2017-10-01

Integrating and testing natural frequencies, naive Bayes, and fast-and-frugal trees.

Woike, Jan Kristian

10.1037/dec0000086
Decision
American Psychological Association (APA)

Integrating and Testing Natural Frequencies, Naïve Bayes, and Fast-and-Frugal Trees

Jan K. Woike,* Ulrich Hoffrage, Laura Martignon*****

* Max Planck Institute for Human Development, Berlin, Germany

** University of Lausanne, Switzerland

*** Ludwigsburg University of Education, Ludwigsburg, Germany

* woike@mpib-berlin.mpg.de

** ulrich.hoffrage@unil.ch

*** martignon@ph-ludwigsburg.de

Corresponding author: Laura Martignon

Version: 22.05.2017

Abstract

This article relates natural frequency representations of cue-criterion relationships to fast-and-frugal heuristics for inferences based on multiple cues. In the conceptual part of this work, three approaches to classification are compared to one another: The first uses a natural Bayesian classification scheme, based on profile memorization and natural frequencies. The second is based on Naïve Bayes, a heuristic that assumes conditional independence between cues (given the criterion). The third approach is to construct fast-and-frugal classification trees, which can be conceived as pruned versions of diagnostic natural frequency trees. Fast-and-frugal trees can be described as lexicographic classifiers but can also be related to another fundamental class of models, namely linear models. Linear classifiers with fixed thresholds and noncompensatory weights coincide with fast-and-frugal trees—not as processes but in their output. Various heuristic principles for tree construction are proposed. In the second, empirical part of this article, the classification performance of the three approaches when making inferences under uncertainty (i.e., out of sample) is evaluated in 11 medical data sets in terms of Receiver Operating Characteristics (ROC) diagrams and predictive accuracy. Results show that the two heuristic approaches, Naïve Bayes and fast-and-frugal trees, generally outperform the model that is normative when fitting known data, namely classification based on natural frequencies (or, equivalently, profile memorization). The success of fast-and-frugal trees is grounded in their ecological rationality: their construction principles can exploit the structure of information in the data sets. Finally, implications, applications, limitations, and possible extensions of this work are discussed.

Keywords: Fast-and-frugal trees, decision trees, heuristics, classification, categorization, Bayesian inferences, computer simulation, lexicographic strategies, probability

1. Introduction

Probability theory emerged during the Enlightenment as a framework for making reasonable inferences under uncertainty (Daston, 1995). According to Pierre-Simon Laplace (1749–1827), probability formalized the intuitions of *l'homme eclairé*, the enlightened man: "… the theory of probabilities is at bottom only common sense reduced to calculus; it makes us appreciate with exactitude that which exact minds feel by a sort of instinct without being able ofttimes to give a reason for it" (Laplace, 1814/1951, p. 196). Laplace began assembling the pieces of a puzzle, namely the many instruments and concepts introduced by multiple mathematicians in the century and a half since the correspondence between Blaise Pascal and Pierre de Fermat in 1654, who are considered to be the founders of probability theory (Ore, 1960). These instruments and concepts included the principles of what was later called Bayesian inference for inverting conditional probabilities or inverse probability. Indeed, Bayes' formula was among the principles of probability that he listed.

Yet Laplace's approach to probabilities has been criticized as circular: for "equipossible" outcomes, the probability of an event was defined as the number of successful outcomes divided by the number of possible outcomes. The circularity consisted in the "equipossible" requirement, which was explained in terms of the physical conditions of the random generator (e.g., an unloaded die) used in the experiment. Another century had to pass before the theory of probability with all its instruments could be embedded in the edifice of formal mathematics. This happened when Kolmogorov (1933) published the *Foundations of the Theory of Probability*, thereby laying the modern axiomatic foundations of probability theory. The price of this celebrated achievement, with its foundational rigor, was a certain loss of natural probabilistic intuition. From that moment on, as Leo Breiman (1968) later commented, probability theory was condemned to having a right and a left hand—the right hand being the measure-theoretical approach that guarantees mathematical rigor, and the left hand meaning 'intuitive probabilistic thinking.' The integration of these two aspects is not easy and sometimes makes modern probability theory both elusive and cumbersome. When probability is treated as a mathematical enterprise based on countably additive measures on sigma-algebras of sets, the intuitive and natural understanding of it is lost.

To illustrate, consider a physician who wants to infer the presence of a particular disease given a set of observations. The Kolmogorov approach to this task uses probabilities: The physician starts with a *prior* probability $P(D)$ that disease D is present. She then

observes evidence E in the form of a set of cues (observations, symptoms, tests) and assesses the probability $P(E \mid D)$ that the evidence will occur if the disease is present and the probability $P(E \mid \overline{D})$ that the evidence will occur if the disease is not present. Based on these probabilities, the physician then uses *Bayes theorem* to find the probability $P(D \mid E)$, also called the *posterior* probability of the disease, given the evidence:

$$P(D \mid E) = \frac{P(E \mid D)P(D)}{P(E \mid D)P(D) + P(E \mid \overline{D})P(\overline{D})} \quad (1)$$

There is a flurry of literature demonstrating how difficult it is to use Bayesian reasoning in this framework—even for experts and in the simple case of a single symptom or test (see, e.g., Eddy, 1982; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000, probability condition). The difficulty appears to reside in the format of information representation in (1).

Simply changing the information format to *natural frequencies* has been shown to facilitate physicians' information processing (Gigerenzer & Hoffrage, 1995). Now the doctor thinks of an absolute (and possibly fictitious) number of people with and without the disease, and then subdivides this population into subclasses (having or not having the disease), which are again subdivided according to the test results (positive and negative). Here, the proportions of these subclasses correspond to the original probabilities involved. The computation to be made at the end is simple arithmetic: the number of true positives divided by the sum of the number of true positives and the number of false positives. As using natural frequencies instead of probabilities facilitates the computation, it is not surprising that representing statistical information in terms of natural frequencies rather than Kolmogorov probabilities improves performance in Bayesian inference tasks (Gigerenzer & Hoffrage, 1995). This beneficial effect of natural frequencies has been demonstrated in a variety of applied domains, such as medicine, law, and education (Gigerenzer, 2002; Hoffrage & Gigerenzer, 1998; Hoffrage et al., 2000).

Most importantly for the present investigation, Hoffrage, Krauss, Martignon, and Gigerenzer (2015) have shown empirically that the beneficial effects of natural frequencies extend to more complex situations than those involving just one cue (symptom, test, or feature). Participants who were given the prior probability of a disease and statistical information about two independent tests (specifically, their sensitivities and false-alarm rates) in terms of probabilities struggled to infer the probability of the disease being present if both

tests were positive. In contrast, when they received the same (i.e., mathematically equivalent) information in terms of natural frequencies, about three quarters of the participants derived the correct (i.e., Bayesian) solution. The authors concluded that natural frequencies also facilitate Bayesian performance in tasks involving two or three cues—but they acknowledged that there are obvious limits to the number of cues people can deal with by means of natural frequencies.

Two questions arise when Bayesian reasoning with a single cue is extended to multiple cues. The first is descriptive and empirical: What is the maximum number of cues the human mind can handle with natural frequency representations? As the number of cues grows, the size of the natural frequency tree explodes. At some point, Bayesian performance can be expected to reach its limits. The limits depend on various factors, such as whether the decisions need to be made *based on experience* or *based on description* (Hertwig, Barron, Weber, & Erev, 2004). The second question is prescriptive: How can and should one make inferences under uncertainty when the number of cues exceeds what people can handle with natural frequency representations? Note that this question is relevant not only because of memory constraints (for inferences based on experience) or because handling huge numbers of cue–value combinations is unfeasible (for inferences based on description). There is another reason one may not want to use natural frequency representations for tasks involving multiple cues, namely the eventual "brittleness"—as opposed to robustness—of the resulting tree: As the tree grows, the number of end nodes becomes ever larger, while the number of cases per end node becomes ever smaller. A decision maker may not even have encountered some of the cue–value combinations, or, for some specific combinations of cue values, she may have encountered so few cases that she finds it questionable to generalize to new cases.

One way to tackle the problems associated with large numbers of cues—limited memory, unfeasibility, and lack of robustness—is to adopt heuristic procedures. As it turns out, various heuristics can be used to classify objects based on multiple cues and to estimate the probabilities of those classifications being correct. The traditional Bayesian paradigm has its "own" heuristic: Naïve Bayes. Naïve Bayes simplifies reality—which is why it is considered a heuristic—by assuming conditional independence of cues given the category in question. Among the Bayesian networks that make use of all cues, Naïve Bayes is the simplest. Nevertheless, it requires much computational effort: prior distributions are updated

by means of Bayes' formula which, if the number of cues is large, involves a large number of multiplications.

A much simpler alternative is offered by fast-and-frugal heuristics (Gigerenzer, Todd, & the ABC Research Group, 1999), specifically, fast-and-frugal trees (Martignon, Katsikopoulos, & Woike, 2008, 2012; Martignon, Vitouch, Takezawa, & Forster, 2003). The fast-and-frugal trees we analyze in this article can be conceptualized as pruned and simplified natural frequency trees. Like other heuristics proposed and studied in the context of the simple heuristics program (Gigerenzer, Hertwig, & Pachur, 2011), fast-and-frugal trees truncate information search and typically make a classification based on just a few of the available cues. They thus reduce memory load and can be set up and executed by the unaided mind, requiring, at most, paper and pencil.

What are the costs of using heuristics, be they Naïve Bayes or fast-and-frugal trees? Precisely this is one of the questions addressed by this article. On the one hand, we aim to illustrate how the two heuristics can be integrated within a naturally Bayesian approach based on natural frequencies; on the other hand, we compare, using computer simulations in settings with multiple cues, the classification performance of three strategies: (1) the natural Bayesian relying on profile memorization, (2) Naïve Bayes, and (3) various fast-and-frugal trees.

The article is structured as follows: We first stress the importance of sampling, both in order to illustrate natural frequencies and to justify our approach of testing the strategies' performance out of sample. We then extend Bayesian inferences with natural frequency representations from situations with one cue to situations with multiple cues. Subsequently, we show how the strategies listed above can be constructed and how they function in the context of classifications with multiple cues, focusing on how they can be integrated into a common framework. In the Methods section, we explain how we set up the simulations: which data sets were used, how the strategies were implemented, and how their performance was measured. We then report the results and, finally, we discuss applications, limitations, and avenues for further research.

2. Natural Sampling and Natural Frequencies: Three Cases

Before we consider classifications with multiple cues, let us examine the relationship between probabilities and natural frequencies more thoroughly than has been done in the past.

These two concepts can be related to the difference between population and sample, which is crucial when it comes to evaluating the performance of strategies, as shown below. We distinguish three possible cases in this context. The first does not involve sampling (simply because the entire population is at hand), the second involves a fictitious sample from an infinite population that reflects the probabilities of the population, and the third involves natural sampling from a population—and hence sampling error. Let us look at these cases in more detail.

In Case 1, inferences are made within a given, finite sample that may also be labelled the reference population. The natural frequencies are the tallies that result when measuring the state of the criterion (here, disease) and of the cue for each object (here, each patient) in the population. These frequencies can be arranged in a tree that first splits the number of patients into those with disease and those without disease. In a second layer, each of these two numbers can be split into the number of patients who test positive and the number who test negative. Because the test result is causally affected by the state of the criterion, this tree is often referred to as a causal tree (Martignon et al., 2003; Waldmann & Martignon, 1998). Note that not only the four joint frequencies at the lowest layer are natural frequencies, but also the margins for disease and no disease are natural frequencies, simply because they can be observed in this population. To infer the state of the criterion for a patient for whom only the test result is available, this tree needs to be inverted, that is, the number of patients must first be split into the numbers of patients who test positive and those who test negative. Each of these two nodes is then split into those who have the disease and those who do not. Again, all numbers in this tree are natural frequencies. A tree with numbers arranged in this way is called a diagnostic tree (Martignon et al., 2003). Note that making inferences with this diagnostic tree amounts to making inferences for known patients. If, for a given patient drawn randomly from this reference population, only the cue information is provided and the criterion needs to be inferred, the normative approach would be to ask "Which is the best inference that can be made for all patients that share exactly this cue information?" The answer can be read off from a diagnostic natural frequency tree. It is the same answer that can be obtained from applying Bayes' theorem after the natural frequencies have been translated into probabilities.

Case 2 does not describe the relationship between variables in a finite sample, but in an infinite population of random results, such as the population of tosses of a pair of dice

(e.g., a red and a blue die, with the criterion being the sum of both dice and the cue being the number of the red die). Whereas, in Case 1, probabilities are Laplace probabilities that correspond to observed frequencies, in Case 2, they are derived from assumptions (e.g., about the physical propensities of objects, such as symmetry or equal distribution of mass within the dice). These probabilities can then be translated into a frequency tree that has the same structure as a natural frequency tree. In this case, however, the top node is no longer a fixed and finite population for which all natural frequencies below this node have been observed empirically. Rather, it is a fictitious number that forms the starting point used to express the expected relationships between all numbers in the tree. In other words, the frequencies in this tree are not natural in the sense that they have been observed empirically; they are basically "expected frequencies" (Spiegelhalter & Gage, 2014), that is, the frequencies one expects in the long run. Predictions made for new chance events would *not* be predictions out of sample—simply because the new chance events can be seen as being sampled from the same, infinite, population. Hence, they would *not* be predictions under uncertainty, with unknown probabilities, but predictions based on known probabilities (or, equivalently, known expected frequencies). The normative solution could therefore be obtained from applying Bayes' theorem (or, equivalently, using the tree with fictitious frequencies). Note that when Gigerenzer and Hoffrage (1995) originally explained the concept of natural frequencies (which, in that publication, were still called frequency formats), they asked readers to imagine "an old, experienced physician in an illiterate society" (p. 686) who observed a fictitious number of women. Running this example with 100 women allowed the authors to translate the probabilities into expected frequencies that were whole numbers. With a fictitious sample of 101 women this would not have been possible. This shows that not only the 100 women in Gigerenzer and Hoffrage's (1995) tree were fictitious, but all other frequencies in this tree were also fictitious, in the sense that they reflected expected frequencies and not observed frequencies. In fact, their example was the present Case 2.

