we show that for some kinds of dataset this ideal machine-learning algorithm can be applied at insubstantial computational cost. It is preferable to popular algorithms such as neural networks, random forests, support-vector machines, which are unsuited to clinical decision-making problems owing to their output limitations. We discuss this matter further in § 6.

✎ Add other advantages of the Turing-Jaynes machine:

- it does not make assumptions, besides natural assumption of smoothness of full-population frequency distribution
- can be used with partially missing clinical data
- can be used with binary or continuous predictands
- it tells us the maximum predictive power of the predictors
- it quantifies how prediction could change if we had more sample data
- it can be applied on-the-fly to each new patient

and goals, results, and some synopsis

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

How does the clinician use the Turing-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 § 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 Turing-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 § 1-§ 2-§ 3 order; whereas readers unfamiliar with these principles (who may include readers with a specialized background in machine learning) read them in § 3-§ 2-§ 1 order.

1 LEARNING

🧩 Maybe the proper learning stage, before application to patients, should be in a separate section?

1.1 Theory and method

In the learning stage the Turing-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

---

Inset: Inferential and decision-making steps

1. Infer the joint, full-population frequencies of predictors and predictand with the Turing-Jaynes machine, using available datasets.
   Input the clinical data of the patient, and select the appropriate output probabilities to be used in the next step.

2. Assess which statistics of the dataset can be applied to the present patient. Take their probabilities from step 1.
   Assess auxiliary clinical information available for this patient; quantify it in a prior probability.
   From these, the final probability of the patient's true medical condition is calculated.

3. Assess the clinical courses of action (treatments, more tests, and so on) available for the present patient.
   Assess the utility (benefit and loss) of each course of action, depending on each possible medical condition of the patient.
   Combine the utilities with the final probabilities from step 2. Choose the course of action having maximal expected utility.
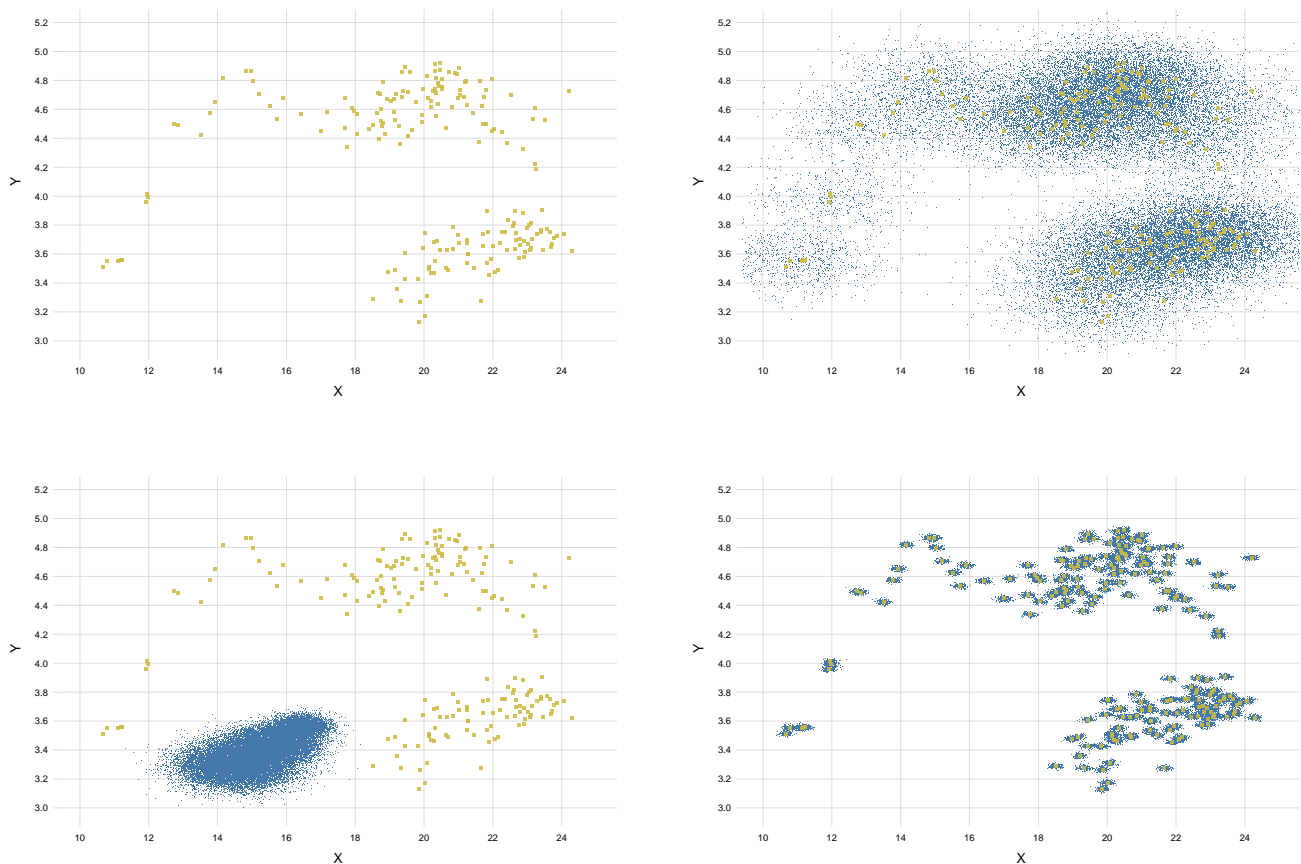

