AMERICAN PSYCHOLOGICAL ASSOCIATION

Decision

ISSN: 2325-9965

2022, Vol. 9, No. 3, 285–319
https://doi.org/10.1037/dec0000160

The Likelihood Difference Heuristic and Binary Test Selection Given Situation-Specific Utilities

Jonathan D. Nelson[1, 2], Christine Rosenauer[2], Vincenzo Crupi[3, 4, 5], Katya Tentori[6, 7, 8], and Björn Meder[2, 9]

[1] Department of Psychology, University of Surrey

[2] Center for Adaptive Behavior and Cognition and iSearch Research Group, Max Planck Institute for Human Development, Berlin, Germany

[3] Department of Philosophy of Science and Education, University of Turin

[4] Center for Logic, Language, and Cognition, University of Turin

[5] Center for Mathematical Philosophy, Ludwig Maximilians University

[6] Department of Psychology and Cognitive Science, University of Trento

[7] Centre for Medical Sciences, University of Trento

[8] Center for Mind-Brain Science, University of Trento

[9] Department of Health, Health and Medical University Potsdam

Consider the task of selecting a medical test to determine whether a patient has a particular disease. Normatively, this requires taking into account (a) the prior probability of the disease, (b) the likelihood—for each available test—of obtaining a positive result if the medical condition is present or absent, respectively, and (c) the utilities for both correct and incorrect treatment decisions based upon each possible test result. But these quantities may not be precisely known. Are there strategies that could help identify the test with the highest utility given incomplete information? Here, we consider the Likelihood Difference Heuristic (LDH), a simple heuristic that selects the test with the highest difference between the likelihood of obtaining a true positive and a false-positive test result, ignoring all other information. We prove that the LDH is optimal when the probability of the disease equals the therapeutic threshold, the probability for which treating the patient and not treating the patient have the same expected utility. By contrast, prominent models of the value of information from the literature, such as information gain, probability gain, and Bayesian diagnosticity, are not optimal under these circumstances. Further results show how, depending on the relationship of the therapeutic threshold and prior probability of the disease, it is possible to determine which likelihoods are more important for assessing tests' expected utilities. Finally, to illustrate the potential relevance for real-life contexts, we show how the LDH might be applied to choosing tests for screening of latent tuberculosis infection.

This article was published Online First May 16, 2022.
Jonathan D. Nelson https://orcid.org/0000-0002-1956-6691
Christine Rosenauer https://orcid.org/0000-0002-6801-9520
Katya Tentori https://orcid.org/0000-0002-5968-9936
Björn Meder https://orcid.org/0000-0002-9326-400X
All authors contributed equally to this work. Jonathan D. Nelson, Christine Rosenauer, Vincenzo Crupi, Katya Tentori, and Björn Meder wrote the manuscript; Christine Rosenauer and Vincenzo Crupi derived the mathematical results; Katya Tentori researched the latent tuberculosis scenario.

We thank Gerd Gigerenzer, Rob Hamm, Laura Martignon, and Shenghua Luan for very helpful ideas and comments. This research was supported by grants NE 1713/2 to Jonathan D. Nelson, ME 3717/2 to Björn Meder, and Christine Rosenauer 409/1-2 to Vincenzo Crupi and Katya Tentori from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516) and by a Ministry of Education, University and Research (MIUR) grant Basic Research Investment Fund (FIRB) D11J12000470001 to Vincenzo Crupi.

Correspondence concerning this article should be addressed to Jonathan D. Nelson, Department of Psychology, University of Surrey, Stag Hill, University Campus, Guildford GU2 7XH, United Kingdom. Email: jonathan.d.nelson@gmail.com or j.d.nelson@surrey.ac.uk

Keywords: likelihood difference heuristic, medical decision-making, information search, diagnosis, utilities