Drawing a random sample from a fixed population, also called *natural sampling*, constitutes Case 3. Such sampling leads to natural frequencies. Repeating this process leads to multiple samples, multiple trees (one for each sample), and hence to variance across samples. When a real sample is drawn from a population, it is extremely unlikely that the proportions in the sample are identical to those in the population—but note that this identity is enforced in Case 2. In Case 3, inferences for new, as-yet-unseen objects are inferences under uncertainty—simply because the probabilities in the populations are not known. The

probabilities derived from the natural frequencies observed so far may be good estimates for the population probabilities, but estimates should not be confused with known probabilities. As a consequence, Bayes' theorem (or the corresponding inference based on the natural frequencies) may not necessarily provide the best response for new objects. Note that this third case, in which inferences need to be made under uncertainty, is the most relevant one for daily decision making—and it is precisely here that heuristics can potentially outperform Bayes' theorem (which is, as we have seen, best and normative for Case 1 and 2).

3. Inferences with Multiple Cues: Approaches to Classification

Classification—or, synonymously, categorization—is fundamental for cognition: Humans combine features or cues into concepts and decompose concepts into features or cues. Furthermore, classification is essential for decision making. Most decisions depend on how people categorize objects, situations, or events. Let us consider the case of a physician who makes decisions on how to treat patients. These decisions hinge on the categorization of those patients as having or not having specified diseases. For each patient, the physician is presented with a set of cues (e.g., direct observations, reported symptoms, or test results) on which to base her diagnosis. First, we assume the physician wants to infer the answer to a simple yes/no question: Does a given patient have the disease under consideration? Second, we assume that the input to make this inference is a set of binary cues. A binary cue is either in its "high" state (cue value is 1, coded such that it indicates the disease being present) or in its "low" state (cue value is 0). The combination of all cue values for a given object (here, patient) will henceforth be referred to as a *cue profile*. Finally, we assume that the physician has to report her inference not as a probability but as a simple "disease present" (denoted as 1) or "disease absent" (denoted as 0), with no opportunity to hedge the result. The relevant question is whether the physician's inference is correct or incorrect. In what follows, we describe different approaches to this classification task, from strictly normative to heuristic.

3.1 Bayes Theorem and Profile Memorization

How does Bayes' theorem stated for probabilities and for situations with one cue extend to situations in which the evidence consists of several pieces of information? As we have mentioned, natural frequencies are also applicable to tasks with two or three cues (Hoffrage et al., 2015). Figure 1A illustrates a causal natural frequency tree that displays, for a finite population of 10,000 women and depending on whether or not they have breast cancer, the numbers of women for each possible combination of results when two tests are conducted,

namely a mammogram (M) and an ultrasound (U). Although the round number 10,000 would seem to be fictitious (and hence Case 2 discussed above), we will describe this tree as if all numbers were not expected frequencies, but observed frequencies. Figure 1B displays the same information in an inverted (i.e., diagnostic) tree, in which the set of 10,000 women is first split according to the mammography result, then according to the ultrasound result, and in which the third layer displays, for each of the eight cue profiles, how many of the patients with the same profile have the disease and how many not. For instance, to infer whether a woman with a positive mammogram and a negative ultrasound has breast cancer, the natural Bayesian would look up, in internal or external memory, how many women share this profile (n = 916, Figure 1B) and would find that four have breast cancer. This procedure is based on the full information contained in the natural frequency tree and is termed profile memorization. Of course, the resulting proportion (4/916), or, equivalently, the posterior probability that results from Equation 1, is not yet a classification. However, by comparing it to a critical threshold, it can easily be turned into one. When fitting known data, no other procedure is able to outperform the profile memorization method of the natural Bayesian.

Figure 1C shows another diagnostic tree—one in which the ultrasound is displayed above the mammography. Panels B and C thus differ with respect to the order of the cues. Note that for n cues, there are $n!$ possible diagnostic trees, each containing $n + 1$ layers (excluding the top node but including the final end nodes with the criterion), and a total of 2^n distinct cue profiles. For 10 cues, for instance, there would be 3,628,800 different diagnostic trees, each with 1,024 distinct cue profiles. As these numbers show, natural frequency representations at some point reach their limits. First, frequencies become hard to memorize. As Massaro (1998) put it, “a frequency algorithm will not work” because “it might not be reasonable to assume that people can maintain exemplars of all possible symptom configurations” (p. 178). Second, even if memory is not a problem because tables or visual displays are at hand, these representations will be complex. Most importantly, however, many of the numbers in natural frequency trees (both causal and diagnostic) will be zero. This is not a problem when inferences are made within a set of known patients (Case 1, as discussed above), because here such objects do not exist. Likewise, it does not constitute a problem when the reference class is infinite (Case 2), because even the smallest probabilities (which are assumed to be known) can easily be represented in a frequency tree that starts from an arbitrarily chosen and huge, fictitious reference population. It is, however, an issue when inferences are made about new patients based on empirical samples (Case 3). What should be

inferred for a new patient if none of the patients observed so far have had the same cue profile?

<<<<<< Figure 1 >>>>>

3.2 Naïve Bayes

A full natural frequency tree with *n* cues is powerful enough to display all interdependencies among cues but it requires the physician to memorize how often the criterion is present in 2^n sets of patients with distinct cue profiles. A commonly applied simplification to the general problem used by Bayesian statisticians and Bayesian modelers is to assume that the evidence items (e.g., cues, tests, or symptoms) are independent conditional on presence or absence of the disease. This assumption marks the transition from the sophisticated, natural Bayesian (who uses profile memorization) to the naïve Bayesian (who seeks to simplify). The naïve Bayesian assesses, for each symptom E_k, only two probabilities: $P(E_k^H|D)$, the probability that the evidence is in its "high" state (E_k^H) given that the disease is present, and $P\left(E_k^H\middle|\overline{D}\right)$, the probability that the evidence is in its "high" state (E_k^H) given that the disease is absent. By the laws of probability, the evidence is in its "low" state (E_k^L) with probability $P(E_k^L|D) = 1 - P(E_k^H|D)$ if the disease is present and $P\left(E_k^L\middle|\overline{D}\right) = 1 - P\left(E_k^H\middle|\overline{D}\right)$=if the disease is absent. For multiple symptoms the evidence consists of the vector of symptom states $(E_1^{j_1}, \ldots, E_n^{j_n})$. With the conditional independence assumption, Equation (1) becomes:

$$P\left(D\middle|E_1^{j_1}, \ldots, E_n^{j_n}\right) = \frac{P(D)\prod_k P\left(E_k^{j_k}\middle|D\right)}{P(D)\prod_k P\left(E_k^{j_k}\middle|D\right) + P(\overline{D})\prod_k P\left(E_k^{j_k}\middle|\overline{D}\right)}, \qquad (2)$$

where j_k denotes either the "high" or "low" state of E_k. The required probabilities may be assessed intuitively by an expert or estimated from data. As stated above, their multiplication according to (2) implements the assumption that the pieces of evidence are conditionally independent. As in the case of profile memorization, the posterior probability obtained by (2) is not yet a classification, but comparing it to a threshold leads to one.

The Naïve Bayes model drastically reduces the number of probability assessments from 2^n (for the 2^n distinct cue profiles in the full tree) to $2n + 1$ (two likelihoods for each cue plus the base rate of the disease); for 10 cues, this represents a reduction from 1,024 to 21. Still, calculating Equation 2 is beyond the reach of intuitive judgment. It is, however,

manageable with paper and pencil—especially if natural (i.e., expected) frequencies are used rather than probabilities. Using the numbers in Figure 1 (Panels A, D, and E), for instance, the probability that a woman with a positive mammogram and a positive ultrasound has breast cancer is

$$\frac{\frac{100}{10000}\times\frac{80}{100}\times\frac{95}{100}}{\frac{100}{10000}\times\frac{80}{100}\times\frac{95}{100}+\frac{9900}{10000}\times\frac{950}{9900}\times\frac{396}{9900}}=\frac{2}{3}.$$

Posterior probabilities calculated with Equation 2 have been shown to be quite robust to generalization (Domingos & Pazzani, 1997). Naïve Bayes is often used as a Bayesian benchmark among simple models when analyzing small data sets (Martignon & Laskey, 1999).

3.3 Fast-and-Frugal Trees

Fast-and-frugal trees go one step further in terms of simplicity. Like Naïve Bayes, they take independence between cues (conditional on the criterion) for granted and they are constructed without checking whether this assumption is, in fact, justified. Unlike Naïve Bayes, however, they do not necessarily use and integrate all information. As shown in Equation 2, Naïve Bayes uses every cue. Fast-and-frugal trees, in contrast, typically consult only a subset of the available cues.

Fast-and-frugal trees were introduced and defined by Martignon et al. (2003; similar models had been previously used by Dhami and Ayton, 2001, Dhami and Harries, 2001, and Fischer, Steiner, Zucol, Berger, Martignon, et al., 2002). Fast-and-frugal trees can be described in terms of their building blocks. First, they have a search rule: They inspect cues in a specific order. Second, they have a stopping rule: Each cue has one value that leads to an exit node and hence to a classification, and another value that leads to consulting the next cue in the cue hierarchy (the exception being the last cue in the hierarchy, which has two exit nodes). Finally, they have a classification rule. They always classify the patient as having the disease if the value of the consulted cue is in its high state (henceforth also referred to as positive) and as not having the disease if this value is in its low state (henceforth also referred to as negative). Trees that classify patients as having the disease if the proportion of patients with the disease in the exit nodes exceeds a threshold are not fast-and-frugal trees and they are not considered here.

We now illustrate how fast-and-frugal trees relate to full natural frequency trees. Subsequently, we discuss the link between fast-and-frugal trees, lexicographic strategies, and linear strategies. At the end of this section, we explain how fast-and-frugal trees can be constructed.

Figure 2A displays the data of Green and Mehr (1997) in terms of natural frequencies, arranged in a diagnostic tree that starts with the cues and finishes with the criterion (occurrence of myocardial infarction) in the last layer. The tree has been constructed based on information about 89 patients, and its classification performance is determined for exactly those 89 patients (hence, Case 1 discussed above). Figure 2B also displays a diagnostic tree. The cues are arranged in the same order as in 2A, but some branches are cut—in fact, tree 2B is a pruned version of tree 2A. The tree in 2B has at least one exit at each level and therefore fulfills Martignon, Katsikopoulos, and Woike's (2012), definition of a fast-and-frugal (classification) tree. Specifically, in the first step, all 89 patients' electrocardiograms are checked for an elevated ST segment. The 33 for whom the answer is positive (ST+) are classified as high risk, without considering any further information. The remaining 56 patients are checked for chest pain as the main symptom. Among those 56, the 29 for whom chest pain is not the main symptom (CP–) are classified as low risk—again, without considering any further information. The remaining 27 patients are checked for the presence of any other symptom of myocardial infarction. Among those 27, the 17 for whom at least one other symptom is present (OS+) are classified as high risk, and the 10 remaining patients (OS–) are classified as low risk.

<<<<<< Figure 2 >>>>>

3.3.1 Comparing Profile Memorization with Fast-and-Frugal Trees

How does classification of new patients based on the full (diagnostic) natural frequency tree in Figure 2A differ from that with the fast-and-frugal classification tree in Figure 2B? A Bayesian physician would first check each cue and determine the patient's cue profile. She would then assess the probability of an infarction for that cue profile based on the information in the fifth level of the natural frequency tree—this is, using what we term profile memorization. For example, the probability of infarction among the 17 patients with a cue profile [0, 1, 1], that is [ST_{-}, CP_{+}, OS_{+}], is 2/17. This probability is then considered as the best estimate of the posterior probability for a new patient with the same cue profile. Finally, a threshold could be used to decide if the probability of infarction is high enough to classify the

new patient as being at high risk. For example, if the threshold is 0.1, new patients with cue profiles [1, 1, 1], [1, 1, 0], [1, 0, 1], or [0, 1, 1] would be classified as high risk, and those with any other cue profile would be classified as low risk.

The fast-and-frugal tree displayed in Figure 2B distinguishes between only four, rather than eight, cue profiles. These are [1, *, *], [0, 1, 1], [0, 1, 0], and [0, 0, *]. The "*" denotes that this particular cue is not consulted. For instance, the cue profile [1, *, *] denotes that all 33 patients with ST_+ are pooled together, without any further differentiation. As a consequence, they all have the same posterior probability, namely 13/33. Note that the fast-and-frugal tree handles the cue profile [1, 0, 0] differently from the full natural frequency tree. In the full tree, a patient with this profile would be classified as a low risk patient (the only two patients with this profile had no infarction). In the fast-and-frugal tree, in contrast, information search stops if the first cue (ST) is positive, and—if the threshold was again .1—all 33 patients for whom this holds (including the two with [1, 0, 0]) would be classified as high risk (because 13/33 > .1).