Figure 1. ✏ Upper-left: Sample data. Upper-right: candidate frequency distribution that fits the data and does not look unnatural. Lower-left: candidate distribution that might look natural but doesn't fit the sample data. Lower-right: candidate distribution that fits the data very well but looks unnatural.

considered a member only *conditionally* on particular variate values. We shall discuss this point with an example in § 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 Turing-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 1 show a fictitious sample data and various candidate frequency distributions ✏ ....

✏ EITHER HERE OR IN APPENDIX A.1: Add some more intuition and details about the maths, principles, and characteristics (Dunson and Bhattacharya, 2011; Rossi, 2014; Rasmussen, 1999).

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

This is the *maximal amount of information* that can be extracted from the dataset. From it we can indeed quickly calculate any quantity typically outputted by specific or approximate algorithms. For example:

- *Conditional probability, "discriminative" algorithms:* if we are interested in the probability of Z given X, Y, we calculate $\mathrm{p}(Z \mid X, Y) \coloneqq \mathrm{p}(X, Y, Z) / \sum_Z \mathrm{p}(X, Y, Z)$.

- *Conditional probability, "generative" algorithms:* if we are interested in the probability of X, Y given Z, we calculate $\mathrm{p}(X, Y \mid Z) \coloneqq \mathrm{p}(X, Y, Z) / \sum_{X,Y} \mathrm{p}(X, Y, Z)$.

- *Regression or classification:* if we are interested in the average value of Z given X, Y, we calculate $\mathrm{E}(Z \mid X, Y) \coloneqq \sum_Z Z \, \mathrm{p}(Z \mid X, Y)$. The "noise" around this average value is moreover given by $\mathrm{p}(Z - \mathrm{E} \mid X, Y)$.

- *Functional regression:* if Z turns out to be a function f of X, Y, then the probability will be a delta distribution: $\mathrm{p}(Z \mid X, Y) = \delta[Z - f(X, Y)]$. Thus, the Turing-Jaynes machine always recovers a functional relationship if there is one, including its noise distribution.

We will see that no one of these quantities can be used *alone* to solve our clinical decision problem for all future patients.

Use of the Turing-Jaynes machine has further advantages and yields additional useful information:

- *Discrete or continuous variates:* the variate to be prognosed can be not only binary or discrete, as in the present case, but also continuous.

- *Partially missing data:* data having missing values for some variates can be fully used, both in the learning dataset and in the prognostic results of new patients.

- *Maximal predictive power of the variates:* from the probability distribution $\mathrm{p}(X, Y, Z, \dots)$ we can calculate the mutual information between any two sets of variates, for example between Z and $\{X, Y\}$. This quantity tells us what is the maximal predictive power – which can be measured by accuracy or other metrics – from one set to the other that can be achieved by any inference algorithm (MacKay, 2005; Good and Toulmin, 1968; Cover and Thomas, 2006). Functional dependence is included as a special case; for example, a binary variate Z is a non-constant function of $\{X, Y\}$ if and only if their mutual information [2] equals $1\,\mathrm{Sh}$.

- *Variability owing to limited sample size:* we can calculate how much any of the quantities listed so far – from average values to mutual information – could change if more sample data were added to our 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 Turing-Jaynes machine.

