Ambiguity, the certainty illusion, and the natural frequency approach to reasoning with inverse probabilities

John Fountain* and Philip Gunby

Department of Economics, University of Canterbury, New Zealand

People have difficulty reasoning with diagnostic information in uncertain situations, especially when an understanding and calculation of inverse conditional probabilities (Bayes theorem) is required. While natural frequency representations of inference tasks improve matters, they suffer from three problems: (1) calculation errors persist with a majority of subjects; (2) the representation suffers from an illusion of certainty that ignores ambiguity; and (3) the costs of repeatedly applying the representation to deal with imprecision and ambiguity in inference are prohibitive. We describe a user friendly, interactive, graphical software tool for calculating, visualizing, and communicating accurate inferences about uncertain states when relevant diagnostic test information (sensitivity, specificity, and base rate) is both imperfect and ambiguous in its application to a specific patient. The software is free, open-source, and runs on all popular PC operating systems (Windows, Mac, Linux).

Keywords: ambiguity; subjective expected utility; certainty illusion; inverse probabilities; choice under uncertainty; natural frequencies

1. Introduction

...clarity and excellence in thinking is very much like the clarity and excellence in the display of data. When principles of design replicate the principles of thought, the act of arranging information becomes an act of insight. (Tufté, 1997, p. 11)

People have difficulty reasoning with diagnostic information, especially understanding and calculating inverse conditional probabilities (Bayes’ theorem). One explanation for these difficulties is that the form in which information is presented in the inference task matters.¹

As Edwards and Gigerenzer (2003, p. 741) put it: ‘Some forms cloud minds, while others foster insight’.

Two formats have been extensively studied: natural frequencies and probabilities. A recent survey of nine studies involving over 600 subjects by Barbey and Sloman (2007, p. 246) found that only 5–20% of subjects are able to make accurate inductive
inferences when information and questions are posed as probabilities. This compares to 30–50% who are able to reason accurately when the same information and questions are posed as natural frequencies. Researchers in the field advocate a simple solution: use natural frequency formats when posing inference task problems rather than probability formats.

However, there are several problems using natural frequency representations in making inferences. First, the data from the survey above indicates that more than 50% of decision-makers in these studies still do not make correct inferences, even when presented with information as natural frequencies. Second, the paradigmatic question for researchers in this literature (Brase, 2002, 2008) is: on the basis of precisely given numerical information can people accurately compute a posterior probability? But is this problem – a precise question based on precise numerical information (however formatted) requiring precise numerical answers – really the problem decision-makers face in practical inference tasks? We think not. The representations for the inference tasks used in this research suffer from, and inadvertently foster, their own form of certainty illusion (Gigerenzer, 2002, Chapter 2).

Tukey’s Law is worth recalling here: ‘Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.’ In effect, while conventional natural frequency representations with precise whole numbers help improve one-shot inductive calculating and thinking, this same format inhibits the ability of people to make repeated and comparative inductive inferences (robustness checks and ‘what-if’ explorations) in the presence of imprecise informational inputs and other ambiguities.

What is the issue here? Real information related to making posterior inferences in practical situations – medical, educational, legal, scientific – is almost always ambiguous, imprecise, and/or incomplete to some extent. Usually there are several informational input measurements to an inference task (e.g. test sensitivity, test specificity, and/or base rates), each of which is assessed with only a limited degree of precision. Measurements on each of these may be imprecise for a task at hand for many reasons. There may only be a small number of observations. The samples may not be randomized nor carefully controlled. Ethical or cost constraints may result in samples that are at best field experiments with only partial knowledge of characteristics of the relevant population, or at worst convenience samples, self-selected self-reported surveys or sets of anecdotal case reports lacking in relevant conditioning information or controls. Moreover, how ambiguities and imprecision in information from multiple sources interact and propagate to create ambiguity in a final posterior inference output is anything but precise. Also, available imprecise measurements often require adjustment in some way to ‘apply’ the aggregate and incomplete information they contain to predict outcomes for the state variable of interest in an idiosyncratic specific case (e.g. will this patient develop this disease based on this particular test result?). It is possible, in principle, to express, calculate, re-calculate, and record hundreds of one-shot natural frequency representations, highlighting the variation in true and false positive/negative results, and present the posterior probability outputs in a series of written paragraphs, tables, or decision trees. But the pages of textual and numerical outputs emerging from this effort would make it very difficult to compare and contrast just how variations in the inputs are associated with variations in posterior inferences. Better methods for using and communicating risk information based on inherently ambiguous diagnostic test information are sorely needed.