What are the advantages of natural frequency trees over fast-and-frugal trees? One advantage is that the classification of a new patient is tailored to the patient's cue profile. The downside is that this procedure needs a lot of data—*all* questions on cue values have to be asked for each patient. However, as the number of questions increases, the number of distinct cue profiles increases (exponentially). As a consequence, the average number of patients per cue profile decreases and for some profiles the classification may be based on a tiny number of observed patients. For example, the classification for the profile [1, 0, 0] in Figure 2A is based on only two patients. A procedure that bases many classifications on small samples may be quite brittle.

In contrast to full natural frequency trees, fast-and-frugal trees are not complete; they truncate some of the possible paths. In fact, a large number of patients are assigned to a category after only one cue value has been checked. This feature makes this strategy much more frugal. Furthermore, in contrast to the Bayesian approach, some classifications are made jointly for patients with different cue profiles. For example, the classification for the profile [1, 0, 0] is made jointly with the profiles [1, 1, 1], [1, 1, 0], and [1, 0, 1]. This practice may contribute to robustness—that is, good generalizations to new patients. Given that the fast-and-frugal tree has fewer nodes than the full tree, the average number of patients per exit node is larger, which on average moves the estimates of the ratios in a given sample closer to that of the underlying population. Finally, because the fast-and-frugal tree is transparent and easy

to apply, it suits the hospital environment and may be more appropriate for guiding behavior in emergencies.

3.3.2 Fast-and-Frugal Trees as Lexicographic Strategies

The fast-and-frugal tree in Figure 2B contains a bold vertical bar that splits the cue profiles in two parts: those to the left and those to the right of it. The cue profiles to the left correspond to patients classified as being at high risk of a myocardial infarction. Transposing this vertical bar to the corresponding position in Figure 2A shows that new patients with cue profiles [1, 1, 1], [1, 1, 0], [1, 0, 1], [1, 0, 0], or [0, 1, 1] are classified as high risk by the fast-and-frugal tree, and those with profiles [0, 1, 0], [0, 0, 1], or [0, 0, 0] as low risk. Inspecting this arrangement of profiles and the cutoff leads to the insight that fast-and-frugal trees can be characterized as models that classify objects *lexicographically*. To show this some mathematical rigor is required. We assume *n* binary cues and two classes or categories, C_1 and C_0 (corresponding to the presence and absence of disease, respectively). Without loss of generality, we also assume that cues are inspected in the order $c_1, c_2, \ldots, c_n$ and that for $i = 1, 2, \ldots, n$, cues are coded so that if an object x exits the tree at the *i*th level, it is assigned to category C_1 if $x_i = 1$ and to C_0 if $x_i = 0$.

> *Definition*: A cue profile x is lexicographically larger than a cue profile y ($x >_l y$) if and only if there exists $1 \leq i \leq n$ such that $x_i = 1$ and $y_i = 0$ and $x_j = y_j$ for all $j < i$. If neither $x >_l y$ nor $y >_l x$, then x and y coincide ($x = y$).

The following result establishes a characterization of fast-and-frugal trees as lexicographic classifiers based on splitting profiles (Martignon et al., 2003; Martignon et al., 2008).

> *Result 1*: For every fast-and-frugal tree f there exists a unique cue profile $S(f)$—called the tree's *splitting profile*—such that f assigns x to C_1 if and only if $x >_l S(f)$. For every cue profile S there exists a unique fast-and-frugal tree f, such that $S(f) = S$.

In Figure 2B, let $x_1 = 1$ if and only if the ST segment is elevated, $x_2 = 1$ if and only if chest pain is the main symptom, and $x_3 = 1$ if and only if any other symptom is present. Also let C_1 represent high risk and C_0 low risk. The splitting profile of this tree is [0, 1, 0]. The bold vertical bar marks the position of the splitting profile. All cue profiles to the left of the bar are lexicographically larger than the splitting profile. As Result 1 states, these cue profiles are assigned to the high risk category C_1. Note that the full natural frequency tree that

follows the same convention of placing positive (negative) cue branches to the left (right) does not have a splitting profile: If a threshold of .1 is adopted, against which the posterior probability is compared, then the classifications for the eight distinct cue profiles are, from left to right, C_1, C_1, C_1, C_0, C_1, C_0, C_0, and C_0. For the fast-and-frugal tree, in contrast, the classifications for the four distinct cue profiles are, from left to right, C_1, C_1, C_0, and C_0. This sequence can be split in the middle.

3.3.3 Fast-and-Frugal Trees as Linear Classifiers with Noncompensatory Weights

Due to their extremely simple structure, fast-and-frugal trees can even be integrated within the class of linear classifiers, specifically, as linear classifiers with noncompensatory weights. In linear models for classification, each cue c_i has a *weight* $w_i > 0$ and the score $R(x) = \sum_i x_i w_i$ is computed for each cue profile $x = [x_1, x_2, \ldots, x_n]$. A scalar *threshold* $h > 0$ defines the categories, so that an item *x* is assigned to C_1 if and only if $R(x) > h$. A linear classifier in which all weights are 1 is called *tallying*. The following result relates linear and lexicographic inferences for classifications (Martignon et al., 2008):

> *Result 2*: For every fast-and-frugal tree *f* there exist $h > 0$ and $w_i > 0$, where $w_i > \sum_{k>i} w_k$ for $i = 1, 2, \ldots, n-1$, such that f makes identical classifications to a linear model with weights w_i and threshold *h*. For every linear model with weights $w_i > 0$, such that $w_i > \sum_{k>i} w_k$ for $i = 1, 2, \ldots, n-1$ and a threshold $h > 0$, there exists a fast-and-frugal tree *f* that makes identical categorizations.

For example, the Green and Mehr (1997) tree in Figure 2B makes identical classifications as a linear model with $R(x) = 4x_1 + 2x_2 + x_3$ and $h = 2$ (both assign [0, 0, 0], [0, 0, 1] and [0, 1, 0] to C_0 and all other cue profiles to C_1).

Linear models with $w_i > \sum_{k>i} w_k$ are called *noncompensatory* (Martignon & Hoffrage, 2002). Result 2 states that fast-and-frugal trees are equivalent to noncompensatory linear models in the sense that the two make the same classifications. Note, however, that Result 2 does not imply that it is impossible to distinguish empirically between fast-and-frugal trees and noncompensatory linear models. The process predictions of fast-and-frugal trees, including ordered and limited information search, are distinct from those of linear models, which use all the available information in no specified order (for a study in which the outcome and process predictions of trees were pitted against those of compensatory models, see Hertwig, Fischbacher, & Bruhin, 2013).

3.3.4 Heuristic Principles for Tree Construction

We now explain how a fast-and-frugal tree such as that displayed in Figure 2b can be constructed. The two key features of a fast-and-frugal tree are its ordering of cues and its exit structure. Given that n cues can be arranged in $n!$ different orders, the task of selecting an order is not trivial. Implementing different orders and seeking to identify that with the best performance would be cumbersome and not in the spirit of the simple heuristics program (Gigerenzer et al., 1999). In the remainder of this section, we propose heuristic principles for the construction of trees. Note the plural: trees. Because trees can be evaluated with respect to a range of measures, and because different domains and applications call for different measures, having a toolbox of trees at one's disposal (instead of just one or two) is in the spirit of ecological rationality.

The ordering of cues is determined by their relationship with the criterion. As stated above, fast-and-frugal trees assume cues to be independent of one another, conditioned on the criterion. They are constructed without checking whether this assumption is justified: each cue's relationship with the criterion is measured without considering the other cues. This is exactly what *take-the-best*, a simple heuristic for paired comparison, also does: it orders cues according to their validity (Gigerenzer & Goldstein, 1996). Cue validity is defined as the proportion of correct inferences divided by the number of total inferences that a cue allows in a particular set of comparisons (Gigerenzer, Hoffrage, & Kleinbölting, 1991)—and this measure is determined for each cue separately, thereby ignoring any cue intercorrelations or dependencies. Whereas in a paired comparison it makes no difference whether one alternative or the other is chosen incorrectly, in the context of medical diagnoses it is important to distinguish these two possible errors: misses and false alarms. In fact, a cue for medical diagnosis can be characterized by two measures: the *positive predictive value* (PPV) of a cue, that is, the proportion of patients with the disease among all patients for whom the cue is positive (or in its high state, i.e., $P(D|E)$), and the *negative predictive value* (NPV), that is, the proportion of patients without the disease among all patients for whom the cue is negative (or in its low state, i.e., $P(\bar{D}|\bar{E})$). In other words, PPV and NPV indicate how diagnostic a cue is given that it has a positive or negative value, respectively.

For illustration, the PPV for mammography in Figure 1D is 80/1030; the NPV for ultrasound in Figure 1E is 9504/9509. A measure that captures the relationship between cue and criterion, merging the two possible classification errors, is the accuracy A (or validity) of

a cue, $P(D \wedge E) + P(\overline{D} \wedge \overline{E})$; for the ultrasound in Figure 1E, the accuracy (A) is $\frac{95+9504}{10000} = 0.9599$.

As stated above, a fast-and-frugal tree entails a classification at each level—for each cue, one of the two possible values will lead to an exit node (i.e., a classification) and the other to consulting more cues (with the exception of the lowest ranked cue, which has two exit nodes). The ordering of cues and the structure of exit nodes are related to each other as follows: If PPV (NPV) has been used to identify which cue comes next in the hierarchy, then the exit node of this cue is placed on the left (right). This means that if the cue is positive (negative), the patient will be classified as having (not having) the disease, and if the cue is negative (positive), information search will continue. Obviously, classifying a patient as having the disease can lead to a false alarm, but it can never be a miss. This means that ordering cues according to their PPVs and constructing the tree such that all exit nodes are on the left will probably produce only a few misses—but it may produce many false alarms (note that misses can occur, but only if a patient is classified as not having the disease after all cues have turned out to be negative). Conversely, ordering cues according to their NPV and placing all exit nodes on the right will probably produce only a few false alarms—but it may produce many misses.

Rakes (R). Following Martignon et al. (2003) and Martignon et al. (2012), we refer to a tree with all exits on the same side as a *rake*. A tree that orders cues by PPV and thus has all exit nodes on the left is denoted by R_+; a tree that uses NPV and has all exit nodes on the right, by R_-. Note that for each of these trees, the ordering of cues does not make a difference to classifications. For instance, R_+ will classify a patient as high risk if (and only if) any of the cues has a positive value (no matter where in the rake this cue is positioned). Obviously, both R_+ and R_- are quite single sided, in the literal sense.

Zigzag trees (Z). A more balanced tree structure is what Martignon et al. (2003) and Martignon et al. (2012) referred to as zigzag tree. We define Z_+ as starting with the cue with the highest PPV and placing it at the top of the diagnostic tree, with the exit node on the left. Subsequently, Z_+ identifies the cue with the highest NPV among all remaining cues and places it second, with the exit node on the right. For the third position, it identifies the remaining cue with the highest PPV (and puts its exit node on the left), and so on. Z_- follows the same logic but it starts with the cue with the highest NPV, continues with the cue with the highest PPV,

and so on. While Z_+ and Z_- determine a priori whether the first exit node is on the left or right, a new tree we call Z_0 is adaptive in the sense that it lets the data set determine whether Z_+ or Z_- will be chosen. If the highest predictive value is a PPV, the zigzag tree starts with an exit to the left (i.e., $Z_0 = Z_+$); if it is an NPV, then $Z_0 = Z_-$.

Base-rate respect (B). We now introduce trees that combine subtrees of the rake and of the zigzag type. These trees are more ecological, because the data set has an even larger impact on their shape. Each of the tree construction principles described above ignores the base rate of positive criterion values. Each would thus lead to one and the same tree for the same set of PPVs and NPVs, irrespective of whether the prevalence of the disease is 50% or, say, as small as 0.01%. It is easy to see how zigzag trees suffer under this base-rate neglect. Similarly, if the disease is rare, it may not be a good idea to construct a rake tree with exit nodes rigidly pointing to the left for every cue in the hierarchy. Such a procedure will typically lead to many false alarms. We therefore propose two trees that combine features of rakes and zigzag trees. In the standard variant, denoted *B*, for a base rate of $b < 0.5$, the first $k = \lfloor |\log_2(b)| \rfloor$ cues have exit nodes on the right, thereby classifying patients into the majority category. These *k* cues, and their ordering, are determined by the maximum NPV. The formula for *k* implies that base rates of a disease between .25 and .5 would result in trees starting with one exit node on the right; base rates between .125 and .25, in trees starting with two such nodes; base rates between .0625 and .125, in trees with three nodes, and so on. Correspondingly, for $b > 0.5$, the tree starts with $k = \lfloor |\log_2(1-b)| \rfloor$ cues, determined by the PPV, that have exit nodes on the left. After these *k* cues with fixed exit directions, the tree continues with a zigzag structure that starts with a classification into the minority class. The remaining cues are ordered by alternating between picking one with maximum PPV and one with maximum NPV. This hybrid is biased against early classifications into the minority class. In fact, the bias results from exploiting the base rate of the disease in a smart way. The B_1 tree is even more biased: It does not start with k, but with $k + 1$ cues that classify objects into the majority category.