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 our case 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 § ✎ *** 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.

---

[2] The "shannon" (Sh) is a measurement unit of information, as specified by the ISO (International Organization for Standardization) (2008).

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,[3] and therefore cannot be used in the decision-making stage, see § 3. It is possible to transform this score into a proper probability (Dyrland et al., 2022b), but this possibility does not solve the 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.

1.2   Application to the example study

✎ Show some graphs about the full-population distribution inferred by the Turing-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 Hippocampus Volume is missing; the Turing-Jaynes machine has no difficulty with this.

Given these predictor values the Turing-Jaynes machine yields the following probabilities of main importance for the application in the next sections:

---

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

	Olivia	Ariel	Bianca	Curtis
Age	75.4	75.4	75.4	63.8
Sex	F	F	F	M
Hippocampus volume (HC)$/10^{-3}$	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-imm ediate memory	36	36	36	20
RAVLT-del ayed recall	5	5	5	0
RAVLT-rec ognition	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
$\mathrm{p}(\text{conversion to AD} \mid \text{predictors})$	0.302	0.302	0.302	0.703
$\mathrm{p}(\text{predictors} \mid \text{conversion to AD})/10^{-12}$	8.97	8.97	8.97	1.14
$\mathrm{p}(\text{predictors} \mid \text{no conversion to AD})/10^{-12}$	18.6	18.6	18.6	0.343

Table 2. ✎ Probabilities computed by the Turing-Jaynes machine

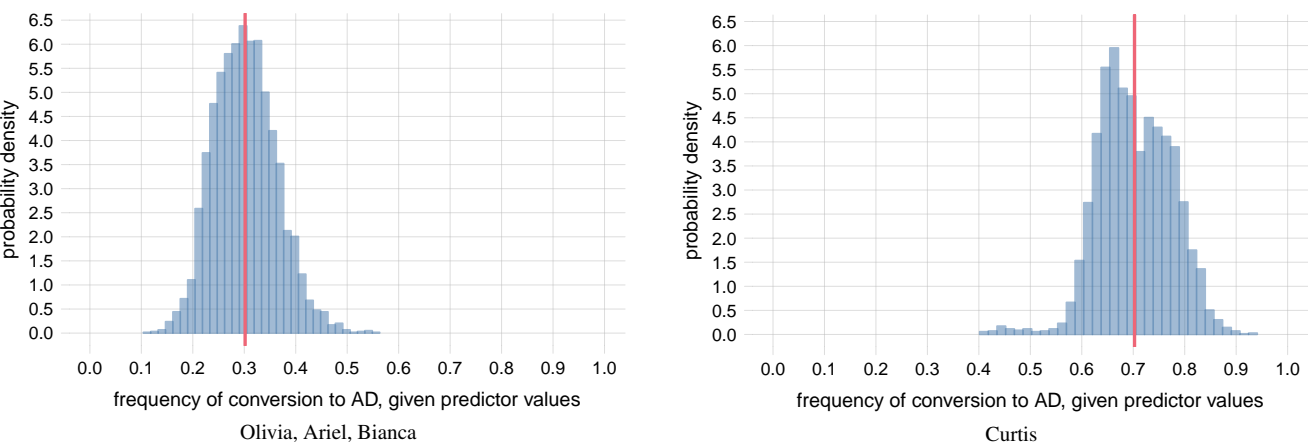Olivia, Ariel, Bianca   Curtis

Figure 2. 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 $\mathrm{p}(\text{conversion to AD} \mid \text{predictors}, \text{dataset})$

As an illustration of the additional information provided by the Turing-Jaynes machine, fig. 2 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 $\mathrm{p}(\text{conversion to AD} \mid \text{predictors}, \text{dataset})$ 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 (Joint Committee for Guides in Metrology), 2008).