196

J. Fountain and P. Gunby
This paper addresses problems with natural frequency representations of inference tasks by presenting an interactive, visual software tool that:

- eliminates calculation errors in practical inference tasks based on diagnostic test outcomes;
- facilitates interactive, comparative robustness checks on inferences ('what-if' reasoning) to guard against misplaced precision;
- can be used with either natural frequencies or single event probabilities, but builds on the demonstrated power of the former;
- has a user-friendly graphical interface involving only sliders, menu buttons, and input fields;
- is based on freely available, open source software capable of running on all widely used operating systems (Linux, Mac, Windows).

The remainder of the paper is organized in the following way. In Section 2 we explain the basic concepts and research designs outlined in the introduction and then argue the case for finding an inference task representation that explicitly recognizes the repetitive nature of inductive inferences in the presence of ambiguities. In Section 3, we describe our visual interface and show how it can be applied to the expression of ambiguities about relevant input information and the exploration of when and how much these ambiguities matter for drawing inferences from diagnostic tests in a practical medical situation. Section 4 concludes, and an Appendix explains the underlying theory behind the interface.

2. Basic concepts defined

First, we need to be clear about what is meant by natural frequency or probability formats for representing uncertainty in an inference task. Table 1, from Gigerenzer and Hoffrage (1995) and Gigerenzer (2002), illustrates how each format frames the ‘same’ inference problem. The standard probability format on the left of Table 1 is a textual representation of an inference task distinguished by (1) the use of the language and concepts of probability; (2) the expression of quantitative information in percentage or real number decimal form; and (3) the output of the inference task, which is a request to express uncertainty about a single-event inverse conditional

Table 1. Example of two standard formats.

<table>
<thead>
<tr>
<th>Standard probability format</th>
<th>Standard frequency format</th>
</tr>
</thead>
<tbody>
<tr>
<td>The probability of disease X is 20% for women at age 40 who participate in routine screening. If a woman has disease X, the probability is 80% that she will get a positive diagnostic test. If a woman does not have disease X, the probability is 30% that she will also get a positive diagnostic test. A woman in this age group had a positive diagnostic test in a routine screening. What is the probability that she actually has disease X?</td>
<td>20 out of every 100 women at age 40 who participate in routine screening have disease X. 16 of the 20 women with disease X will get a positive diagnostic test. 24 out of the 80 women without disease X will also get a positive diagnostic test. Here is a new representative sample of 100 women at age 40 who got a positive diagnostic test in routine screening. How many of these women do you expect to actually have disease X?</td>
</tr>
</tbody>
</table>
probability (e.g. what is the probability that this patient has this disease given this diagnostic test result?). The standard frequency format is also a textual representation but it avoids probability language altogether, using instead the language and concepts of sampling and reference classes. Quantitative information is expressed in whole number integers adding to simple round number totals (e.g. ‘100’ or ‘10,000’, not ‘3951’). The inference question is posed as a request for an expectation of a number of cases from a new sample from a well-defined reference class (e.g. how many out of a new sample of \( n \) women testing positive do you expect to have this disease, where \( n \) is usually a simple whole number divisible by 10 or 100).