Maximum predictive value (M). Our next tree is also quite sensitive to the environment. As for all trees, the position of the exit is determined by whether the cue at this position has been picked by its PPV or its NPV. The tree we refer to as *M* chooses the maximum predictive value among all remaining cues. But unlike Z_0, which does this only for the first cue position, *M* does it for each position.

Accuracy (A). The final tree construction principle we want to introduce here differs from the others in that it uses the cues' accuracy to establish their ordering; hence the abbreviation *A*. The PPV and the NPV are still considered, however—not to determine the position in the cue ordering, but the direction of the exit node. If for a given cue PPV > NPV, the exit is to the left; otherwise, it is to the right. For each of the tree construction principles proposed above, ties in the process are broken randomly.

For the sake of completeness, we note that Woike (2008) also tested trees for which the ordering of cues was established not by PPV and NPV (measures that can be read off from diagnostic trees; see, e.g., Figures 1D and 1E), but by sensitivity, $P(E|D)$, and specificity, $P(\overline{E}|\overline{D})$ that is, measures that stem from causal trees (e.g., Figure 1A). As their performance was generally weaker than that of those we listed above, they are not reported in this article (for more information, see Martignon et al., 2003, and Woike, 2008).

3.4 Complex Classification Methods

Other classification methods include CART, logistic regression, Bayesian networks, neural networks, support vector machines, and exemplar models. Although some of them are simple in execution, their construction often requires complex computations. We therefore decided to not include them in the present work that focusses on linking and comparing the natural Bayesian with simple heuristics, specifically, with Naïve Bayes and fast-and-frugal trees. Note that each model tested in this article can be set up with paper and pencil alone, without the need for complex computations. We will, however, return to some of these complex classification methods in the General Discussion below.

4. Method

We now describe how we set up the simulations and implemented the strategies described in the previous section.

4.1 Data Sets

Most of the data sets used were taken from the UCI Machine Learning Repository (Lichman, 2013). Table 1 displays these 11 data sets. In each data set, a number of objects (ranging from 62 to 768) are described by the state of a dichotomous criterion variable (e.g., disease) and by a number of cues (ranging from 5 to 22).

<<<<<< Table 1 >>>>>

Many of the cues were dichotomous; others were continuous. The latter were dichotomized by assigning values larger than the median to the "high" category and values less than or equal to the median to the "low" category. Which cue value (e.g., male/female) was coded as a high state cue value was determined within each training sample, such that the predictive accuracy of a cue always exceeded .5.

4.2 Procedure

People have to make most real-life decisions in novel situations, and they make those decisions based on past experience. We therefore set up our simulations so that parameters need to be estimated from small samples, and classifications need to be made for new cases. Technically, this means that a classifier was estimated for each method in each data set by (1) sampling a random subset of the data as a training sample, (2) dichotomizing continuous variables (if any) based on the information contained in the training sample, (3) fitting the classifier to the training sample, and then (4) classifying each of the remaining objects (i.e., those that were not included in the training sample; henceforth, the test sample). This procedure was repeated 2,500 times for each classifier. The whole process was carried out for training samples of 15%, 50%, and 85% of the data set. Classifying a different set of observations from those used to train the classifiers is standard practice in classification research; this approach tests the method's ability to generalize to new data. Testing strategies according to this procedure is clearly an instantiation of Case 3 discussed above.

4.3 Implementation of Strategies

We now describe how we implemented the three approaches introduced above.

4.3.1 Profile Memorization Based on Natural Frequencies

Classification based on diagnostic frequency trees such as the one displayed in Figure 1B was achieved by determining, separately for each cue profile in the training sample, the proportion of patients with the disease, and then classifying a new patient with this profile as having the disease if this proportion was greater than a critical threshold. This critical threshold is typically determined based on cost–benefit considerations of the cells in the 2*2 table (classifications × criterion values). In our simulations, we realized all possible thresholds, that is, we varied the threshold from 0 to 1 in small increments. For new patients whose profile had not been observed in the training sample, the classification was made by comparing the proportion of patients with the disease across all profiles in the training sample with the threshold that maximized accuracy within the training sample.

4.3.2 Naïve Bayes

The Naïve Bayesian method is regarded by Bayesian statisticians as a heuristic approach, because it assumes statistical independence of cues for a given criterion value. We used the Matlab implementation of Naïve Bayes, more specifically the "fitcnb" function in Matlab 2016Rb with standard parameters and all variables entered as categorical predictors. This function was given the same cue values that entered the tree construction algorithms, with the same cue directions and the same split of continuous cues. Predictions for objects from the test sample were generated by the "predict" function with the fitted model passed as parameter. These predictions are individual values between 0 and 1; these values were used to compute the Receiver Operating Characteristics (ROC) curves for Naïve Bayes. We varied the classification threshold from 0 to 1 in small increments, as we did for the profile memorization method.

4.3.3 Fast-and-Frugal Trees

We determined the performance for each of the trees introduced above. Ties in the process were broken randomly. When a PPV or an NPV was undetermined, because its computation would have required zero to be divided by zero, this predictive value was placed at the very end of the hierarchy. In contrast to profile memorization and Naïve Bayes, the classification of fast-and-frugal trees hinges on the exit structure of the tree and not on the comparison of posterior probabilities with a threshold.

4.4 Dependent Variables

The performance of the strategies was evaluated in two ways. The first is based on ROC diagrams. These diagrams include the ROC curves of profile memorization and of Naïve Bayes. The ROC curves plot the sensitivity (percentage of correct identifications among patients with the disease) against the false-alarm rate (percentage of incorrect identifications among patients without the disease; the complement of specificity) as the classification threshold glides from 0 to 1. As fast-and-frugal trees do not use such thresholds, their classification performance cannot be represented as a curve, but by one point per construction principle. To illustrate, the tree in Figure 2B would be plotted at $x = 35/50$ (1 – specificity = false-alarm rate: among all $20 + 15 + 10 + 29$ patients who had no infarction, $20 + 15$ were classified as high risk) and $y = 1$ (sensitivity: all 15 who had an infarction were classified as high risk). There is one difference, though, to the entries in the ROC diagrams shown below: The points in these diagrams do not represent the classification

performance of trees, but that of tree construction *principles*. In 2,500 trials, each construction principle generated a total of 2,500 trees and most likely not all of them were the same.

Although all objects in the test sample are included when computing sensitivities and false-alarm rates, the data points in the ROC diagrams do not convey the base rates of patients with and without disease. Imagine a strategy that performs well with respect to one of the two possible classification errors (i.e., this proportion is quite low) but poorly with respect to the other error (i.e., this proportion is quite high). Each data point in the ROC diagram reveals these two proportions, but it is mute about the underlying absolute frequencies. The same data point can signal high or a low accuracy depending on which of the two errors occurs how often. In other words, ROC curves display what Hoffrage, Gigerenzer, Krauss, & Martignon (2002) called normalized frequencies, rather than natural frequencies.

Our second measure, the percentage of correct classifications, is not fooled by such differences with respect to base rates. On the contrary, this measure is sensitive to base rates: It computes *one* single measure across all cases, thereby weighing each correct classification equally, no matter whether a patient with or without the disease has been classified correctly. However, this measure does not consider the different types of errors, namely misses (classifying a diseased patient as healthy) and false alarms (classifying a healthy patient as diseased). While each tree allowed each object in the test sample to be classified without any further specification, profile maximization and Naïve Bayes required a classification threshold. Each threshold included in our simulations led to a point in the ROC curve and to a percentage of correct classifications. The percentages reported below are those that result from a threshold that was arbitrarily set at .5.

In sum, the advantage of one dependent variable is the disadvantage of the other, and vice versa, and it is hard to establish which of the two variables is superior. This observation is important because, as we will see below, the two methods of measuring performance arrive at different conclusions when evaluating our strategies.

5. Results

In this section, we compare the performance of the three classification approaches introduced above on the 11 data sets presented in Table 1.

5.1 ROC Diagrams

Figure 3 displays the ROC curves (and data points) for our strategies, separately for the 11 data sets, when the training sample consists of 50% of the objects (Panels A–K), and the ROC curve (or points) resulting from averaging those data points across the data sets (Panel L). The major findings from Figure 3 are the following.

<<<<<< Figure 3 >>>>>

First, in two data sets (Figures 3H and 3K), both the curves of profile memorization and Naïve Bayes, and the points of the fast-and-frugal trees were close to the diagonal, the areas under the curves of the former were close to .5, and consequently, a strategy's *d'* (a measure capturing how well the strategy can differentiate between the two categories, disease present versus absent) should be close to zero. In other words, in these two data sets, performance of all strategies is at chance level. Nevertheless, it is interesting to see that all of them were equally affected by the nondiagnosticity of the cues. In the other nine data sets, the cues were diagnostic and the strategies performed above chance.

Second, the ROC curve of Naïve Bayes by and large dominated that of profile memorization. Although profile memorization, an instantiation of Bayes' rule, provides optimal classifications when fitting known data (Case 1), for which it is considered to be normative, it is outperformed by its heuristic sibling, Naïve Bayes, when classification performance is evaluated for new patients (Case 3). This can be seen with the naked eye; there is no need to compute the area under the curve. Naïve Bayes hence lives up to its reputation as a simple and robust benchmark for classifications under uncertainty—that is, classifications of new objects and when the probabilities in the population are not known but are estimated from samples.

Third, the fast-and-frugal trees competed well with Naïve Bayes. In 9 of 11 data sets, the trees' performance was by and large on the ROC curve of Naïve Bayes; for the other two data sets, the trees' performance was clearly less good (with the exception of R_+, R_- and A, which were always on Naïve Bayes' ROC curve).

Fourth, the fast-and-frugal trees generally outperformed the profile memorization method: Of the 99 points representing the performance of the nine tree construction principles across the 11 data sets, 74 were above the ROC curve for profile memorization, 19 were on it, and only six below.

Fifth, comparing the fast-and-frugal trees among each other revealed that, as expected, R_+ and R_- (which minimize different classification errors) were located in the upper right and lower left corner, respectively. Z_+ and Z_- (which strike a balance between the two classification errors by alternating the exit nodes) moved away from the corners and were located more in the middle. In each data set, Z_0 was identical to either Z_+ or Z_-. This was necessarily the case for each of the 2,500 trials in a given data set, and reflects how Z_0 is designed. The observation that this was also true, on average, across 2,500 trials within a given data set, suggests that there is little variance, within a data set, with respect to the direction of the first exit node. Across data sets, however, Z_0 was "between" Z_+ and Z_-; this was again as expected because the points in Panel L result from averaging the points in Panels A–K. B and B_1 (which are ecologically rational in the sense that they take the base rate of the disease into account) lie, with the exceptions of Panels F and J, on the ROC curve of Naïve Bayes.

Sixth, for data sets with a base rate of the disease below .4 (Figure 3A–E), A and M were located in the lower left corner; for base rates above .6 (Panels H–K), they were in the upper right corner; and for base rates between .4 and .6 (Panels F and G), no clear picture emerged. For the other tree construction principles, no such dependency of location on the base rate of the disease was observed.

Seventh, it is interesting to observe that all trees in Panel L, which displays the strategy-specific midpoints across data sets, were close to Naïve Bayes' ROC curve—with the exception of M and A. Hence, at least at first glance, ordering cues according to their maximum predictive value (whether this is PPV or NPV) or the cues' predictive accuracy does not appear to perform as well as the other ordering principles. However, this conclusion would be misleading. Note that the points representing the performance of M and A are located either in the lower left corner or in the upper right corner. The midpoint of these 11 points will therefore be somewhere in the middle of the diagram, close to the diagonal, and hence well below Naïve Bayes' ROC curve. This was not the case for the other trees, whose points were not as dispersed as those of M and A. Therefore, computing averages across data points might lead to wrong conclusions and is thus problematic. To illustrate, note that the performance of tree A lies on Naïve Bayes' ROC curve for each of the 11 data sets—but when averaged across data sets, it is well below this curve (see Figure 3L).

<<<<<< Figure 4 >>>>>

Bearing this problem in mind, let us now turn to Figure 4 which displays, across the 11 data sets, the strategies' performance for different sizes of training sample (note that Figures 3L and 4B are identical). As shown, performance is quite robust across sizes of training sample. The strategy most affected by this dimension is classification based on profile memorization: As the size of the training sample increases, in particular from 15% to 50%, the ROC curve moves away from the diagonal. We have already mentioned the brittleness of profile memorization's estimates when these are based on low numbers for a given cue profile. In contrast, the heuristics—both Naïve Bayes and the fast-and-frugal trees—did not seem greatly affected.

5.2 Percentage of Correct Classifications

Figure 5 displays the performance with respect to our second dependent variable: percentage of correct classifications, irrespective of how the incorrect classifications were distributed across the two possible errors. For each size of training sample, the clear winner here is Naïve Bayes. The difference between Naïve Bayes and profile memorization was about five percentage points. Of the fast-and-frugal trees, the two ecological variants, *B* and B_1, performed best. None of the rakes, R_+ and $R_–$, or zigzag trees, Z_+, $Z_–$, and Z_0, outperformed classification via profile memorization. In contrast, *M* did so, and *A* matched its performance, at least when the training sample encompassed 15% of the population. Note that these were exactly the strategies that seemed to perform worse in the averaged ROC diagram (Figure 3L), at least at first glance. In each of the 11 strategies, increasing the training sample from 15% to 50% led to higher accuracy gains than did increasing it from 50% to 85%.