In medical diagnosis, as in many other domains, it is rarely possible to obtain all potentially relevant information before deciding what course of action to take. Nonetheless, a limited number of carefully chosen tests can greatly facilitate diagnosis and treatment. But how should these tests be selected? When the relevant probabilities and utilities (i.e., benefits and harms of diagnostic decisions associated with tests' outcomes) are fully known, the utility-maximizing test can be determined using Bayesian decision theory (Savage, 1954). *Outcome-based* (or *situation-specific*) *utility functions* measure the extent to which an agent values the outcomes of actions, for instance, the benefits and harms associated with correct and incorrect treatment decisions. In practice, such utility functions can be derived from objective measures (e.g., monetary costs), subjective measures (e.g., side effects experienced by a patient or treatment benefits and harms as judged by doctors; Christensen-Szalanski & Bushyhead, 1981), or a mix of both (e.g., quality-adjusted life years; Weinstein et al., 2009). In this article, when we speak of utilities, unless denoted otherwise we mean such outcome-based, noninformational, utilities. Accordingly, the *expected utility gain* of a test refers to the average improvement in noninformational outcome-based utility. The best test is defined as the test that among available tests maximizes expected utility gain, given the applicable utility function.

While mathematically straightforward, this kind of analysis faces two challenges in real diagnostic situations. First, from a descriptive point of view, people may have difficulty weighting all probabilities and utilities and acting accordingly. This raises the question of how people intuitively evaluate tests and to what extent their test-selection strategies are sensitive to normative principles. Second, a precise quantification of appropriate utility functions is often challenging. Difficulties include measurement issues and comparability of subjective valuations, incommensurability of different metrics, and diverging views on treatment benefits and harms among multiple stakeholders (e.g., patients, health care providers, policy makers). Thus, attempts to derive appropriate utilities for evaluating medical tests are both important and notoriously difficult.

One proposal to circumvent these difficulties is to evaluate medical tests using *epistemic* (or *pure information*) utility functions (Benish, 1999, 2003; Good & Card, 1971). Such utility functions are also sometimes called *Optimal Experimental Design* (*OED*; Nelson, 2005) or *Optimal Data Selection* (Oaksford & Chater, 1994) models of the value of information. Epistemic utility functions consider the probabilities of diseases and test outcomes but disregard any further utilities associated with subsequent treatment decisions. Within this perspective, test selection should be determined solely by tests' ability to reduce epistemic uncertainty about the true state of nature, such as whether a patient has a particular medical condition or not. *Information gain* (Lindley, 1956), which is based on the expected reduction in Shannon's (1948) entropy, *Bayesian diagnosticity* (Good, 1950; Good & Card, 1971), and *probability gain* (Baron, 1985; Baron et al., 1988) are examples of epistemic utility functions.

Under what conditions—if any—can pure-information test-selection strategies identify the highest utility test? We analyze the *Likelihood Difference Heuristic* (LDH; Slowiaczek et al., 1992), a simple strategy that is applicable to the selection of binary tests (e.g., medical tests that can have only positive or negative outcomes) in a binary hypothesis space (e.g., disease vs. no disease, where the possible actions are to treat or not to treat the disease). The LDH requires only two pieces of information for each test, namely the likelihood that the test is positive when the disease is present (the true positive rate) and the likelihood that the test is positive when the disease is absent (the false positive rate). Both quantities are typically available for routine medical tests. The LDH then deterministically selects the test with the highest likelihood difference, ignoring any further available information, a feature shared with other heuristics from the judgment and decision-making literature (Gigerenzer & Gaissmaier, 2011). Because of its simplicity, the LDH can be used even when the prior probability of the disease or the outcome-based utilities are completely unknown. Of course, the fact that the LDH can be used does not demonstrate that the LDH—or any heuristic—would be sensible to use in any particular test-selection situation. Theoretical analyses (Nelson, 2005) show that the LDH invariably selects tests in accordance with *impact*, an epistemic utility function that quantifies belief change as the average absolute difference from prior to posterior probabilities. However, the LDH has not yet been studied with respect to outcome-based utility functions.