The information in Figure 1 is used by some researchers to explain why most people find frequencies easier to work with than the probabilities. It contains two visual representational supplements to the standard format questions of Table 1: a labeled and highlighted decision tree diagram on the left, with frequency counts at the terminal nodes, and a symbolic algebraic template in conventional mathematical notation for inverse probability calculations on the right. The bold highlighted numbers 16 and 24 in the bottom row of the decision tree on the left identify all those cases with positive test diagnostics, 40 in total. Examining the terminal nodes of the decision tree, only 16 of those came from cases with the disease. Simple integer arithmetic can then be used to compute a ratio, \( \frac{16}{40} \), an acceptable answer in itself, or a percentage, 40%, or a decimal, 0.4, all indicating the fraction of those diagnosed positive who actually have the disease. By contrast, the algebra that uses standard probability notation and the series of computations involving products and ratios of decimal numbers on the right of Figure 1 is likely to frighten off (hence the sad face emoticon) anyone except a probability theorist or statistics student familiar with the algebraic version of Bayes Theorem and comfortable with numerical calculations.

 Appropriately trained authors, editors, and journal readers use the augmented representations of Figure 1 (absent the emoticons) as ‘mind tools’ to organize and communicate insights about computational complexity in the paradigmatic inductive reasoning tasks in Table 1. But research subjects are not provided with either the augmented visual representations (e.g. highlighted decision trees) when presented with their inductive reasoning tasks in standard formats (Table 1) or the training to

![Figure 1. Augmented frequency and probability representation formats.](image-url)
use them as cognitive scaffolds. Why not? It seems odd to exclude such supplementary material if, as Giggenzler and Hoffrage (1995, p. 387) state, one objective ‘. . . is to lead research on Bayesian inference out of the present conceptual cul-de-sac and to shift the focus from human errors to human engineering’. And, as we noted in the introduction, and argue in the remainder of this section, supplemental dynamic visual representations are absolutely necessary in order to deal with the problem of ambiguity in inductive inference.

The paradigmatic question for researchers in this area needs rethinking. Trying to ascertain if people can accurately compute a posterior probability from precisely given numerical information may be a misleading line of enquiry. Non-controversial, numerical diagnostic information is rarely if ever ‘given’ in the kinds of practical inference problems being studied in this research. In social and health sciences in particular, information available at reasonable cost can be expected to be partial, incomplete, imprecise, diverse in its quality and quantity, and possibly of dubious relevance and/or credibility. In these circumstances the real question is one of finding mind-tools (i.e. task representations) that enable and facilitate decision-makers to express, calculate and communicate inferences on the basis of ambiguous information. A useful representation should both permit and facilitate robustness checks, tentative and curiosity driven explorations, and what-if counterfactuals to discover the implications of ambiguities and imprecision. Repeated calculation of posterior (post-test) probabilities based on different assumptions should be easy to make and to compare with one another. As we show in the next section, the dynamic, interactive representational format we have developed enables anyone with an ability to manipulate a graphical slider to do both of these things well. Trying to perform the same tasks with serially repeated presentations of the paradigmatic questions as either natural frequencies or probabilities would be tedious, difficult, error prone, and probably confusing to most subjects no matter what their quantitative background and expertise.

