frontiers

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

Patients with Mild Cognitive Impairment have an increased risk of a trajectory toward Alzheimer's Disease. Early identification of patients with a high risk of Alzheimer's Disease is essential to provide treatment before the disease is well-established in the brain. great importance to study how well different kinds of predictors allow us to prognose a trajectory from Mild Cognitive Impairment towards Alzheimer's Disease in an individual patient.

But more is needed for a personalized approach to prognosis, prevention, and treatment, than just the obvious requirement that prognoses be as best as they can be for each patient. Several situational elements that can be different from patient to patient must be accounted for:

- 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. The main result underlying statistical decision theory is that decision-making *must* take this particular mathematical form in order to be optimal and logically consistent.

The present work investigates the prognostic power of a set of neuropsychological and Magnetic Resonance Imaging examinations, demographic data, and genetic information about Apolipoprotein-E4The present work investigates the prognostic power of a set of neuropsychological and Magnetic Resonance Imaging examinations, demographic data, and genetic information about Apolipoprotein-E4 (APOE) status, for the prediction of 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.

(APOE) status, for the prediction of 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.

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

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 some clinical analyses and cognitive tests. Today they received the results of their analyses. From these results, available clinical statistical data, and other relevant information, their clinician will assess their risk of developing Alzheimer. The clinician and each patient will then decide among a set of possible preventive treatments.

Besides this shared condition and worry, 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:

- 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 with a higher rate of conversion, and from a family with a heavy history of Alzheimer's Disease, unlike Olivia. Because of this 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 clinical results. In particular, Olivia does not have the risky Apolipoprotein-E4 (APOE4) allele (Liu et al., 2013) whereas Curtis has, and Olivia is more than 10 years older than Curtis. But 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 as "difference in auxiliary information" (Olivia and Ariel), "difference in preventive benefits" (Olivia and Bianca), "difference in clinical predictors" (Olivia and Curtis). Table 1 reports the clinical results and demographic data common to Olivia, Ariel, Bianca, as well as those of Curtis 🔧 need to explain the variates and refer to (Rye et al., 2022). The figure on its side summarizes the similarity and differences between Olivia and the other three patients.

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

	Olivia	Ariel	Bianca	Curtis
Age	75.4	75.4	75.4	63.8
Sex	F	F	F	M
$\mathrm{HC}/10^{-3}$	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

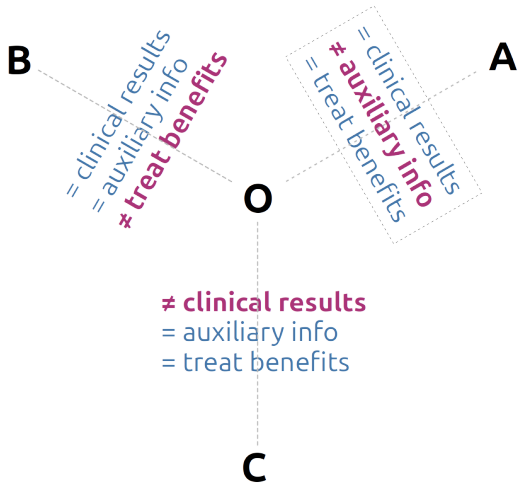


Table 1. Clinical results & demographic data

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 such as those listed in table 1. The solution method integrates available clinical statistical data with each new patient's unique combination of clinical results, auxiliary information, and treatment benefits.

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 these principles 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, special modelling assumptions, and limitations in its informational output; not in the sense of being impracticable. In the present work we indeed show that for some kinds of dataset such ideal machine-learning algorithm is a reality. 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 § ***.

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 table 2.

To be redone

Table 2. Inferential and decision-making steps. Steps in **boldface** represent patient-dependent, personalized steps that cannot be obtained from the learning dataset

1. Infer the full-population frequencies of predictors and predictand with the Turing-Jaynes machine, using available datasets.
2. **Assess in respect of which variates the present patient can be considered as belonging to the same population underlying the learning dataset.**
3. **Assess the prior probability of the predictand for the present patient.** This step allows us (a) to consider additional clinical information outside of the dataset's variates, and available for the present patient only; (b) to correct for mismatches between the dataset's underlying population and the patient's one.
4. Calculate the *likelihood* of the predictand for a specific patient, given the patient's predictor values. Combine this likelihood with the prior from step 3, to obtain the final probability for the predictand, for the present patient.
5. **Assess the clinical courses of action available for the present patient, together their benefits and costs.** This step is fundamentally patient-dependent and is the one open to most variability from patient to patient.
6. Choose the course of action having maximal expected benefit for the present patient, given the benefits assessed in step 5 and the final probability assessed in step 4.