<<<<<< Figure 5 >>>>>

5.3 Discussion: Comparing ROC Diagrams with Percentage of Correct Classifications

Comparing averaged performance in the ROC diagram (Figure 3L) with predictive accuracy (Figure 5) yielded some seemingly paradoxical observations. To take the most extreme example, in the ROC diagram, Z_0 showed higher specificity and sensitivity than the construction principles *M* and *A*, but lower overall predictive accuracy. Note that it is logically impossible that one algorithm, in any given trial, achieves higher accuracy across all patients than another if this other algorithm achieves higher accuracy than the former for both patients with and patients without the disease (as these comprise the total number of patients in the data set). So why do the zigzag trees dominate *M* and *A* in the ROC diagrams when

averaged across data sets (Figure 3L), if, conversely, *M* and *A* outperform Z_+ and Z_- with respect to predictive accuracy (Figure 5)?

As we have mentioned, the dispersion of the performance measures of a given construction principle across the data sets (Figure 3A–K) must be considered when interpreting the averages in Figure 3L. But there is more to it than this. The relevant question is why some tree construction principles are more dispersed than others. Observe that, when moving from the lower left to the upper right corner of Naïve Bayes' ROC curve, the following tree construction principles always appear in this order: R_-, Z_-, Z_0, Z_+, and R_+. In contrast, *B* and B_1 seem to "dance" along this strict ordering, sometimes appearing before Z_- and sometimes after Z_+. The same can be said for *M* and *A*, but the range these principles span is even wider. Moreover, the positioning of *B*, B_1, *M*, and *A* within the strict ordering of R_-, Z_-, Z_0, Z_+, and R_+ seems to be related to the base rate of the disease. As we shall see, this pattern is related to the paradox mentioned above (e.g., any *Z* is better than *M* in Figure 3L, but worse in Figure 5).

One important piece of the puzzle is what we said when introducing the dependent variables in the Methods section: Any point in an ROC diagram plots the two classification errors but it does not contain the base rates of the disease. Predictive accuracy, in contrast, ignores which type of classification error occurs how often and reports only the complement, namely correct classifications. Thereby it takes the base rate of the disease into account, albeit only implicitly: When the percentage of correct classifications is computed, the objects in the majority class (e.g., patients with disease) contribute more to the score than do those in the minority class. In other words, an ROC diagram contains normalized frequencies, whereas predictive accuracy reports the sum of natural frequencies. The other important piece of the puzzle is that data sets differ with respect to base rates of the disease (see Table 1). Let us now put the pieces together.

Sensitivity and specificity contribute equally to the overall accuracy if (and only if) the prevalence of the two classes, that is, patients with disease and without disease, is equal. When the base rate of the disease is above .5 (as in Figure 3G–K, see Table 1), sensitivity is more important than specificity for achieving high accuracy; when the base rate is below .5, specificity is more important than sensitivity.

Now imagine that we add a line with a slope of 45º (i.e., a parallel to the identity line) to a given ROC curve (say, for Naïve Bayes), with the line touching the ROC curve at one point. In Figure 3L, this point would be close to B, B_1, and Z_0. Moving this point along the ROC curve would involve an uneven tradeoff: An increase in specificity (achieved when moving toward the lower left corner) is coupled with a comparatively greater decrease in sensitivity. Conversely, and formulated differently, the price to be paid for one additional percentage point in correctly classifying patients with disease (i.e., moving toward the upper right corner) is a decrease of *more* than one percentage point in correctly classifying patients without disease. The loss in specificity can still be advantageous in the sense that it can boost overall accuracy: A loss in specificity implies a gain in sensitivity, and if there are more patients with the disease, then such a tradeoff will, overall, lead to more correct classifications.[1]

However, one can overdo it: R_+ is designed to maximize sensitivity and R_- to maximize specificity. As a consequence, they end up in the corners (upper right and lower left, respectively) of the ROC diagrams. But being located close to one of these corners will translate into high total accuracy only if the disease base rate, and hence the slope of the iso-accuracy curve, is extreme. This was not the case in any of our 11 data sets. Hence, R_+ and R_- not only sacrifice specificity and sensitivity, respectively, but also total predictive accuracy, as Figure 5 confirms. Conversely, B and B_1, the two construction principles that exploit the base rate of the disease (by the design of their exit structures), are the winners among all trees with respect to predictive accuracy (Figure 5). This observation is consistent with the fact that these principles, together with Z_0, are closest to the 45º tangent point in the averaged ROC diagram. This should not be taken to imply that R_+ and R_- always build inferior trees. For a problem, in which it is much more important to avoid classifying patients as negative, when

[1] Note that the iso-accuracy line has a slope of 45º if and only if the base rate of the disease is .5. For base rates higher (lower) than .5, the iso-accuracy line is flatter (steeper). If, for example, 90% of the patients have the disease, then increasing sensitivity by one percentage point (and hence moving toward the upper right corner) will be enough to compensate for a loss of nine percentage points in specificity. Therefore, the tangents of those construction principles that are sensitive to the base rate of the disease are generally flatter (steeper) and hence closer to the upper right (lower left) in those data sets in which the base rate is above (below) .5.

they are in fact positive than it is to avoid making the reverse error, R_+ might offer a viable and desirable solution.

Remember that B and B_1 are designed to exploit extreme base rates. The higher the base rate of the disease, the more nodes with exits to the left the tree will initially have. Exits to the left lead to classifications of the disease as being present. If patients with disease form the majority category, then a tree that is biased towards classifying patients as having the disease will still produce misses, but these errors will become relevant for only a minority. This is why and how the ecological rationality of the B tree (here, in the sense of base-rate respect) is rewarded in terms of overall accuracy. Now we can also understand why the B trees for data sets with high disease base rates are close to R_+ (see Figures 3K, 3J, 3I, 3H, …) and those for data sets with low base rates are close to R_- (3A, 3B, 3C, 3D, …). Not surprisingly, the B_1 tree, which pushes the enforced rake structure at the beginning of the tree one step further, is always closer to the corner than the B tree. Ironically, moving closer to the corners (and using base rates to determine towards which to move) not only has a positive impact on overall accuracy (note that it has a higher accuracy than B, see Figure 5) but also increases the dispersion of the B_1 trees, relative to the B trees, which, in turn, moves the B_1 average below Naïve Bayes' ROC curve (we explained why this happens when discussing how averaging across data sets affects the placements of M and A).

The results for M and A also offer some interesting insights. Their performance is also affected by the base rate of the disease, yet for a different reason and in a more extreme way than for B and B_1. Remember that the values for continuous cues were dichotomized at the median, yielding (roughly) as many positive as negative cue values. This has implications for the PPVs and NPVs: If, for instance, the base rate of the disease is above .5, then the number of patients without the disease (which will be less than half) will necessarily be lower than the number of patients with a negative cue value (after median-split, hence about half of the patients). This mismatch makes it likely that a cue's PPV will be higher than its NPV. Why? As there are more diseased patients than positive cue values, it is possible to achieve a PPV of 1; with fewer healthy patients than negative cue values, however, it is impossible to achieve an NPV of 1. If PPVs tend to be larger than NPVs, this will create a bias towards exit nodes pointing to the left and hence to classifying patients as having the disease. The stronger this bias, the higher the sensitivity. Indeed, note that for all data sets with a base rate of disease above .5, the positioning of the M and A trees is very close to the upper right corner of the

ROC diagrams where sensitivity is high (Figure 3G–3K); for the other data sets, these trees are typically found in the lower left corner.

The corners are also the "home" of R_+ or R_-. But there is an important difference: Whereas R_+ is always upper right and R_- is always lower left without any base-rate sensitivity, the classification performance of M and A is base-rate sensitive. Whereas the rigidity of R_+ or R_- is penalized in terms of predictive accuracy, M's and A's adaptive bias toward the majority class is rewarded and they by far outperform the rake trees (Figure 5). At the same time, their bias is more extreme than that of B and B_1, and their accuracy suffers in comparison.

To conclude this discussion, the positioning of the tree construction principles in the 11 ROC diagrams and, in particular, in the averaged ROC diagram can be properly interpreted only when base rates of the disease are taken into account, when the tree construction principles are well understood, and when the danger of averaging is recognized. Our two dependent variables—ROC diagrams (which distinguish the two classification errors) and overall accuracy (which is base-rate sensitive)—are complementary. One has what the other lacks. The ROC diagrams alone would not allow us to infer the rank ordering of the strategies with respect to predictive accuracy; the accuracy results would not give us any idea of what the ROC diagrams might look like. The full picture emerges only if both are seen jointly. We chose to present the results for percentage correct, because it is an intuitive measure and stays within the framework of natural frequencies. For any concrete problem with known base rate and error costs, the ROC diagram might offer a better basis for deciding between construction algorithms, as the percentage of correct classifications does not differentiate between error types and those can differ in severity or importance (see Verde, Macmillan & Rotello, 2006, for a discussion of alternative unidimensional measures combining hit rates and false-alarm rates motivated by signal detection theory).

6. General Discussion

This article has integrated two lines of research that had previously developed largely separately, namely research on the advantages of using natural frequency representations rather than probabilities to communicate risk (Gigerenzer & Hoffrage, 1995) and research on heuristics (Gigerenzer et al., 1999; Hertwig, Hoffrage, & the ABC Research Group, 2013; Todd, Gigerenzer, & the ABC Research Group, 2012). We have shown that Bayesian inference tasks with more than one cue can be solved by means of natural frequency

representations. We have also argued that such representations have limits. These limits can be addressed by two heuristic approaches. First, Naïve Bayes simplifies the task by assuming independence between cues—but still integrates across all cues. When combined with a classification threshold, Naïve Bayes' computation of posterior probabilities can lead to classifications based on multiple cues. Second, fast-and-frugal trees also assume independence between cues, but they go one step further and truncate information search. The trees we proposed do not require a classification threshold, and they do not compute posterior probabilities. Our article integrates these three approaches to classification—natural frequencies, Naïve Bayes, and fast-and-frugal trees—into a common framework that makes it possible to conceive (1) Naïve Bayes as a tree that enforces independence between cues and (2) fast-and-frugal trees as pruned versions of such trees, as lexicographic strategies, and as linear classifiers with noncompensatory weights.

Previous research for other tasks, such as paired comparison (Czerlinski, Gigerenzer, & Goldstein, 1999; Martignon & Hoffrage, 2002) or estimation (Hertwig, Hoffrage, & Sparr, 2012; Woike, Hoffrage, & Hertwig, 2012), has shown that simple heuristics, when making inferences out of sample, can outperform models that are optimal when it comes to making inferences within the *same* sample for which the strategies fitted their parameters. We found the same pattern in our simulations: When making classifications out of sample with multiple cues, the two heuristic approaches, Naïve Bayes and some of the fast-and-frugal trees, outperformed the model that was normative for the case of fitting known data, namely classification based on natural frequencies (or, equivalently, profile memorization).

What are the advantages of fast-and-frugal trees? First, they are easy to set up. Their parameters can be estimated with paper and pencil (remember, PPV and NPV can be obtained by counting and forming simple ratios; see the frequency trees in Figures 1D and 1E). Note that fourth graders can already understand the concept of cue validity through enactive education approaches—manipulating towers of colored tinker cubes to represent the relationship between cues and outcomes—and can answer elementary questions on the validity of cues (Martignon & Monti, 2010; Till, 2014). Once cue validities have been estimated, tree construction involves only a few simple rules. Second, fast-and-frugal trees are transparent and therefore easy to use and communicate. Their graphical representation allows for straightforward application in practical settings (Fokkema, Smits, Kelderman, & Penninx, 2015). Their execution does not require any complex computation. Naïve Bayes uses the same

input needed for tree construction, then uses all information, and finally requires some computation on the execution level—albeit only multiplication that a teenager should be able to perform with paper and pencil alone. Third, fast-and-frugal trees are, unlike Naïve Bayes, frugal. They generally use only a fraction of the potentially available information because each of these trees has an exit on every level. Fourth, these trees allow for fast classifications. In emergencies, in particular, it is often essential to make a diagnosis quickly, with whatever information is at hand, and without the possibility to conduct any further tests. Fifth, the trees are robust. In our analyses, fast-and-frugal trees (just as Naïve Bayes) outperformed profile memorization when tested under uncertainty, that is, when we predicted the class membership of new cases. Walsh, Einstein, and Gluck (2013) have stressed the importance of robustness in various areas, including decision making, and proposed a method for quantifying it (see also Gigerenzer & Brighton, 2009).

Given these advantages, it is not surprising that fast-and-frugal trees have been garnering much attention among researchers and practitioners alike. In the remainder of this Discussion, we first give an overview of selected contributions from research and practice that, in many instances, go hand in hand. We then address limitations of the present investigation and discuss avenues for further research.