The frequency or probability numbers ‘given’ in the paradigmatic questions as informational inputs – test specificity, test sensitivity, and base rates – are objects about which people will usually be uncertain. Neither of the standard formats recognize this problem of uncertainty about uncertainties, i.e. ambiguity, or the problem of ambiguity aversion (see Al-Nijjar & Weinstein, 2009; Klibanoff, Marinacci & Mukerji, 2005; Mukerji, 2000; Nau, 2007; Nehring 2009). Instead precise numbers are presented in the standard research paradigm as givens, non-contestable ‘truths’ to inform a precise inverse probability inference. An illusion of (second-order) certainty about stochastic relationships between outcome states and diagnostics is created in order to resolve a (first-order) uncertainty problem about outcome states and test diagnostics. Of course, the researchers comparing natural frequency and probability representations don’t advocate this sort of uncritical perception or thinking. But the two paradigmatic standard representational formats don’t do anything to prevent the illusion of (second-order) certainty, and, unwittingly, the precise numerical forms in which the questions are stated and answers are expected, combined with the cost of repeated calculations involved in robustness checks, cultivate it. Our dynamic interactive graphical format, on the other hand, is a mind-tool that builds on one of the key ideas in the rapidly developing research on ambiguity and ambiguity aversion: imprecise, but coherent, sets of probabilities, expressed as upper and lower bounds on relevant marginal or conditional probabilities.
There is a cost in cognitive effort in recognizing that there are uncertainties about uncertainties. Something new – a graphical representation for numerical information involving diagnostic information and conditional probabilities – has to be learned and experienced in order for it to be useful for understanding and communicating risks. But this learning cost is a fixed cost, and, once paid, the ‘mind-tool’ involved can be used repeatedly at low incremental effort cost. In a one-off situation where there are no ambiguities in the underlying diagnostic information and its relation to relevant output state measures, the dynamic graphical interactive representation probably isn’t worth the effort to learn. In other situations, where repeated calculations need to be made and compared, and where insights need to gained and explored as to how ambiguities in the underlying inputs to an inference interact and propagate through to inferences about output states, the effort cost to learn the graphical representation may well be worth it. For example, doctors, lawyers, or professional educators in repeated risk communication transactions predicting health state, legal state, or educational state outcomes for individual patients, clients, or students, may find it worthwhile to incur the learning costs. There is a tradeoff here: better quality of inferences under ambiguity – our interactive graphical method – comes at a (higher fixed, lower variable) cognitive effort cost compared with the (lower fixed, higher variable) cognitive effort costs of repeated serial application of standard inference questions in natural frequency formats.

An exact analogy from education in introductory economics may be helpful. One can ask students to calculate a competitive market price and quantity on the basis of a presentation of two forms of information: written word paragraphs with precise numbers embedded serially in the text or in tables of ‘given’ prices, quantities supplied, and quantities demanded in neatly labeled rows and columns. A precise question with a precise answer, in two formats. Can subjects do the calculations at all? Which format facilitates more accurate calculations? Academic economists all expect the tabular representation to improve decision-making compared with lengthy written word paragraphs containing demand and supply numbers. But virtually every elementary economics text and course doesn’t stop (and many don’t even start) with written word or tabular representations with precise numbers. Those representations are usually just steps in the process of developing the famous demand and supply graph of Alfred Marshall: two curves crossing in a two-dimensional diagram with price on the vertical axis and quantity on the horizontal axis. This diagram has a proven track record of facilitating insight and avoiding confusion in trying to understand and predict how competitive markets work, and especially so when trying to communicate basic principles of market interaction when underlying conditions of demand and supply are relatively vague and only imprecisely known. A demand and supply graph can be used to make precise statements and predictions about demanders, suppliers and their interactions, but that is not its primary role. Rather, the demand and supply graph functions as a visual heuristic that simultaneously focuses attention on relevant quantities, prices, and their interrelationship but also facilitates thinking about the impact of misspecifications of these variables and relationships: what will happen if income, or taxes, or prices of substitutes or complements, or technologies of consumption or production are different from what is presumed in the initial graph?

We rest our case for the need to incorporate principles of ambiguity and robustness into the representation of inductive inference tasks. The proof is really in the pudding, all explained in the next section.
3. Explaining and using the dynamic interface for the software

This section describes the dynamic interface for the interactive software tool we have developed. It is best read in conjunction with hands on use of the software. Instructions for downloading the software and installing the program needed to run it, Mathematica Player, can be found at http://uctv.canterbury.ac.nz/post/4/1132. Mathematica Player is a freely downloadable program made available by Wolfram Research to run the thousands of open-source visualizations available at the Wolfram Demonstration Project. No experience with Mathematica or any other programming language is necessary – all interactive user controls are in the form of user friendly sliders, menus, buttons, and text input fields.