Human Inquiry and Test Selection

Carefully deciding which information to acquire is a central ability in many domains. Early research on how humans intuitively select queries was inspired by Popper's (1959) philosophy of science and the idea that one should seek out potentially disconfirming evidence (Wason, 1966, 1968). More recent approaches conceptualize human query selection as probabilistic inductive inference, where the goal is not to obtain potentially disconfirming information for a single hypothesis, but rather to acquire information in order to discriminate among multiple hypotheses (Coenen et al., 2019; Crupi et al., 2018; Gureckis & Markant, 2012; Meder et al., in press). This approach, sometimes referred to as the OED framework, has theoretical roots in Chamberlin's (1890) Method of Multiple Working Hypotheses, with instantiations in information theory (Lindley, 1956; Shannon, 1948), Bayesian philosophy of science (Good, 1950), and automatic Bayesian experiment design (Myung & Pitt, 2009).

In the cognitive and decision sciences, OED models are widely used as normative benchmarks against which human information acquisition can be considered. Domains studied include hypothesis testing (Austerweil & Griffiths, 2011; Crupi et al., 2009; Oaksford & Chater, 1994); eye movements for perception (Najemnik & Geisler, 2005), concept formation (Nelson & Cottrell, 2007), and reading (Legge et al., 1997); categorization (Markant & Gureckis, 2014; Nelson et al., 2010); causal induction (Bramley et al., 2015; Steyvers et al., 2003); the neural value of obtained or expected information (Filimon et al., 2020; Nakamura, 2006); eyewitness identification (Wells & Lindsay, 1980); medical diagnosis (Baron et al., 1988; Benish, 1999); and children's behavior on 20 questions games (Meder et al., 2019; Nelson et al., 2014; Ruggeri & Lombrozo, 2015). Most of this work (Coenen et al., 2019) suggests that at least to a first approximation, the OED framework provides a good basis for understanding human behavior. It is important to remember, however, that in all of the above tasks, the goals are purely epistemic, and the OED models employed disregard any benefits and costs associated with subsequent decisions based upon the obtained information.

However, in many domains such as medical diagnosis, the goals of the searcher are not purely epistemic. Rather, information acquisition is a means to an end, for instance, to decide about alternative medical treatments based on a test outcome. From a normative perspective, in such cases, the outcome-based utilities, and not only the informational value of tests, should be considered in test selection (Schlaifer & Raiffa, 1961). This is important because optimizing a particular epistemic utility function (e.g., using information gain or probability gain to select tests) will not necessarily identify the test with the highest utility gain, given the applicable rewards and costs (Markant & Gureckis, 2012; Meder & Nelson, 2012).

Can people adapt their test-selection strategies to different kinds of reward structures? In comparison to the great deal of research on human test selection in tasks where the goals are purely epistemic, little is known about people's ability to consider outcome-based utilities in information acquisition. It is important to note that in Signal Detection Theory (Green & Swets, 1966), there is a rich body of theoretical work on how expected reward should be considered when making classification decisions. Many experiments have investigated the circumstances under which people can adapt their decision thresholds to the rewards and costs associated with different categorization decisions (e.g., Maddox, 2002; Maddox & Bohil, 1998, 2003; Trommershäuser et al., 2003a, 2003b). However, that work investigates the decisions that are made after information is at hand and does not address the questions about whether, or how, people take outcome utilities into account when selecting tests to conduct. Below, we briefly review three articles that do address human test selection given outcome-based utilities.

Baron and Hershey (1988) used hypothetical medical diagnosis scenarios. Test selection was broadly sensitive to relevant variables, such as the varying costs of false-positive versus false-negative errors, but also showed deviations from normative principles. Choices of tests often reflected utilities qualitatively, for instance, in a preference to select tests that minimized the most harmful kind of errors.

Meder and Nelson (2012) used a probabilistic classification task, where experience-based learning was used to convey the relevant probabilities. In a subsequent search phase, participants had to choose between two queries before making a categorization decision. Varying monetary rewards were associated with different kinds of correct classification decisions; for instance, a correct Category A classification may have been worth 10 times as much as a correct Category B classification. The test that maximized classification accuracy was in these cases different from the test that maximized average payoffs. Participants' test-selection strategies were influenced by the applicable payoffs during the learning phase. However, if the payoffs were changed for the test phase, participants were not able to take this into account. Moreover, if probability information was conveyed using words and numbers, rather than experientially, neither utility gain nor probability gain could explain test-selection behavior.