6.1 Previous Research and Applications

The predictive accuracy of some of the heuristics studied here surpasses that of profile memorization, the model that is normative when evaluated within its training sample. But how do these heuristics compare with other benchmarks when it comes to predicting new cases? Laskey and Martignon (2014) compared the predictive accuracy of classification methods using the same 11 data sets used here. The heuristics they considered were Naïve Bayes and a tree they called ZigZag-val (a variant of our *B* tree); as benchmarks, they used Classification and Regression Trees (CART; Breiman, Friedman, Olshen, & Stone, 1984), which are simple in execution but whose construction often requires complex computations, and logistic regression, adapted to classification by introducing a classification threshold against which probability estimates were compared. For each of three sizes of the training sample (15%, 50%, and 90%), Naïve Bayes outperformed both logistic regression and CART. Even more surprisingly, when trained on small sets (15%), the ZigZag-val tree also outperformed logistic regression and CART. These simulations thus showed that, in an uncertain world, fast-and-frugal trees can reduce estimation error and have a competitive

advantage not only over profile memorization (as shown here) but also over other benchmarks.

In another study in which the predictive accuracy of heuristics was benchmarked, Woike, Hoffrage, and Petty (2015) included another important aspect in their simulations, namely environmental properties. Instead of using real-world data sets, these authors artificially generated data sets and could hence control features such as the number of cues and the skewness of the cues' validities. Simulated venture capitalists had to decide whether or not to invest money in a business plan that was described on a set of cues. The four strategies—which continuously updated their parameters based on feedback they received on the outcome of previously seen plans—were a linear model with equal weights and fast-and-frugal trees as heuristics, and logistic regression and CART as complex benchmarks. It turned out that strategies were differentially affected by environmental properties: The simple equal-weighting strategy was able to compete with the two complex benchmark strategies across most types of environments, and it even outperformed them when all cues were equally valid. In precisely this condition, the performance of the fast-and-frugal tree was miserable; however, the tree excelled, even outperforming all other strategies, when cue validities were quite skewed (more specifically, followed a geometric distribution). This set of simulations makes a strong case for ecological rationality: simple heuristics—here, equal weighting and fast-and-frugal trees—can perform very well, especially if their architecture mirrors the structure of information in their decision making environment (see also Todd et al., 2012).

An exciting area within the emerging field of research on fast-and-frugal trees addresses how trees can be constructed. The calculations of PPV and NPV necessary for the constructions we proposed here could all be performed on the back of a napkin. Instead of counting cases, as reflected in PPV and NPV, one could count the number of reasons why a positive or negative cue value may be misleading and use these numbers as alternative criteria for the ranking of cues (Cummins, 1995; Stenning, Martignon, & Varga, 2017). More complex but pertinent construction methods, notably methods based on signal detection theory, have been proposed by Luan, Schooler, and Gigerenzer (2011). These authors focused on trees with a relatively small number of cues and compared tree architectures in terms of performance across a range of simulated environments. They proposed that the *d'* measure, which takes both sensitivity and specificity of cues into account, be used to order cues. One may ask whether cues should be ordered based on a measure that takes all cases into account

if they are used within the tree to make classifications for only one type of case: Why should the classification of patients with positive cue values be influenced by properties of those with negative cue values, and vice versa? (We hasten to add that the same concern could be raised for our *A* tree.) Yet this critique can also serve as an argument *for* using *d'* (and *A*): These measures capture the quality of cues based on the total training sample, whereas the measurement of PPV and NPV is based only on the subset of cases with positive and negative cue values, respectively. Luan et al. (2011) compared the performance of tree algorithms against majority-of-cue models and models that are normative for fitting tasks. They found that fast-and-frugal trees reached performance levels close to that of majority-of-cue models and outperformed complex models for small sample sizes. In related work, Hozo, Djulbegovic, Luan, Tsalatsanis, and Gigerenzer (2016) have discussed principles for cue ordering along a decision criterion; this approach links to what we said above about the splitting profile and allows for comparison with classification models that include a threshold.

Providing an overview of the work on fast-and-frugal trees is beyond the scope of this article. To give an impression of the popularity they have acquired in recent years, however, we would like to mention a few, selected studies. Fast-and-frugal trees have been tested, notably, as predictors of depression (Jenny, Pachur, Williams, Becker, & Margraf, 2013). They have been praised as an ideal data structure for information technology implementation, particularly in pharmacogenomics, where complex data need to be used at point of care (Van Rooij et al., 2015). They have been proposed as suitable for social workers seeking to stabilize critical situations until fuller services are available (Taylor, 2016; see also Kirkman & Melrose, 2014). They allow swift decisions to be made in potentially dangerous situations where time is of the essence—their performance on past data indicates that their use might reduce the number of civilian casualties at military checkpoints (Keller & Katsikopoulos, 2016). They have been discussed as smartly designed strategies and as tools to boost decision makers' performance and autonomy (Grüne-Yanoff & Hertwig, 2016). They have been used by the Bank of England to assess bank vulnerability (Aikman et al., 2014; see also Haldane, 2015).

While fast-and-frugal trees can convey one or more of the advantages listed above to many practitioners in various domains, some of these practitioners may experience noteworthy tradeoffs. For instance, Wegwarth, Gaissmaier, and Gigerenzer (2009) observed an interesting dilemma faced by some doctors, who felt pressured to hide the simple way they

make decisions. Anticipating that their patients would not trust the fast-and-frugal approach, they preferred to pretend that their decisions stemmed from relatively complex and sophisticated strategies. We hope that analyses such as those reported in this article can help to legitimize the use of fast-and-frugal trees and other simple heuristics.

6.2 Limitations and Avenues for Further Research

Our results derived from analyses with 11 medical data sets. The sole focus on the medical domain is clearly a limitation; future studies should run similar tests with other data sets, from medicine but also beyond. Such analyses will not only test the reliability of our findings, but they will also provide a broader basis for studying the ecological rationality of classification strategies. Additional data sets and analyses will shed more light on how features that describe the data affect strategy performance. Data characteristics that could be considered include the number of cues, number of objects, distribution of the predictive values of the cues (overall level and skewness), measurement level and distribution of cue values, costs of obtaining cue value information, base rate of criterion values, costs of different classification errors, reversibility of decisions based on the classifications, and timing and quality of feedback on the criterion. Future research could study the ecological rationality of the strategies either analytically, with empirical data sets, or with artificially created data sets. A major advantage of real-world data sets is that they come with naturally existing information structures and that the results may be of interest to practitioners. The major advantage of artificial data sets is that they allow for perfect control over environmental properties and, in turn, for making causal claims about how such properties affect a strategy's performance. The work of Hertwig, Fanselow, and Hoffrage (2003), Luan et al. (2011), Martignon and Hoffrage (2002), and Woike et al. (2015), illustrates the value of studying purposely created environments (see also the discussion of representative sampling and formal sampling in Dhami, Hertwig, & Hoffrage, 2004).

As mentioned, it can be advantageous to have a toolbox of trees (rather than just one or two) at one's disposal. First, trees can be evaluated with respect to a range of measures, and different domains and applications may call for different measures. Second, the performance of trees depends on the environment, and different environments may favor different trees. The ecological rationality of fast-and-frugal trees is reflected in the observation that some of them are sensitive to environmental properties (note the different locations of *M*, *A*, and the *B* trees in Figure 3A–K) while others are not (e.g., the rake trees are always located in the same corner). Knowing how the classification performance of trees hinges on environmental

properties (such as the base rate of a disease) and knowing the properties of the data set at hand thus allows the practitioner to pick the best tree for a particular data set. Note that our Figure 5 does not reflect how the construction principles differ across environments with respect to accuracy. As is apparent from Figure 3, however, a tree should not be selected from the adaptive toolbox based on averages such as those reported in Figures 4 and 5; rather, it should be selected for a given data set.

In a similar vein, when selecting a tree from the toolbox, practitioners may consider not only the structure of information in the environment, but also which feature is most important to them (and clearly, the latter may depend on the former). If predictive accuracy is their main concern, then *M* and *A* should be in the consideration set; if it is sensitivity, then R_+ seems best; if it is specificity, then R_-. But which of these measures should they adopt? Predictive accuracy ignores the types of errors, ROC diagrams ignore the absolute frequencies of errors, and both ignore the costs of errors. Thus far, we have also disregarded the costs of errors. In fact, different errors have different costs, particularly in medical applications, but probably in most other domains as well (e.g., on psychological or economic dimensions). Practitioners may want to take these costs into account when selecting "their" strategy for "their" domain from the adaptive toolbox. Researchers could potentially assist practitioners by integrating costs of errors into tree construction principles, or by investigating how such costs affect classification thresholds (e.g., of profile memorization or Naïve Bayes; for a descriptive account, see Johnson, Blumstein, Fowler, & Haselton, 2013).

Another potential limitation is that fast-and-frugal trees are built only with dichotomous cues (Fokkema et al., 2015). Yet this limitation can also be seen as a research opportunity: How should one deal with non-binary cues? One possibility is to dichotomize them. This can be achieved in multiple ways. In our simulation, we have opted for median-splits, whereas Hozo et al. (2016) have discussed the possibility of using ROC curves to determine the cutting point. Another possibility is to construct trees that can handle non-binary cues as input. This path has been taken by Berretty, Todd, and Martignon (1999) and Blythe, Todd, and Miller (1999). Note that the algorithm proposed by these authors, *Categorization-by-Elimination*, can assign objects to one of several categories on the output side. For related work that constructs fast-and-frugal trees with more than two output categories, see Keller, Czienskowski, and Feufel (2014).

Finally, we would like to mention a project that has a great potential to aid researchers and practitioners aiming to tackle such open issues: Phillips, Neth, Woike and Gaissmaier (2017) are in the process of developing a software package for fast-and-frugal trees in R.

A new and exciting field is currently developing within the emerging science of heuristics (Gigerenzer et al., 2011): research on fast-and-frugal trees. We classify this domain as low risk and high potential for academic careers, and we look forward to its future developments.

Acknowledgments

We would like to thank Kathryn Laskey, Julian Marewski, and three anonymous reviewers for helpful comments on a previous version of this article, Susannah Goss for her careful editing, and the Swiss National Science Foundation (grant SNF 100014–140503/1) and the German Research Foundation (DFG, grant SPP 15-16, MA15-44-12) for their financial support.

7. REFERENCES

Aikman, D., Galesic, M., Gigerenzer, G., Kapadia, S., Katsikopoulos, K., Kothiyal, A., Murphy, E., & Neumann, T. (2014). *Taking uncertainty seriously: Simplicity versus complexity in financial regulation* [Financial Stability Paper, Bank of England No. 28]. London, United Kingdom: Bank of England.

Berretty, P. M., Todd, P. M., & Martignon, L. (1999). Categorization by elimination: Using few cues to choose. In G. Gigerenzer, P. M. Todd, & the ABC Research Group, *Simple heuristics that make us smart* (pp. 235–254). New York, NY: Oxford University Press.

Blythe, P. W., Todd, P. M., & Miller, G. F. (1999). How motion reveals intention: Categorizing social interactions. In G. Gigerenzer, P. M. Todd, & the ABC Research Group, *Simple heuristics that make us smart smart* (pp. 257–287). New York, NY: Oxford University Press.

Breiman, L. (1968). *Probability Theory*. New York, NY: SIAM.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, P. J. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth International Group.

Cestnik, G., Konenenko, I., & Bratko, I. (1987). Assistant-86: A knowledge-elicitation tool for sophisticated users. In I. Bratko & N. Lavrac (Eds.), *Progress in machine learning* (pp. 31–45). Wilmslow: Sigma.

Cummins, D.D. (1995). Naïve theories and causal deduction. *Memory and Cognition, 23*(5), 646-658. http://dx.doi.org/10.3758/BF03197265

Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics? In G. Gigerenzer, P. M. Todd, & the ABC Research Group, *Simple heuristics that make us smart* (pp. 97–118). New York, NY: Oxford University Press.

Daston, L. (1995). *Classical probability in the enlightenment* (Reprint edition). Princeton, NJ: Princeton University Press.

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., ... Froehlicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology, 64*(5), 304–310. http://dx.doi.org/10.1016/0002-9149(89)90524-9Dhami, M. K., & Ayton, P. (2001). Bailing and jailing the fast and frugal way. *Journal of Behavioral Decision Making*, *14*(2), 141-168. http://dx.doi.org/10.1002/bdm.371

Dhami, M. K., & Harries, C. (2001). Fast and frugal versus regression models of human judgement. *Thinking & Reasoning*, *7*(1), 5-27. http://dx.doi.org/10.1080/13546780042000019

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin, 130,* 959–988. http://dx.doi.org/10.1037/0033-2909.130.6.959

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning, 29*(2), 103–130. http://dx.doi.org/10.1023/A:1007413511361

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty:*

Heuristics and biases (pp. 249–267). Cambridge, United Kingdom: Cambridge University Press.