Figure 2 is an annotated screenshot of the interface that you will see on opening the program. It has three interconnected parts. The left panel consists of user controlled diagnostic input variables. The right panel contains two outputs, one tabular, the other graphical, both derived from the input sliders in the left panel. A brief explanation of each of the panels in Figure 2 is followed by an example in Figure 3 of how decision-makers can use the software to vary the input parameters and observe the effect on diagnostic inferences.

We assume that a decision-maker, DM, is uncertain about the disease state and test outcome variables \((S,T)\) for a specific patient, labeled ‘\(A\)’. Will \(A\) have the disease? Will she have a positive test result? Will she have both, or neither, or some combination in between? That decision-maker might be \(A\) herself, or her clinician (doctor, nurse, counselor), or the two of them in consultation. The inference task is
to interpret and predict A’s risk of having the disease based on either a positive or negative test result, or no test at all, when sources of information and evidence relevant to predicting are imprecise and ambiguous.

The frequency information in the bottom row of the truth table in the top right panel is a way of expressing uncertainty about possible values for \((S, T)\) in the form of counts of cases in a hypothetical population (the size of which can be chosen by the menu buttons in the top left of the diagram). The initial default values in the table of Figure 2 assume a population of 100 cases, 16 of which are true positives, four of which are false negatives, 24 of which are false positives, and 56 of which are true negatives. Of course the table can also be used with the language of probability to express uncertainties, since scaling whole number frequencies (the column counts) by dividing by 100 or its multiples is a relatively transparent operation that produces a discrete probability mass function \(P(S, T) = (0.16, 0.04, 0.24, 0.56)\) on the four possible events.

The initial values for the frequency numbers in the table are controlled by the sliders and menu buttons on the left hand side of Figure 2. These sliders represent numerical inputs for conventional ways of expressing uncertainty about state variables \(S\) and the outcomes of diagnostic tests \(T\): test sensitivity (e.g. accuracy in detecting people with a disease), test specificity (e.g. accuracy in detecting people without a disease), and the base rate of the state variable (e.g. underlying rate of people with a disease). There are actually two rows of frequencies, corresponding to the two sets of sliders in the left panel, one a benchmark set of sliders. Having two sets of sliders and associated frequency representations is useful for comparing the implications of imprecise inputs into the inference process.

The graphical display in the bottom right panel of Figure 2 is bounded by a unit square, with variables in the \(x\)-direction indicating conditional and marginal probabilities for the state variable \(S\) and variables in the \(y\)-direction indicating conditional and marginal probabilities for the test variable \(T\). The graphic presents dynamically updated numerical values for two output variables, two posterior (post-test) probabilities for the state variable \(S\), as well as for three input variables, the test sensitivity and specificity and the state variable base rate. The posterior probability for the state variable \(S\) given a positive test outcome \(T = 1\), \(P(S|T = 1)\), is shown as a square box on the top \(x\)-axis of the diagram. The posterior probability for the state variable \(S\) given a negative test outcome \(T = 0\), \(P(S|T = 0)\), is shown as a square box on the bottom \(x\)-axis. Both of these conditional probabilities are joined by a solid line. As explained in the Appendix, this straight line is a constraint required by the laws of probability (a coherency constraint) on the relationship between the two end-point conditional probabilities and marginal probabilities (base rates) for the state and test variables. The dynamically updated numerical values for these posterior probabilities are calculated from Bayes theorem based on the values of the sensitivity and specificity of the test, as specified by the sliders in the left-hand panel of Figure 2. The sensitivity of the test, the probability for the test \(T\) being positive given that the disease state variable is \(S = 1\), \(P(T|S = 1)\), is shown as a circle on the right-hand side, \(y\)-axis of the diagram. The specificity of the test is the conditional probability \(P(T = 0|S = 0)\) for the test variable \(T\) being negative given the absence of the disease, \(S = 0\). On the graph, specificity \(P(T = 0|S = 0)\) is effectively indicated by the circle on the left-hand side \(y\)-axis, at a vertical height equal to \(P(T|S = 0)\), but its magnitude has to be measured down from the point \((0,1)\) on the \(y\)-axis, not up from the origin \((0,0)\). Both of the conditional probabilities \(P(T|S = 0)\), and \(P(T|S = 1)\),
are joined by a dashed line, another coherency constraint required by the laws of probability (explained in the Appendix). The vertical line in the figure through the bold triangle on the bottom x-axis at 0.2 is the base rate $P(S)$ of the disease state $S = 1$, set by the base rate slider in the left panel. It necessarily cuts the intersection of the two coherency constraints (the dashed and solid lines in the graph) – see the Appendix for why. The vertical height of that intersection point is the marginal probability or base rate of having a positive test $T = 1$, $P(T)$.