These steps will be explained and illustrated in the next three sections, presented in chronological order as the clinician would apply them. They could also be presented in reverse order, more suited to their logical dependence, because the procedure in each one is 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

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 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 ✎ ....

✎ 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.

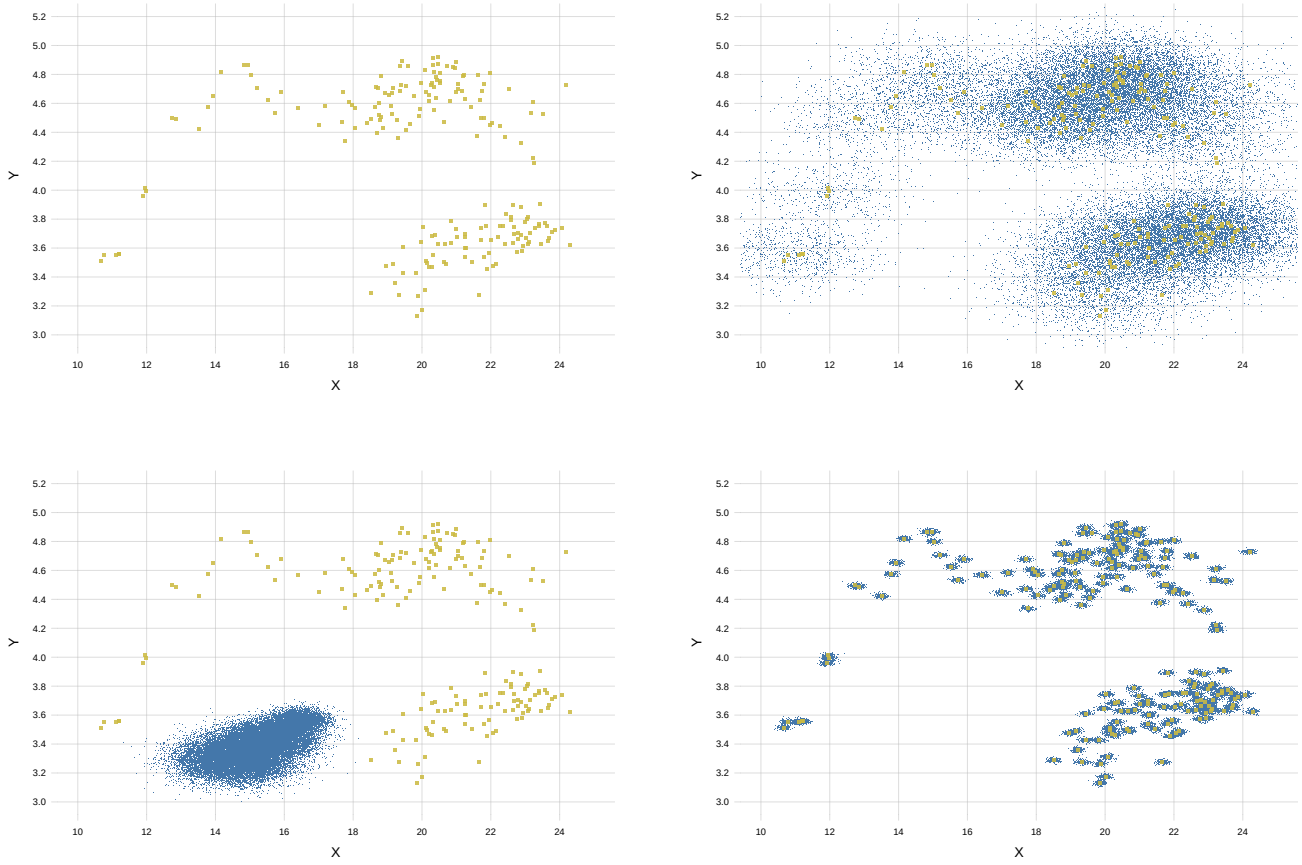


Figure 1. Upper-left: Sample data. Upper-right: candidate frequency distribution that fits the data and does not look unnatural. Lower-left: candidate distribution that might look natural but doesn't fit the sample data. Lower-right: candidate distribution that fits the data very well but looks unnatural.

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 function of $\{X, Y\}$ if and only if their mutual information [2] equals $1\,\mathrm{Sh}$.

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

Figure 2. Examples of a-priori probable candidates of frequency distribution for a variate such as AVDEL30MIN_neuro or AVDELTOT_neuro

- *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.