Fischer, J. E., Steiner, F., Zucol, F., Berger, C., Martignon, L., Bossart, W., Altwegg, M., & Nadal, D. (2002). Use of simple heuristics to target macrolide prescription in children with community-acquired pneumonia. *Archives of Pediatrics & Adolescent Medicine*, *156*(10), 1005-1008. http://dx.doi.org/10.1001/archpedi.156.10.1005

Fokkema, M., Smits, N., Kelderman, H., & Penninx, B. W. (2015). Connecting clinical and actuarial prediction with rule-based methods. *Psychological Assessment*, *27*(2), 636–644. http://dx.doi.org/10.1037/pas0000072

Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you*. New York, NY: Simon and Schuster.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143. http://dx.doi.org/10.1111/j.1756-8765.2008.01006.x

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650–669. http://dx.doi.org/10.1037/0033-295X.103.4.650

Gigerenzer, G., Hertwig, R., & Pachur, T. (Eds.). (2011). *Heuristics: The Foundations of adaptive behavior*. New York, NY: Oxford University Press.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684–704. http://dx.doi.org/10.1037/0033-295X.102.4.684

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506–528. http://dx.doi.org/10.1037/0033-295X.98.4.506

Gigerenzer, G., Todd, P. M., & the ABC Group (1999). *Simple heuristics that make us smart*. New York, NY: Oxford University Press.

Green, L., & Mehr, D. R. (1997). What alters physicians' decisions to admit to the coronary care unit? *The Journal of Family Practice*, *45*(3), 219–226.

Grüne-Yanoff, T., & Hertwig, R. (2016). Nudge versus boost: How coherent are policy and theory? *Minds and Machines*, *26*, 149–183. http://dx.doi.org/10.1007/s11023-015-9367-9

Haldane, A. G. (2015). Multi-polar regulation. *International Journal of Central Banking*, *11*(3), 385–401.

Hertwig, R., Barron, G., Weber, E. U., & Erev I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(8), 534–539. http://dx.doi.org/10.1111/j.0956-7976.2004.00715.x

Hertwig, R., Fanselow, C., & Hoffrage, U. (2003). Hindsight bias: How knowledge and heuristics affect our reconstruction of the past. *Memory, 11*, 357–377. http://dx.doi.org/10.1080/09658210244000595

Hertwig, R., Fischbacher, U., & Bruhin, A. (2013). Simple heuristics in a social game. In R. Hertwig, U. Hoffrage & the ABC Research Group, *Simple heuristics in a social world* (pp. 39–65). New York, NY: Oxford University Press.

Hertwig, R., Hoffrage, U., & Sparr, R. (2012). How estimation can benefit from an imbalanced world. In P. M. Todd, G. Gigerenzer, & the ABC Research Group, *Ecological Rationality: Intelligence in the World* (pp. 379–406). New York, NY: Oxford University Press.

Hertwig, R., Hoffrage, U., & the ABC Research Group (2013). *Simple heuristics in a social world*. New York, NY: Oxford University Press.

Hoffrage U., & Gigerenzer G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, *73*, 538–540. http://dx.doi.org/10.1097/00001888-199805000-00024

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition, 84*, 343–352. http://dx.doi.org/10.1016/S0010-0277(02)00050-1

Hoffrage, U., Krauss, S., Martignon, L., & Gigerenzer, G. (2015). Natural Frequencies improve Bayesian reasoning in simple and complex tasks. *Frontiers in Psychology, 6*(1473)*,* 1–14. http://dx.doi.org/10.3389/fpsyg.2015.01473

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science, 290,* 2261–2262. http://dx.doi.org/10.1126/science.290.5500.2261

Hozo, I., Djulbegovic, B., Luan, S., Tsalatsanis, A., & Gigerenzer, G. (2016). Towards theory integration: Threshold model as a link between signal detection theory, fast- and-frugal trees and evidence accumulation theory. *Journal of Evaluation in Clinical Practice*. http://dx.doi:10.1111/jep.12490

Jenny, M. A., Pachur, T., Williams, S. L., Becker E., & Margraf, J. (2013). Simple rules for detecting depression. *Journal of Applied Research in Memory and Cognition, 2*(3)*,* 149–157. http://dx.doi.org/10.1016/j.jarmac.2013.06.001

Johnson, D. D., Blumstein, D. T., Fowler, J. H., & Haselton, M. G. (2013). The evolution of error: Error management, cognitive constraints, and adaptive decision-making biases. *Trends in Ecology & Evolution, 28*(8)*,* 474–481. http://dx.doi.org/10.1016/j.tree.2013.05.014

Keller, N., Czienskowski, U., & Feufel, M. A. (2014). Tying up loose ends: A method for constructing and evaluating decision aids that meet blunt and sharp-end goals. *Ergonomics*, *57*(8), 1127–1139. http://dx.doi.org/10.1080/00140139.2014.917204

Keller, N., & Katsikopoulos, K. V. (2016). On the role of psychological heuristics in operational research; and a demonstration in military stability operations. *European Journal of Operational Research*, *249*(3), 1063–1073. http://dx.doi.org/10.1016/j.ejor.2015.07.023

Kirkman, E., & Melrose, K. (2014). *Clinical judgement and decision-making in children's social work: An analysis of the 'front door' system* [Research Report of the Department for Education, Behavioral Insight Team]. Retrieved from https://www.gov.uk/government/publications/clinical-judgement-and-decision-making-in-childrens-social-work

Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin, Germany: Springer.

Krivokapich, J., Child, J. S., Walter, D. O., & Garfinkel, A. (1999). Prognostic value of dobutamine stress echocardiography in predicting cardiac events in patients with known or suspected coronary artery disease. *Journal of the American College of Cardiology, 33*(3), 708–716. http://dx.doi.org/10.1016/S0735-1097(98)00632-9

Kurgan, L. A., Cios, K. J., Tadeusiewicz, R., Ogiela, M., & Goodenday, L. S. (2001). Knowledge discovery approach to automated cardiac SPECT diagnosis. *Artificial Intelligence in Medicine, 23*(2)*,* 149–169. http://dx.doi.org/10.1016/S0933-3657(01)00082-3

Laplace, P. S. (1951). *A philosophical essay on probabilities* (F. W. Truscott & F. L. Emory, Trans.). New York, NY: Dover. (Original work published 1814).

Laskey, K., & Martignon, L. (2014). Comparing fast and frugal trees and Bayesian networks for risk assessment. In K. Makar (Ed.), *Proceedings of the Ninth International Conference on Teaching Statistics*. Flagstaff, AZ: International Statistical Institute and International Association for Statistical Education. Retrieved from http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_8I4_LASKEY.pdf.

Lichman, M. (2013). UCI Machine Learning Repository [Repository]. Irvine, CA: University of California, School of Information and Computer Sciences. Retrieved from http://archive.ics.uci.edu/ml.

Luan, S., Schooler, L. J., & Gigerenzer, G. (2011). A signal-detection analysis of fast-and-frugal trees. *Psychological Review*, *118*, 316–338. http://dx.doi.org/10.1037/a0022684

Mangasarian, O. L., & Wolberg, W. H. (1990). Cancer diagnosis via linear programming. *SIAM News, 23*(5)*,* 1–18.

Martignon, L., & Hoffrage, U. (2002). Fast, frugal and fit: Simple heuristics for paired comparison. *Theory and Decision, 52*, 29–71. http://dx.doi.org/10.1023/A:1015516217425

Martignon, L., Katsikopoulos, K. V., & Woike, J. K. (2008). Categorization with limited resources. A family of simple heuristics. *Journal of Mathematical Psychology*, *52*, 352–361. http://dx.doi.org/10.1016/j.jmp.2008.04.003

Martignon, L., & Laskey, K. B. (1999). Bayesian benchmarks for fast and frugal heuristics. In G. Gigerenzer, P. M. Todd, and the ABC Research Group. *Simple heuristics that make us smart* (pp. 169–188). New York, NY: Oxford University Press.

Martignon, L., & Monti, M. (2010). Conditions for risk assessment as a topic for probabilistic education. Presented at the ICOTS 8, Ljubljana, Slovenia. *International Statistical Institute and International Association for Statistical Education*. Retrieved from http://iase-web.org/documents/papers/icots8/ICOTS8_8C2_MARTIGNON.pdf.

Martignon, L., Vitouch, O., Takezawa, M., & Forster, M. (2003). Naïve and yet enlightened: From natural frequencies to fast and frugal decision trees. In D. Hardman, & L. Macchi (Eds.), *Thinking: Psychological perspectives on reasoning, judgment, and decision making* (pp. 189–211). Chichester, United Kingdom: John Wiley and Sons.

Martignon, L. F., Katsikopoulos, K. V., & Woike, J. K. (2012). Naïve, fast, and frugal trees for classification. In P. M. Todd, G. Gigerenzer & the ABC Research Group,

Ecological Rationality: Intelligence in the world (pp. 360–378). New York, NY: Oxford University Press.

Massaro, D. (1998). *Perceiving talking faces*. Cambridge, MA: MIT Press.

Ore, O. (1960). Pascal and the invention of probability theory. *The American Mathematical Monthly*, *67*(5), 409–419.

Phillips, N. D., Neth, H., Woike, J. K., & Gaissmaier, W. (2017). FFTrees: An R package to create, visualize, and use fast and frugal decision trees. *Manuscript submitted for publication*.

Salzberg, S. (1988). Exemplar-based learning: Theory and implementation [Technical Report TR-10-88]. Cambridge, MA: Harvard University, Center for Research in ComputingTechnology, Aiken Computation Laboratory.

Spiegelhalter, D., & Gage, J. (2014). What can education learn from real-world communication of risk and uncertainty. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education*: *Proceedings of the 9th International Conference on Teaching Statistics (ICOTS9)*. Flagstaff, AZ: International Statistical Institute. Retrieved from http://icots.info/9/proceedings/pdfs/ICOTS9_PL2_SPIEGELHALTER.pdf

Stenning, K., Martignon, L., & Varga, A. (2017). Probability-free judgement: integrating fast-and-frugal heuristics with a logic of interpretation. *Decision*.

Taylor, B. J. (2016). Heuristics in professional judgement: A psycho-social rationality model. *British Journal of Social Work*, bcw084. http://dx.doi.org/10.1093/bjsw/bcw084

Till, C. (2014) Fostering risk literacy in elementary school. *EJME-Mathematics Education*, *9*(2), 83–96.

Todd, P. M., Gigerenzer, G., & the ABC Research Group (2012). *Ecological rationality: Intelligence in the world*. New York, NY: Oxford University Press.

Van Rooij, T., Roederer, M., Wareham, T., Van Rooij, I., McLeod, H. L., & Marsh, S. (2015). Fast and frugal trees: Translating population-based pharmacogenomics to medication prioritization. *Personalized Medicine*, *12*(2), 117–128. http://dx.doi.org/10.2217/pme.14.66

Verde, M. F., Macmillan, N. A., & Rotello, C. M. (2006). Measures of sensitivity based on a single hit rate and false alarm rate: The accuracy, precision, and robustness of d′, Az, and A'. *Perception & Psychophysics*, *68*(4), 643-654. http://dx.doi.org/10.3758/BF03208765

Waldmann, M., & Martignon, L. (1998). A Bayesian network model of causal learning. *Proceedings of the 20th annual conference of the Cognitive Science Society* (pp. 1102–1107). Madison, WI: Cognitive Science Society.

Walsh, M. M., Einstein, E. H., & Gluck, K. A. (2013). A quantification of robustness. *Journal of Applied Research in Memory and Cognition*, *2*(3), 137–148. http://dx.doi.org/10.1016/j.jarmac.2013.07.002

Wegwarth, O., Gaissmaier, W., & Gigerenzer, G. (2009). Smart strategies for doctors and doctors- in- training: heuristics in medicine. *Medical Education*, *43*(8), 721–728. http://dx.doi.org/10.1111/j.1365-2923.2009.03359.x

Woike, J. K. (2008). *Die paradoxe Rationalität einfacher Heuristiken* [The paradoxical rationality of simple heuristics] (Doctoral dissertation). Bochum, Germany: Ruhr-University.

Woike, J. K., Hoffrage, U., & Hertwig, R. (2012). Estimating quantities: Comparing simple heuristics and machine learning algorithms. In A. Villa, W. Duch, P. Erdi, F. Masulli, & G. Palm (Eds.), *Artificial Neural Networks and Machine Learning, ICANN 2012 – 22nd International Conference on Artificial Neural Networks, Lausanne, Switzerland, September 11–14, 2012, Proceedings Part II* (pp. 483–490). Heidelberg, Germany: Springer Verlag.http://dx.doi.org/10.1007/978-3-642-33266-1_60

Woike, J. K., Hoffrage, U., & Petty, J. S. (2015). Picking profitable investments: The success of equal weighting in simulated venture capitalist decision making. *Journal of Business Research*, *68*(8), 1705–1716. http://dx.doi.org/10.1016/j.jbusres.2015.03.030

Woolery, L., Grzymala-Busse, J., Summers, S., & Budihardjo, A. (1990). The use of machine learning program LERS-LB 2.5 in knowledge acquisition for expert system development in nursing. *Computers in Nursing, 9(6),* 227–234.

Tables

Table 1: The 11 data sets used to compare the performance of the three approaches to classification (profile memorization, Naïve Bayes, and various fast-and-frugal trees). Data sets are ordered according to the base rate of positive criterion values.