The specific numbers shown in Figure 2 are derived from the slider settings, which simultaneously determine the counts in the columns of the natural frequency table. The computer does all of the numerical calculations for the user. But the presence of the natural frequency table prevents the interface from being a ‘black-box’ technology for users. Straightforward whole-number arithmetic can be used along with the numbers in the natural frequency table to confirm (and discuss, in the case of multiple users) the measurements being expressed in any dynamic plot. For example, using Figure 2, the sensitivity slider set to 0.8 (80%) corresponds to the count of 16 cases with $S = 1$ and $T = 1$ in column 1 divided by $16 + 4 = 20$ the sum of the counts in columns 1 and 2 where the disease state is present, $S = 1$.

Of course, the numbers and pictures in Figure 2 on its own are only initial, default values for a dynamic interactive process. The major advantage of the software interface over the static natural frequency table is to facilitate the expression and exploration of ambiguities and robustness checks. Figure 3 shows the impact on posterior probabilities of decreasing the base rate from 20% to 5% when test sensitivity and specificity remain unchanged. As the base rate slider is manipulated, the position of the vertical line marking the base rate changes (with a more opaque line at the original base rate at a benchmark level for comparison). The corresponding posterior probabilities clearly decrease with this change in the base rate, and markedly so, from 40% to 12%, for the positive test-posterior probability $P(S|T = 1)$, but also for the negative post-test test-posterior probability, $P(S|T = 1)$, from 7% to 1%. It only takes a few moments playing with the base rate slider to convince even the most inexperienced user that (1) knowledge of the base rate is fundamentally important for posterior inferences based on diagnostic test results; (2) ambiguities about the size of the base rate imply corresponding ambiguities about the posterior inferences that can be made; and (3) imprecisely known or ‘ball park’ (interval) maximum and minimum base rates translate into corresponding imprecisely known or ‘ball park’ (interval) limits on both positive and negative posterior probabilities.

Not only are the levels of the new post-test probabilities instantly recalculated on the graph as the base rate slider is manipulated, but the relative difference between the two post test probabilities is also immediately transparent. In general, the difference between positive and negative posterior probabilities is a measure of the amount of information to be gained from actually doing a diagnostic test (see Bernardo & Smith, 1994, pp. 43–44). This difference is reflected in the slope of the solid line, with steeper lines indicating smaller differences between the two posterior probabilities and flatter lines indicating larger differences between the two posterior probabilities. The original (OLD) and the altered (NEW) solid lines are marked in the graph of Figure 3, with the NEW one, at a base rate of 5%, much steeper (implying much less to be learned from this diagnostic test) in comparison to the OLD line, corresponding to a base rate of 20%.
Take a closer look at this insight from the dynamic graph. One performs a diagnostic test with an expectation that the test result will change one’s prior or pre-test uncertainty $P(S)$, here 5%, significantly. The question ‘is the change in beliefs from pre-test to post-test enough to make the test worthwhile doing’ is just as important a question as ‘what should I infer from a specific – e.g. positive – test result’? The dynamic interface helps explore and resolve the first question as well as the second question. It is obvious from Figure 3 that the base rate $P(S) = 5\%$ for the disease state is bracketed by the positive (12%) and negative (1%) posterior probabilities for the disease. This is generally true because the base rate $P(S)$ is a weighted average of the two numbers for positive and negative posterior inferences (see the Appendix for an explanation). When the difference between those two posterior probabilities is small (the slope of the solid line large) both positive and negative posterior inferences will then be close to the pre-test base rate. If the information gain isn’t deemed to be very large, perhaps a costly and risky diagnostic test isn’t worth doing?