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

2 ASSESSMENT OF POPULATION AND AUXILIARY INFORMATION

2.1 Theory and method

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.
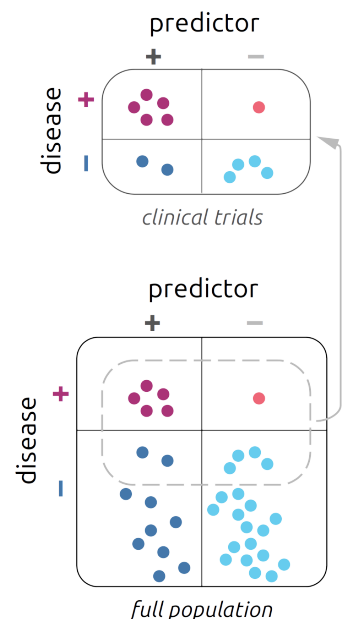
There is a discrepancy in the frequencies of predictand given predictors for the sample dataset and for the full population, because the proportion of positive vs negative disease cases in the latter has some value, 16.7%/83.3% in our example, whereas the samples for the trials (dashed line in the lower 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.

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:

$$
\mathrm{p}(\mathsf{predictand}|\mathsf{predictors}) = \frac{\mathrm{p}(\mathsf{predictors} \mid \mathsf{predictand}, \mathsf{dataset}) \cdot \mathrm{p}(\mathsf{predictand} \mid \mathsf{population})}{\sum\limits_{\mathsf{predictand}} \mathrm{p}(\mathsf{predictors} \mid \mathsf{predictand}, \mathsf{dataset}) \cdot \mathrm{p}(\mathsf{predictand} \mid \mathsf{population})} \quad (1)
$$

In our example we find

$$
\mathrm{p}(\mathsf{disease+} \mid \mathsf{predictor+}) = \frac{\mathrm{p}(\mathsf{predictor+} \mid \mathsf{disease+}, \mathsf{trials}) \cdot \mathrm{p}(\mathsf{disease+} \mid \mathsf{population})}{\left[\begin{array}{l}\mathrm{p}(\mathsf{predictor+} \mid \mathsf{disease+}, \mathsf{trials}) \cdot \mathrm{p}(\mathsf{disease+} \mid \mathsf{population}) + \\ \quad \mathrm{p}(\mathsf{predictor+} \mid \mathsf{disease-}, \mathsf{trials}) \cdot \mathrm{p}(\mathsf{disease-} \mid \mathsf{population})\end{array}\right]}
$$

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

$$(2)$$

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 $\mathrm{p}(\mathsf{predictand} \mid \mathsf{predictors}, \mathsf{dataset})$ from the dataset would be the appropriate one to use. But the probabilities $\mathrm{p}(\mathsf{predictors} \mid \mathsf{predictand}, \mathsf{dataset})$ together with Bayes's theorem as in eq. (1) would also lead to exactly the same probability. We thus see that *using the probabilities*

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

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

The use of dataset probabilities different from $\mathrm{p}(\mathsf{predictand} \mid \mathsf{predictors}, \mathsf{dataset})$ 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 (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

---

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.

No systematic sampling of any variates was made in this partition, besides the conversion variate.

We then make a prognosis for each of the 343 new patients with four approaches: (a) using the probabilities $\mathrm{p}(\mathsf{predictand} \mid \mathsf{predictors}, \mathsf{dataset})$, as typical of machine-learning algorithms; (b) using $\mathrm{p}(\mathsf{predictors} \mid \mathsf{predictand}, \mathsf{dataset})$ 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. Here is the accuracy (that is, the number of prognoses giving more than 50% probability to the correct course) of each approach, ranked:

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 instead 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 Turing-Jaynes machine, are incapable of extracting this useful information.

---

the specific patient. For a patient coming from a different region, for example, it may be judged that the conditional probabilities $\mathrm{p}(\mathsf{predictand} \mid \mathsf{predictors}, \mathsf{dataset})$ still apply. In other words, the patient may still be considered of a member of the subpopulation having those specific predictor values.

This topic is complex and of extreme importance for inference, but its study is not the goal of the present work. We refer the readers to the brilliant paper by Lindley & Novick (1981) for further discussion, and the works by Malinas and Bigelow (2016) and Sprenger and Weinberger (2021) about Simpson's paradox, to which this topic is related. 🧩 Maybe add refs to Pearl (and Russell) about the notion of causality – which is somewhat circular, however.

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 § 1, the Turing-Jaynes machine allows us to quickly calculate conditional probabilities $\mathrm{p}(Y \mid X, \mathsf{dataset})$ for any desired variate subsets Y and X required by the patient's relevant population.