Markant and Gureckis (2012) also studied search under situation-specific reward conditions, where participants had to identify different shapes on a game board by sampling targets on a grid. Their task involved both test costs (i.e., sampling itself was costly) and costs for specific kinds of errors. The task was designed in such a way that there was a conflict between obtaining information to minimize overall costs and selecting queries in accordance with information gain. In two experiments, searchers' behavior was better accounted for by information gain than utility gain, lending support to the hypothesis that people often focus on reducing epistemic uncertainty even if situation-specific utilities apply. Thus, the psychological literature to date, sparse though it is, suggests that people have at best very limited ability to use task-specific payoff structures into account when choosing which tests to conduct.

The LDH, known also as the *feature-difference heuristic* (Nelson, 2005) has to our knowledge not yet been studied, either theoretically or empirically, in medical test selection or other domains with outcome-based utilities. However, behaviors consistent with the LDH have been reported in various experimental tasks with purely epistemic goals. One such task is the Planet Vuma scenario, in which participants have to rate the usefulness of a set of binary-outcome features in order to categorize fictitious aliens into one of two species (Nelson, 2005; Skov & Sherman, 1986; Slowiaczek et al., 1992). Another task involves more naturalistic crime scenarios, in which participants select one of two queries they find most useful for an investigation (Liefgreen et al., 2020).

The use of likelihood subtraction has also been documented in the literature on belief revision. In particular, it is a nonnormative strategy that can lead to base rate neglect (e.g., Domurat et al., 2015; Gigerenzer & Hoffrage, 1995; McDowell et al., 2018). Both laypeople and physicians have been reported (Hoffrage & Gigerenzer, 1998) to estimate the positive predictive value of diagnostic problems by computing the difference between the sensitivity of a test, $P(e|h)$, and the false-positive rate of the test, $P(e|\neg h)$. The likelihood subtraction algorithm is also a prominent model in covariation assessment (often referred to as ΔP model), both in the Bayesian confirmation framework, where it represents a specific measure of evidential support (Nozick, 1981), and in the causal induction literature, where it corresponds to the probabilistic contrast model (Cheng & Novick, 1990).

To summarize, use of likelihood differences have been observed in empirical research on test selection, belief updating, covariation assessment, and causal induction. This suggests that the likelihood difference is a signal that people could use in a wide range of circumstances. In this article, we explore the circumstances, in situations with outcome-based payoff utilities, under which it could be sensible (or not) to use the LDH when choosing among binary tests to conduct.

Goals and Scope

We provide a formal analysis of the LDH as a strategy for selecting tests in situations where outcome-based utilities apply (e.g., medical diagnosis). We focus on the simplest possible test selection scenario, which is characterized by the option to conduct one binary test (i.e., a test that will have a negative or positive result), after which it is necessary to decide which course of action to take (i.e., to treat a patient or not). We use a simple medical diagnosis scenario to explain our mathematical results. However, our results apply without loss of generality to all situations with binary tests, hypotheses, and decisions. An analogous situation would be deciding what piece of information is most important for deciding whether an accused person is guilty or not guilty, in criminal law.

Our analyses are based on mathematically well-defined probabilities and utilities. What utilities are appropriate for a given situation or for a particular person (e.g., a patient or a doctor) is an important question, but is not part of our analyses. Similarly, we will not take into account the costs of the tests per se, but only the outcome-based utilities resulting from correct or incorrect treatment decisions. Nor will we consider the implications of planning a sequence of tests, although the best first test in a sequence can be different from the best single test if just one test can be conducted (Geman & Jedynak, 2001; Hyafil & Rivest, 1976; Meder et al., 2019; Meier & Blair, 2013; Nelson et al., 2018).