The test sensitivity and specificity values used for the calculations underlying Figures 2 and 3 are low, 80% and 70% respectively. What if the test was more (or less) sensitive or more (or less) specific, or if the test involves a tradeoff between sensitivity and specificity? These sorts of ‘what-if’ questions are quickly asked and answered using the software tool we have developed.

The software facilitates the expression of ambiguity in key underlying uncertainties as well as an exploration of the implications for posterior inferences of any ambiguities. A range for the base rate, the sensitivity, or the specificity is one
way of expressing a decision-maker’s ambiguity about a specific testing process. Perhaps there is some vital piece of clinical history missing for this patient, such as recent sexual activity in case of a possible herpes or AIDS infection investigation, or a missing family history of breast cancer in case of a breast cancer investigation, or the quality control or threshold values in the processes used to measure and calibrate the test sensitivity and/or specificity are suspect. Varying sliders one at a time permits the exploration and assessment of the marginal impacts of these changes on posterior beliefs as well as global impacts when there is more than one source of ambiguity. Answering these kinds of ‘what-if’ questions quickly using only conventional natural frequency representations would be tedious and error prone. Notice the caveat ‘only’ here. The trick is not to abandon natural frequency representations with whole number arithmetic but to augment their strengths using an interface that lowers the cognitive costs and errors of manual calculation and comparison. Our interface does just that.

4. Conclusion

The format that we present can be seen as a dynamic extension of Gigerenzer’s insight as to why natural frequency formats help people make better inferences from diagnostic information in uncertain situations: ‘the representation does part of the reasoning’ (Gigerenzer, 2002, p. 48). Our extension of static natural frequency representations follows Tufte’s principles for good visual displays: ‘When principles of design replicate the principles of thought, the act of arranging information becomes an act of insight’ (Tufte, 1997, p. 11, emphasis added). In a repeated, and repetitive, inference task environment our representation avoids sequences of error-prone intermediate manual calculations and data entry tasks, clarifies visually important systemic elements of the inference task, and enables decision-makers to focus on the important task at hand: representing and exploring uncertainty about relationships between diagnostic tests and state variables of interest. While we have concentrated on examples from the medical literature where the state variable is the presence or absence of a disease, the software interface is equally useful in educational, legal, and scientific contexts when the relationship between underlying informational inputs and posterior inferences is both uncertain and ambiguous. This simple yet flexible generic graphical tool avoids the illusion of (second-order) certainty that plagues the two standard (natural frequency or probability) representational formats, and has a solid foundation in the theory of imprecise probabilities and the developing research on ambiguity and ambiguity aversion.

Notes

3. Coherency of probabilities means satisfying the laws of probability at a formal level. Coherency of imprecise probabilities means that whatever imprecise and/or incomplete assessments of conditional or marginal probabilities are made are capable of being supported by a coherent probability distribution – even if that full distribution is not precisely assessed. See Lad (1996, chs. 2 and 3) or Walley et al. (2004) for a detailed exposition of the theory of imprecise probabilities.
4. This is because upward movements away from the origin on the vertical axis measure larger magnitudes for probabilities of the test being positive and the specificity itself is a
conditional probability for the test being negative. Since the two probabilities $P(T|S = 0)$ (shorthand for $P(T = 1|S = 0)$) and $P(T = 0|S = 0)$ must sum to 1, once we plot a point at height $P(T|S = 0)$ we can measure the size of the specificity $P(T = 0|S = 0)$ downwards by the amount $1 - P(T|S = 0)$ from the point $(0,1)$ on the $y$-axis.