2.2 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 $\mathrm{p}(\mathsf{conversion} \mid \mathsf{predictors}, \mathsf{dataset})$ 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 $\mathrm{p}(\mathsf{conversion\ to\ AD} \mid \mathsf{aux\ info})$	0.463	0.65	0.463	0.463
final probability $\mathrm{p}(\mathsf{conversion\ to\ AD} \mid \mathsf{predictors}, \mathsf{dataset}, \mathsf{aux\ info})$	0.302	0.47	0.302	0.703

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

3 ASSESSMENTS OF TREATMENTS AND BENEFITS; FINAL DECISION

3.1 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 Turing-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 $U_{ia}, U_{ib}, \ldots$ with respect to the probabilities $\mathrm{p}(a), \mathrm{p}(b), \ldots$ of the clinical conditions $a, b, \ldots$:

$$\bar{U}_i := U_{ia}\,\mathrm{p}(a) + U_{ib}\,\mathrm{p}(b) + \cdots \tag{3}$$

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)

3.2 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.

	Olivia		Ariel		Bianca		Curtis	
	$\neg\mathrm{AD}$	AD	$\neg\mathrm{AD}$	AD	$\neg\mathrm{AD}$	AD	$\neg\mathrm{AD}$	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 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 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

$$\bar{U}_\beta = 9 \cdot (1 - 0.463) + 3 \cdot 0.463 = 7.19 \tag{4}$$

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

4 ADDITIONAL INFORMATION PROVIDED BY THE TURING-JAYNES MACHINE

As already said, the output of the Turing-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, sensitivity checks, and the investigation of the predictors' importance in the prognosis.

4.1 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 Turing-Jaynes machine can answer this question probabilistically, and obviously with great uncertainty. Yet there may be surprising details that can be assessed with a high degree of certainty. The answer is illustrated in fig. 3, assuming that the Turing-Jaynes machine will be regularly updated with the new patients' data (the calculation could also be made without this assumption). The dark-blue curve represents, for each of the 50 probability bins, the median value of the fraction of patients who will have conversion probability in that bin. The light-blue bands around the median are 50%, 75%, 95% uncertainty coverage intervals [4]. Despite the evident large uncertainties, we see that it is very probable that few patients will be prognosed with probabilities around 30% or larger than 90%. We also see that the full population is likely to be grouped into three or four "conversion-probability clusters", as evident from the peaks. Note that there are correlations, which cannot be shown in this kind of plot, between the fractions at different bins.

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) will be prescribed – assuming that the full population has on average the utility matrix of Olivia, table 4. We find these coverage intervals with a 90% probability that the true future proportion will be within each:

$$\alpha: 3\%\text{--}8\% \qquad \beta: 1\%\text{--}13\% \qquad \gamma: 22\%\text{--}59\% \qquad \delta: 34\%\text{--}63\% \tag{5}$$

Again, despite the obvious large uncertainties, we can be quite sure that treatments γ and δ will be prescribed ten times or more often than α and β. This kind of semi-quantitative forecast can be very useful for resource and production planning.

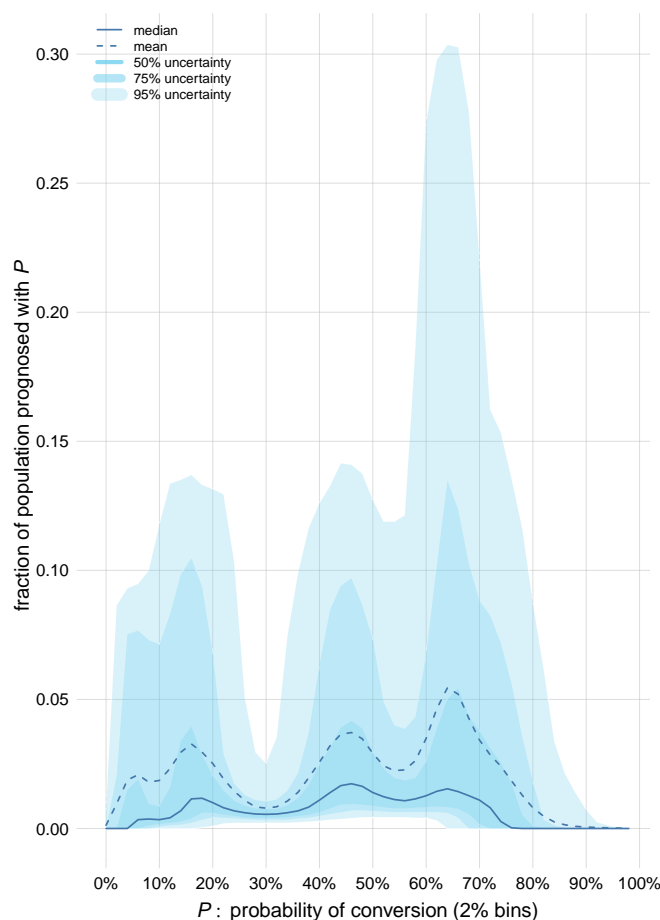[4] for terminology see JCGM (Joint Committee for Guides in Metrology), 2008, C.2.30