The following sections of this article are structured as follows:

- In "Determining the Expected Utility Gain of a Test", we introduce relevant definitions and illustrate how to compute a test's expected utility, given the applicable probabilities and utility function.
- In "Utilities Influence Tests' Relative Value Only via the Therapeutic Threshold", we then show that an outcome-based utility function influences tests' relative utility only via a single number, the *therapeutic threshold*. Introduced by Pauker and Kassirer (1975), the therapeutic threshold is the probability of disease above which the best (utility-maximizing) course of action is to treat a patient and below which the best course of action is not to treat.
- In "Analytic Results on the Likelihood Difference Heuristic", we analyze the LDH mathematically. We prove that, whenever there is maximal need for information, the LDH is guaranteed to select the test with the highest expected utility. The condition of maximal need for information applies when the probability of the disease equals the therapeutic threshold, such that a utility-maximizing decision maker would be indifferent among the available actions (i.e., to treat or not to treat). Further analyses apply to situations where the prior probability of disease and the therapeutic threshold are not equal. We show analytically that depending on whether the probability of disease is above or below the threshold, the utility of a test is either a function of the likelihood of obtaining a positive test result given the presence of the disease (the test's sensitivity) or of the likelihood of obtaining a negative test result given the absence of the disease (the test's specificity).
- In "The Bigger Picture: Simulation Results and Test Selection Based on OED Models", we show via counterexamples and simulations that prominent OED models are suboptimal under conditions of maximal need for information. We also show that the LDH is reasonable to use somewhat beyond these conditions, given that its performance degrades only gradually as we move away from the situation where the prior probability of disease exactly matches the therapeutic threshold. Interestingly, the situations in which the LDH performs well highly overlap with situations in which the choice of test to conduct especially matters.
- In "A Real-World Example: Latent Tuberculosis Testing", we consider the LDH in the context of latent tuberculosis diagnosis. This allows us to show the potential relevance of the LDH to a real-life screening setting as well as to discuss some limitations of our analyses.
- In the General Discussion, we conclude by highlighting key issues for future research on human information acquisition and test selection.

Determining the Expected Utility Gain of a Test

We next define our terminology and describe how to calculate the expected utility gain of a test, that is, how much utility can be gained on average from carrying out the test, compared to making a treatment decision without conducting the test. This calculation is based on the prior probability of the disease and the test characteristics (i.e., likelihood of obtaining a positive test result, given that the disease is present or absent, respectively) as well as the applicable outcome-based utilities. We show that the utilities associated with different kinds of correct and incorrect decisions influence a test's expected utility gain only via a single number, the therapeutic threshold (Pauker & Kassirer, 1975), the probability of disease above which it would be best to treat the patient, and below which it would be best to not treat the patient, if no further information could be acquired before deciding whether or not to treat.

Terminology

Table 1 gives an overview of the terms and definitions that we use throughout this article. We use the nomenclature of Bayesian decision theory, but related terms and concepts are used in other fields, in particular in the medical and Signal Detection Theory literature.

Probability Model

Let $H = \{h, \neg h\}$ be a binary random variable (hypothesis space) associated with a particular disease, where h denotes the hypothesis that a person has the disease and $\neg h$ denotes the hypothesis that she does not have the disease. The probability $P(h)$ is the *prior probability* of the disease, with $P(\neg h) = 1 - P(h)$. We assume that $0 < P(h) < 1$, meaning that it is a priori uncertain whether the person has the disease or not. Another term from the medical

Table 1

Terminology, Definitions, and Related Terms From the Medical Decision-Making and Signal Detection Literature