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 § **??**. It is possible to transform this score into a proper probability (Dyrland et al., 2022), but this possibility does not solve the

[3] 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.

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.

1.1 Application to case study

In our case study, the Turing-Jaynes machine yields the following probabilities of main importance for the application in the next sections:

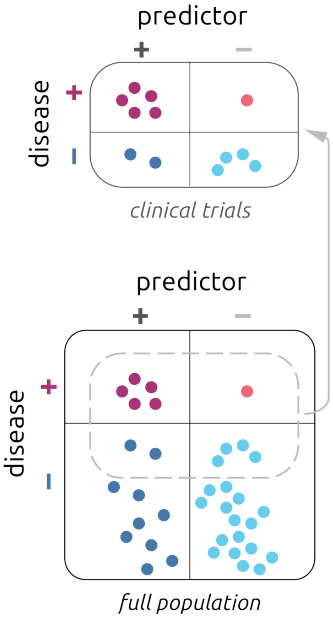
	Olivia	Ariel	Bianca	Curtis
$\mathrm{p}(\text{conversion to AD} \mid \text{predictors})$	0.302	0.302	0.302	0.691
$\mathrm{p}(\text{predictors} \mid \text{conversion to AD})/10^{-12}$	8.97	8.97	8.97	1.54
$\mathrm{p}(\text{predictors} \mid \text{no conversion to AD})/10^{-12}$	18.6	18.6	18.6	0.50

Table 3. Probabilities computed by the Turing-Jaynes machine

As an illustration of the additional information provided by the Turing-Jaynes machine, fig. **??** 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 "error" on these probabilities. There is an error on the probabilities, caused by finite computational precision, but it affects at most the third significant digit of the reported probabilities (all relative numerical errors are below 0.8%, Curtis's two likelihoods being an exception at 2%).

"

2 POPULATION ASSESSMENT AND PRIOR PROBABILITY



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:

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

In our example we find

$$\begin{aligned} \mathrm{p}(\mathsf{disease+} \mid \mathsf{predictor+}) &= \frac{\mathrm{p}(\mathsf{predictor+} \mid \mathsf{disease+}, \mathsf{trials}) \cdot \mathrm{p}(\mathsf{disease+} \mid \mathsf{population})}{\left[\begin{array}{r} \mathrm{p}(\mathsf{predictor+} \mid \mathsf{disease+}, \mathsf{trials}) \cdot \mathrm{p}(\mathsf{disease+} \mid \mathsf{population}) + \\ \mathrm{p}(\mathsf{predictor+} \mid \mathsf{disease–}, \mathsf{trials}) \cdot \mathrm{p}(\mathsf{disease–} \mid \mathsf{population}) \end{array}\right]} \\ &\approx \frac{0.833 \cdot 0.167}{0.833 \cdot 0.167 + 0.333 \cdot 0.833} = 0.33 \end{aligned} \tag{2}$$

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 can be illustrated with a toy example. 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 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. 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 $\mathrm{p}(\mathsf{predictand} \mid \mathsf{predictors}, \mathsf{dataset})$ 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.

2.1 Application to case 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 3 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 3 and the prior probability of 0.65.

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

	Olivia	Ariel	Bianca	Curtis
prior probability $\mathrm{p}(\text{conversion to AD} \mid \text{aux info})$	0.463	0.65	0.463	0.463
final probability $\mathrm{p}(\text{conversion to AD} \mid \text{predictors, dataset, aux info})$	0.302	0.47	0.302	0.691

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

3 TREATMENT & BENEFIT ASSESSMENTS, AND DECISION

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 our method, 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 the differences between patients matter and vary 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 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).

The quantification of utilities is not within the scope of the present work. The clinical books cited above offer several guidelines and rules , but the present approach

(Lindley, 1982)

4 DISCUSSION

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?

— Luca, old pieces of text —

Personalized diagnosis, prognosis, treatment, and prevention strategies must make allowance for several fundamental differences among patients:

- the *kinds* of clinical data and evidence available for diagnosis or prognosis can be different;
- the *values* of the same kind of clinical data and evidence can be different;
- the kinds of treatment or prevention options can be different;
- the advantages and disadvantages, benefits and costs of the same kinds of treatment or prevention can be different;
- finally, the evaluation of the clinician – which often relies on too subtle clues (family history, regional history, case experience) to be considered as measurable data – can be different.