Figure 3. Possible distribution of prognostic probabilities of conversion in the full patient population

4.2 Sensitivity checks

The kind of forecasts illustrated in the previous subsection can also be applied to individual patients 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 Turing-Jaynes machine can give a probabilistic answer to this question.

The clinician uses the Turing-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:

$$\alpha: 2.97\%\text{--}2.98\% \qquad \beta: 4.78\%\text{--}4.79\% \qquad \gamma: 5.89\%\text{--}5.90\% \qquad \boldsymbol{\delta: 7.02\%\text{--}7.03\%} \tag{6}$$

(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.
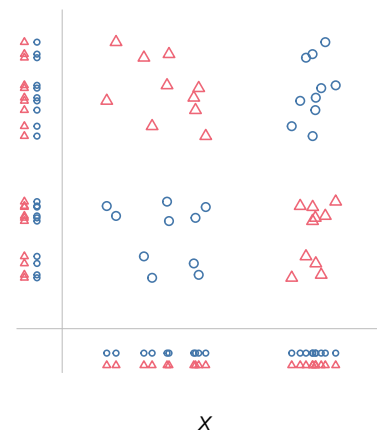
4.3 Predictor importance

🔧 [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 Turing-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 Turing-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 Turing-Jaynes machine achieves this maximal power*. Thus, if the Turing-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 Turing-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; Good and Toulmin, 1968; Kelly, 1956; Kullback, 1978) and international standards (ISO (International Organization for Standardization), 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. 4, 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 § 6 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 (Rye et al., 2022, contrast this with).

The Turing-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
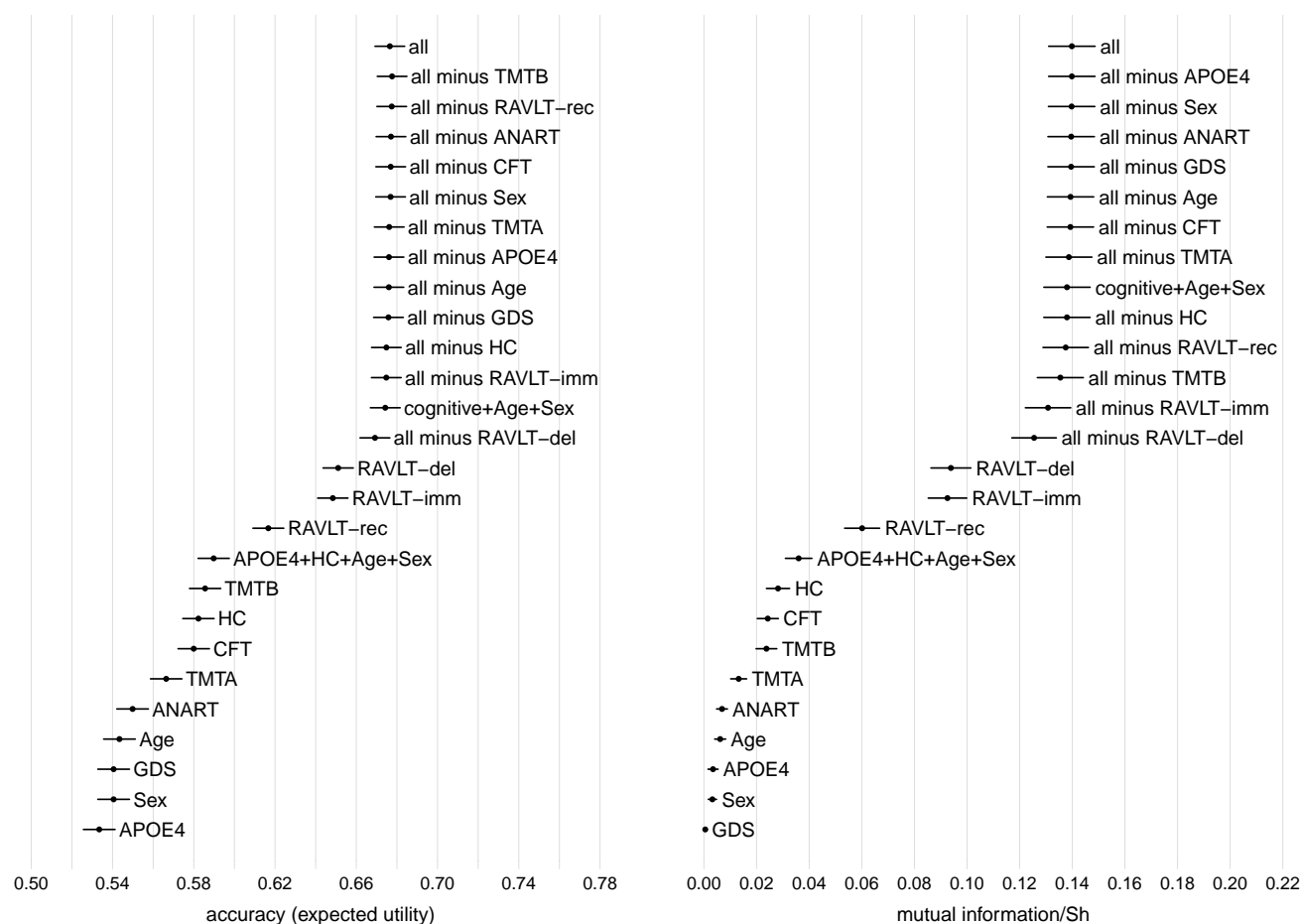
Figure 4. Expected accuracy (for the next new patient) and mutual information of several sets of predictors for the prognosis of conversion to Alzheimer's Disease. Both graphs have been vertically ordered according to increasing values; the two rankings agree within the respective uncertainties. The 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).