5. Lad (1996) has a clear and insightful development of the inference issues associated with upper and lower limits on component ‘pieces’ of coherent probability distributions. See also Walley et al. (2004). For recent research on ambiguity and ambiguity aversion see Klibanoff (2005), Nau (2007), Al-Najjar and Weinstein (2009), and Nehring (2009).

References


Appendix

This Appendix outlines the role of the coherency constraints in the software interface, the solid and dashed lines in Figures 1–5. Consider first the solid line in Figure 3. From elementary probability theory we know that the marginal probability for $S$, the disease being present, is the sum of the joint probabilities:

$$P(S) = P(S = 1, T = 1) + P(S = 1, T = 0)$$

(A1)

Note that our notations uses $P(S)$ and $P(S = 1)$ interchangeably, but the shortened form will be used when the context permits.
Each joint probability on the right-hand side of equation (A1) can be expressed as the product of a marginal and a conditional probability:

\[ P(S) = P(S = 1 \mid T = 1)P(T = 1) + P(S = 1 \mid T = 0)P(T = 0) \]  

(A2)

Equation (A2) says that the marginal probability for \( S \) is a weighted average of two conditional probabilities \( P(S = 1 \mid T = 1) \) and \( P(S = 1 \mid T = 0) \), with the weights \( P(T = 1) \) and \( P(T = 0) = 1 - P(T = 1) \) being the marginal probability of the positive test result \( T \). If the conditional probabilities \( P(S = 1 \mid T = 1) \) and \( P(S = 1 \mid T = 0) \) are the fixed endpoints of the solid line, say \( P(S = 1 \mid T = 1) = a \) and \( P(S = 1 \mid T = 0) = b \) and \( P(S) = P(S = 1) \) is represented by \( x \) and \( P(T) = P(T = 1) \) by \( y \), equation (A2) becomes a simple linear equation in \( x \) and \( y \):

\[ x = a \cdot y + b \cdot (1 - y) \]  

(A3)

This is the equation for the solid line in Figure 3, with generic endpoints \( P(S = 1 \mid T = 1) = a \) and \( P(S = 1 \mid T = 0) = b \).

Similarly the marginal probability for \( T \), the test result being positive, is the sum of the joint probabilities:

\[ P(T) = P(S = 1, T = 1) + P(S = 1, T = 0) \]  

(A4)

Each joint probability on the right-hand side of equation (A4) can be expressed as the product of a marginal and a conditional probability:

\[ P(T) = P(S = 1 \mid T = 1)P(S = 1) + P(S = 1 \mid T = 0)P(S = 0) \]  

(A5)

Equation (A5) says that the marginal probability for \( T \) is a weighted average of two conditional probabilities \( P(T = 1 \mid S = 1) \) the sensitivity of the test, and \( P(S = 1 \mid T = 0) \), one minus the specificity of the test, with the weights \( P(S = 1) \) and \( P(S = 0) = 1 - P(S) \) being the marginal probability of the presence of the disease, \( S \). If the conditional probabilities \( P(S = 1 \mid T = 1) \) and \( P(S = 1 \mid T = 0) \), are the fixed endpoints of the dashed line, say \( P(S = 1 \mid T = 1) = c \) and \( P(S = 1 \mid T = 0) = d \) and \( P(S) \) is represented by \( x \) and \( P(T) \) by \( y \), equation (A5) becomes a simple linear equation in \( x \) and \( y \):

\[ y = c \cdot x + d \cdot (1 - x) \]  

(A6)

This is the equation for the dashed line in Figure 3, with generic endpoints \( P(S = 1 \mid T = 1) = c \) and \( P(S = 1 \mid T = 0) = d \).

Simple algebra and the rules of probability can show that the intersection of the solid and dashed lines of equations (A2) and (A5) occurs precisely at the marginal probabilities \( P(S) \) and \( P(T) \).