Panel in Figure 3	Name	Criterion	Base rate of positive criterion	Objects	Cues
A	Alcohol	>2.5 pints per day consumption	0.255	345	5
B	Echocardiogram	Survival after heart infarction	0.290	62	7
C	Diabetes	Diabetes	0.349	768	6
D	Breast Cancer	Malignant tumors	0.350	683	9
E	Heart Disease Hungarian	Heart disease	0.374	262	10
F	Heart Disease Cleveland	Heart disease	0.461	297	13
G	Horse Colic	Treated with surgery (yes/no)	0.587	218	6
H	Post-Operative	Decision to keep patient in hospital	0.721	86	8
I	Hepatitis	Survival of hepatitis patients	0.791	153	7
J	SPECT	Heart disease	0.794	267	22
K	Cardiac*	Heart disease	0.841	558	8

Note: All data sets are from the UCI Machine Learning Repository (Lichman, 2013), with the exception of Cardiac, which is taken from UCLA Department of Statistics (http://www.stat.ucla.edu/datasets). Original sources: (A) BUPA Medical Research Ltd., (B) Salzberg (1988), (C) National Institute of Diabetes and Digestive and Kidney Diseases, (D) Mangasarian & Wolberg (1990), (E & F) Detrano et. al. (1989), (G) see Lichman (2013), (H) Woolery, Grzymala-Busse, Summers, & Budihardjo (1990), (I) Cestnik, Konenenko, & Bratko (1987), (J) Kurgan, Cios, Tadeusiewicz, Ogiela, & Goodenday (2001), (K) Krivokapich, Child, Walter, & Garfinkel (1999).

Figures

Figure 1: Natural frequency trees in a task with one disease (breast cancer, B) and two conditionally independent tests: mammography (M) and ultrasound (U). "+" and "–" denote positive and negative test results, respectively. Panel A displays a causal tree; Panels B and C display the two diagnostic trees with all cues involved; Panels D and E each display a diagnostic tree for one of the cues, ignoring the other cue. Panels A–C are adapted from Martignon, Vitouch, Takezawa, and Forster (2003).

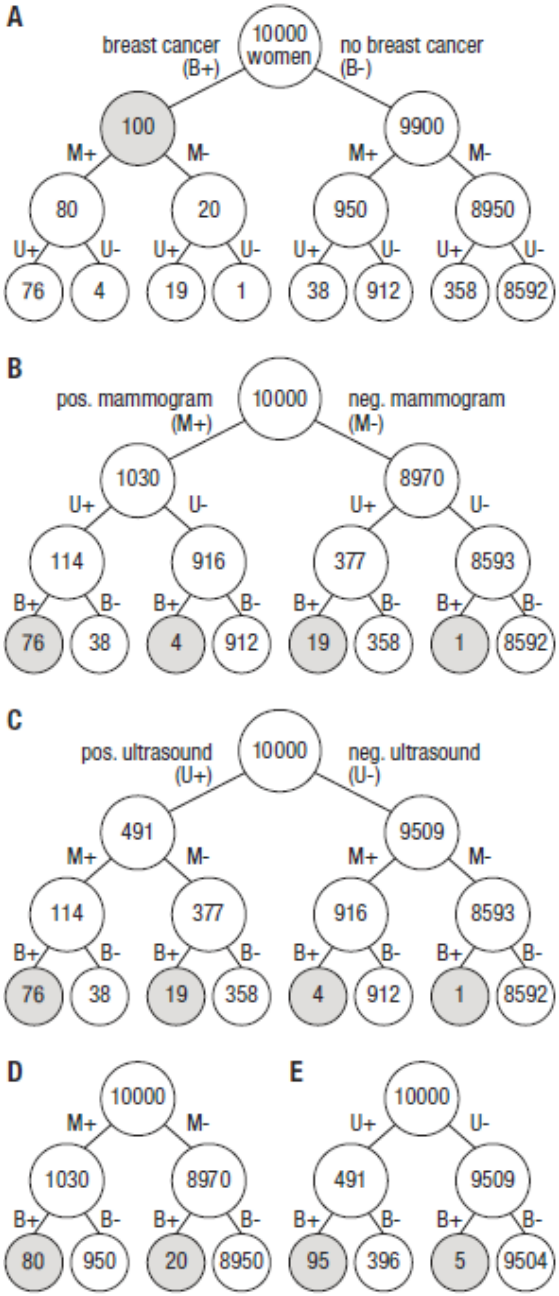

Figure 2, Panel A: Full natural frequency tree for the Green and Mehr (1997) data on 89 patients with severe chest pain. The goal is to determine whether these patients are at high or low risk for myocardial infarction. ST denotes a particular pattern in the electro cardiogram, CP denotes chest pain, OS denotes "at least one other symptom," "+" denotes present, and "–" denotes absent. Numbers in circles denote numbers of patients. Panel B: Fast-and-frugal classification tree obtained by pruning the natural frequency tree. Questions in rectangles specify which cues are consulted for each patient in the corresponding circle in Panel A. Depending on whether this cue value is positive or negative, either a new question is asked or the tree is exited and a classification decision is made (oval). The accuracy of these classification decisions is shown by the patient numbers below these oval exit nodes: The number of patients who actually had a heart attack is displayed in a gray circle; the number of those without heart attack, in a white circle. All patients to the left of the vertical bar in Figure 1B are classified as high risk: all patients to the right, as low risk (Figure adapted from Hoffrage et al., 2015).

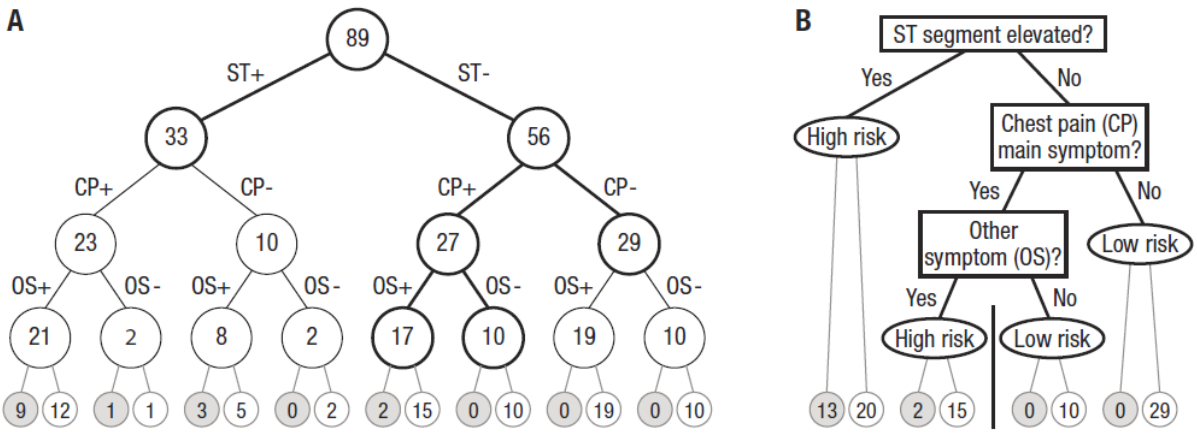

Figure 3: ROC curves for profile memorization classification (Natural Bayesian classification based on natural frequencies) and Naïve Bayes, and the corresponding data points for various fast-and-frugal trees, separately for the 11 data sets (Panels A–K) and averaged across all data sets (Panel L), with a training sample of 50% of the objects and performance tested on the remaining 50%. The points on the ROC curves for profile memorization and Naïve Bayes mark .1 increments on the classification threshold. The construction principles of the fast-and-frugal trees are denoted as Z_+, Z_-, and Z_0 (zigzag trees), R_+, R_- (rakes), B and B_1 (base-rate sensitive trees), M (maximum predictive value), and A (accuracy-based ranking).

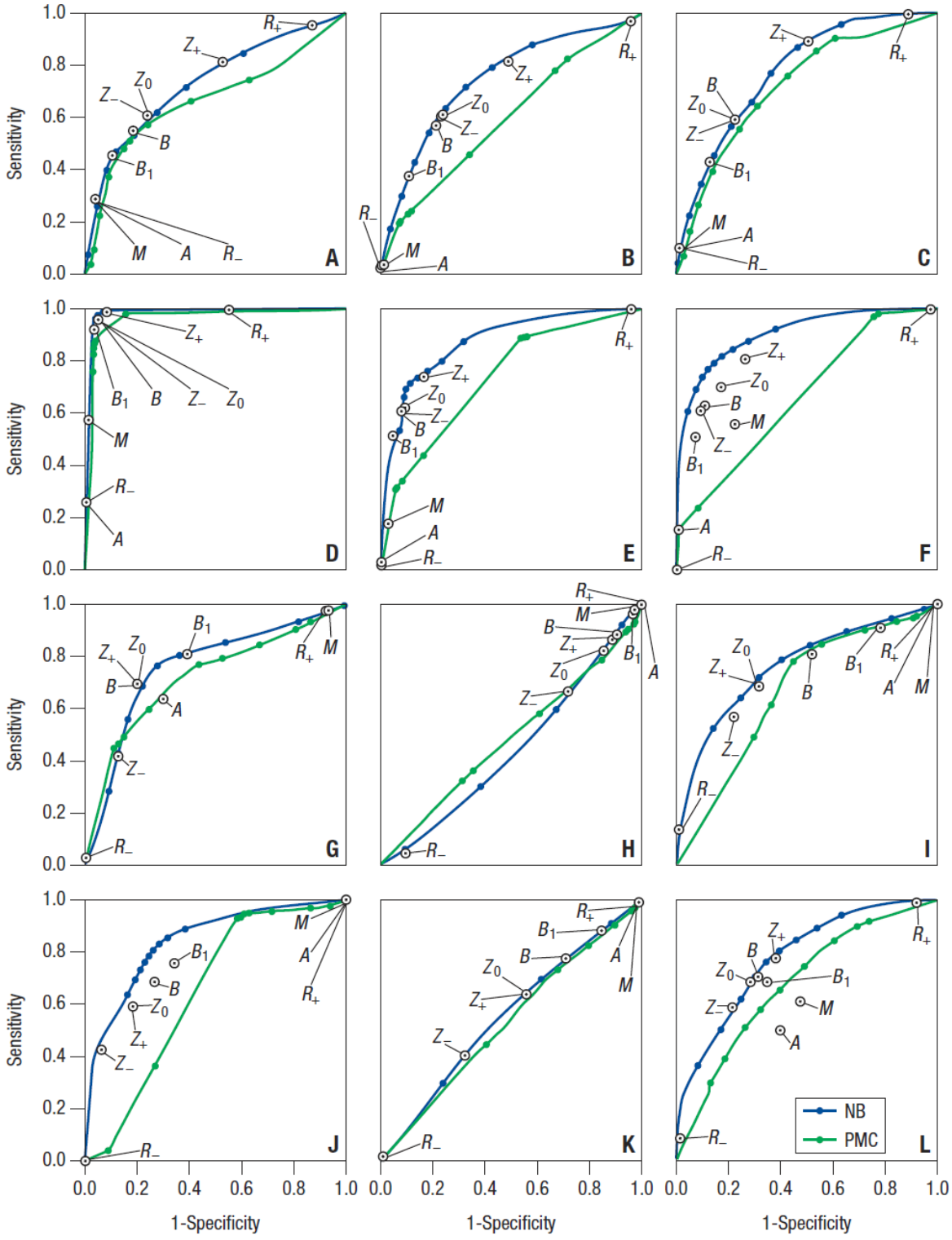
Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
E
F
G
H
I
J
K
L
NB
PMC

Figure 4: ROC curves for profile memorization classification, Naïve Bayes, and the corresponding data points for various fast-and-frugal trees across all 11 data sets, with size of training sample of 15%, 50%, and 85%, in Panels A, B, and C, respectively. The performance of the classifier was evaluated in the remaining 85%, 50%, and 15% of cases, respectively.

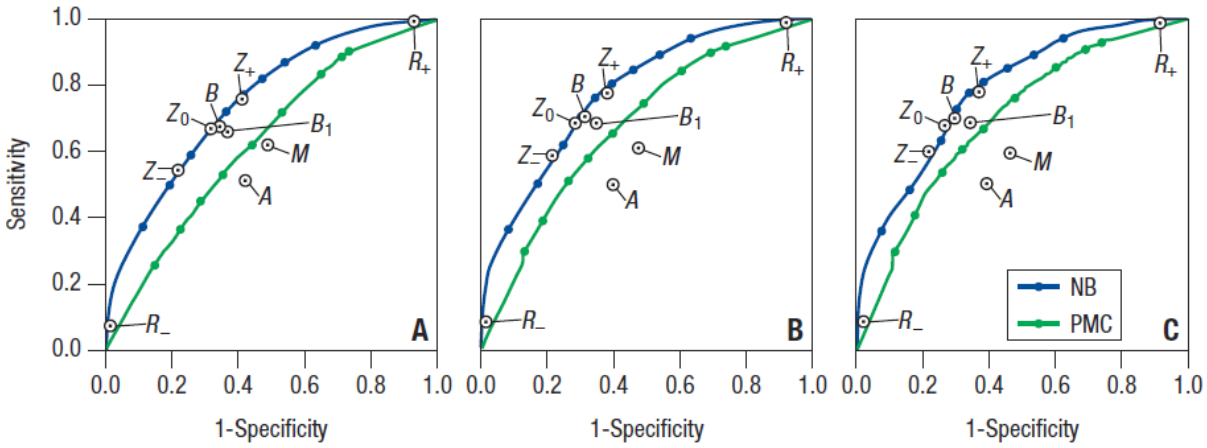

Figure 5: Predictive accuracy for profile memorization classification, Naïve Bayes, and various fast-and-frugal trees across all data sets, with size of training sample of 15%, 50%, and 85%. The performance of the classifier was evaluated in the remaining 85%, 50%, and 15% of cases, respectively.