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
HC/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-im	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

	¬AD AD				
treatment α		$\begin{bmatrix} 10 & 0 \\ 9 & 3 \\ 8 & 5 \\ 0 & 10 \end{bmatrix}$	$\begin{bmatrix} 10 & 0 \\ 9 & 3 \\ 8 & 5 \\ 0 & 10 \end{bmatrix}$	$\begin{bmatrix} 10 & 0 \\ 8 & 3 \\ 7 & 5 \\ 0 & 10 \end{bmatrix}$	$\begin{bmatrix} 10 & 0 \\ 9 & 3 \\ 8 & 5 \\ 0 & 10 \end{bmatrix}$
treatment β					
treatment γ					
treatment δ					

Outputs of Turing-Jaynes machine				
p(AD | predictors)	0.302	0.302	0.302	0.703
p(predictors | AD)/10^{-12}	8.97	8.97	8.97	1.14
p(predictors | ¬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.

5 SUMMARY OF THE CASE STUDY

Summary in table 6.

6 DISCUSSION

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 § 2, p. 11, 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

✏ Difficulty in assessing and quantifying additional info: Just ignoring it is not a solution and is unethical

✏ 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?

6.1 Range of application of Turing-Jaynes machines

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

As mentioned in § 1 and explained in appendix A.1, the fact that the Turing-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 Turing-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 § 2, the essentially sole assumption underlying the Turing-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 Turing-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 available in the dataset, the Turing-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 Mathematical details about the Turing-Jaynes machine

As discussed in § 1, the Turing-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 Turing-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 Turing-Jaynes machine does a full *space exploration* and *averaging*, as explained in § 1. Inference and generalization in fact essentially rely on averaging operations in problems such as the present one (de Finetti's 1930 theorem; de Finetti, 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 5. ✏ Examples of a-priori probable candidates of frequency distribution for a variate such as RAVLT-del or RAVLT-rec

This difference explains why the Turing-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.2 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 § 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

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)

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

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

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`

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`

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. (1979). A. M. Turing's statistical work in World War II. *Biometrika* 66, 393–396. DOI:`10.1093/biomet/66.2.393`

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`

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

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 (Joint Committee for Guides in Metrology) (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

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. and Lusted, L. B. (1959). 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`

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

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

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

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

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

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

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

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:`10.1002/0470033312`

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

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

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`

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

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

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

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

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

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

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`

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`

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`

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