Is there really a methodological framework that can take all these differences into account? Yes, there is, and Medicine has the distinction of having been one of the first fields to adopt it (Ledley and Lusted, 1959): Statistical Decision Theory. Its application in Medicine is explained and exemplified in several, brilliant, old and new textbooks (Weinstein and Fineberg, 1980; Sox et al., 2013; Hunink et al., 2014). This theory has mathematical and logical foundations and its principles constitute indeed the foundations for the definition and realization of Artificial Intelligence (Russell and Norvig, 2022)

The basics of clinical decision making ..basics: each piece of evidence contributes with a likelihood or odds; they combine together and together with the clinician's pre-data evaluation. Then they are combined with the different benefits/costs of treatments or prevention strategies to find the optimal one. Decision trees can be necessary (but don't change this framework). Costs & benefits are evaluated by clinician & patient together.

$$\overbrace{\mathrm{p}(\text{health condition} \mid \text{results of all tests},\ \text{prior info})}^{\textit{post-test probability}} \propto \overbrace{\mathrm{p}(\text{health condition} \mid \text{prior info})}^{\textit{pre-test probability by clinician}} \times \textit{likelihoods of tests} \begin{cases} \mathrm{p}(\text{result of 1st test} \mid \text{health condition},\ \text{prior info}) \times \\ \mathrm{p}(\text{result of 2nd test} \mid \text{health condition},\ \text{prior info}) \times \\ \dots \end{cases} \tag{3}$$

5 ARTICLE TYPES

For requirements for a specific article type please refer to the Article Types on any Frontiers journal page. Please also refer to Author Guidelines for further information on how to organize your manuscript in the required sections or their equivalents for your field

6 MANUSCRIPT FORMATTING

6.1 Heading Levels

6.2 Level 2

6.2.1 Level 3

6.2.1.1 Level 4

6.2.1.1.1 Level 5

6.3 Equations

Equations should be inserted in editable format from the equation editor.

$$\sum x + y = Z \tag{4}$$

6.4 Figures

Frontiers requires figures to be submitted individually, in the same order as they are referred to in the manuscript. Figures will then be automatically embedded at the bottom of the submitted manuscript. Kindly ensure that each table and figure is mentioned in the text and in numerical order. Figures must be of sufficient resolution for publication see here for examples and minimum requirements. Figures which are not according to the guidelines will cause substantial delay during the production process. Please see here for full figure guidelines. Cite figures with subfigures as figure **??** and **??**.

6.4.1 Permission to Reuse and Copyright

Figures, tables, and images will be published under a Creative Commons CC-BY licence and permission must be obtained for use of copyrighted material from other sources (including re-published/adapted/modified/partial figures and images from the internet). It is the responsibility of the authors to acquire the licenses, to follow any citation instructions requested by third-party rights holders, and cover any supplementary charges.

6.5 Tables

Tables should be inserted at the end of the manuscript. Please build your table directly in LaTeX.Tables provided as jpeg/tiff files will not be accepted. Please note that very large tables (covering several pages) cannot be included in the final PDF for reasons of space. These tables will be published as Supplementary Material on the online article page at the time of acceptance. The author will be notified during the typesetting of the final article if this is the case.

7 NOMENCLATURE

7.1 Resource Identification Initiative

To take part in the Resource Identification Initiative, please use the corresponding catalog number and RRID in your current manuscript. For more information about the project and for steps on how to search for an RRID, please click here.

7.2 Life Science Identifiers

Life Science Identifiers (LSIDs) for ZOOBANK registered names or nomenclatural acts should be listed in the manuscript before the keywords. For more information on LSIDs please see Inclusion of Zoological Nomenclature section of the guidelines.

8 ADDITIONAL REQUIREMENTS

For additional requirements for specific article types and further information please refer to Author Guidelines.

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

This is a short text to acknowledge the contributions of specific colleagues, institutions, or agencies that aided the efforts of the authors.

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

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*. Wiley series in probability and mathematical statistics (New York: Wiley), repr. edn. DOI:10.1002/9780470316870. First publ. 1994

Cifarelli, D. M. and Regazzini, E. (1979). Considerazioni generali sull'impostazione bayesiana di problemi non parametrici. Le medie associative nel contesto del processo aleatorio di Dirichlet. *Riv. mat. sci. econ. soc.* 2, 39–52, 95–111

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 (Cifarelli and Regazzini, 1979). 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)

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

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

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

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

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

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. (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

[Dataset] 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

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

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

Raiffa, H. and Schlaifer, R. (2000). *Applied Statistical Decision Theory*. Wiley Classics Library (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*. The Econometric and Tinbergen Institutes lectures (Princeton: Princeton University Press). DOI:10.1515/9781400850303

Russell, S. J. and Norvig, P. (2022). *Artificial Intelligence: A Modern Approach*. Pearson series in artificial intelligence (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

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

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

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