Term	Definition	Explanation	Related terms
Prior probability	$P(h)$	A priori probability of the disease being present (with $P(\neg h) = 1 - P(h)$)	Base rate, prevalence of disease, pretest probability
Diagnostic test	E	A test labeled such that a positive outcome e is positively associated (if at all) with disease, that is, $P(h\|e) \geq P(h)$ and $P(h\|\neg e) \leq P(h)$	
Likelihoods	$P(e\|h)$	Likelihood of positive test given disease	True positive rate, sensitivity $sens(E)$
	$P(e\|\neg h)$	Likelihood of positive test given no disease	False-positive rate, $1 - spec(E)$
	$P(\neg e\|h)$	Likelihood of negative test given disease	False-negative rate, $1 - sens(E)$
	$P(\neg e\|\neg h)$	Likelihood of negative test given no disease	True negative rate, specificity $spec(E)$
Posterior probabilities	$P(h\|e)$	Posterior probability of disease given positive test	Positive predictive value
	$P(\neg h\|e)$	Posterior probability of no disease given positive test	
	$P(h\|\neg e)$	Posterior probability of disease given negative test	
	$P(\neg h\|\neg e)$	Posterior probability of no disease given negative test	Negative predictive value
Likelihood difference	$\lambda(E)$	Difference in outcome likelihoods of a diagnostic Test E, $P(e\|h) - P(e\|\neg h)$	Difference between the test's true positive rate and false-positive rate
Utilities	u_{tp}	Utility of treating a person with a true positive test	Utility for a "hit"
	u_{fp}	Utility of treating a person with a false-positive test	Utility for a "false alarm"
	u_{fn}	Utility of not treating a person with a false-positive test	Utility for a "miss"
	u_{tn}	Utility of not treating a person with a true negative test	Utility for a "correct rejection"
Utility of test result	$u(e)$	Utility gained from a positive test result	
	$u(\neg e)$	Utility gained from a negative test result	
Expected utility of test	$eu(E)$	Average utility gained from Test E. If $eu(E) > 0$, then E is a useful test. The best test is the test that has highest expected utility gain among available tests.	
Therapeutic threshold	t_x	Probability above which treating the patient has higher expected utility than not treating, given the applicable utility function	Threshold probability, decision threshold

literature is the *pretest probability*, denoting the probability of disease before a test is conducted. A frequentist estimate for the prior probability of a disease, absent any individuating information for a particular patient, is the proportion of people in a given reference class that have the disease, often referred to as the disease's *base rate* or *prevalence*.

We use $E = \{e, \neg e\}$ to denote a *diagnostic test*, another binary random variable. We only consider tests for which the outcome is probabilistic, that is, $0 < P(e) < 1$ and $0 < P(\neg e) < 1$, which means that there is some chance of a positive result and some chance of a negative result. In our scenario, e corresponds to a *positive test result*, and $\neg e$ corresponds to a *negative test result*. The test results are labeled such that $P(h|e) \geq P(h)$ and $P(h|\neg e) \leq P(h)$. In other words, a positive test result increases (or does not change) the probability of the disease, while a negative test result decreases (or does not change) the probability of the disease. Thus, we do allow for nondeterministic yet uninformative tests, for which no test result changes the probability of the disease.

The conditional probabilities of test results e and $\neg e$ under true states h and $\neg h$ are the *likelihoods*. The two likelihoods involving positive test results are $P(e|h)$, the conditional probability of a positive test result given that the disease is present, and $P(e|\neg h)$, the conditional probability of a positive test result given that the disease is absent. Frequentist estimates of these probabilities are the proportion of people who have the disease who obtain a positive test result (*true positive rate*) and the proportion of people who do not have the disease yet nonetheless have a positive test result (*false positive rate*). In the medical literature, the true positive rate is often called the *sensitivity* of Test E, $sens(E)$.

The likelihoods involving negative test results are $P(\neg e|h)$, the conditional probability of a negative test result given that the disease is present, and $P(\neg e|\neg h)$, the conditional probability of a negative test result given that the disease is absent. Frequentist estimates of these probabilities are the proportion of people who do have the disease with a negative test result (*false negative rate*), and the proportion of people without the disease who correctly obtain a negative test result (*true negative rate*), respectively. In the medical literature, the true negative rate is often called the *specificity* of Test E, $spec(E)$.

If we administer a Test E, we get a positive test result e with probability $P(e) = P(h)P(e|h) + P(\neg h)P(e|\neg h)$, and a negative test result $\neg e$ with probability $P(\neg e) = P(h)P(\neg e|h) + P(\neg h)P(\neg e|\neg h)$. The *posterior probability* of disease given a positive test result, $P(h|e)$, and the posterior probability of absence of disease given a negative test result, $P(\neg h|\neg e)$, can be calculated using Barnard and Bayes (1763/1958) theorem, that is, $P(h|e) = P(e|h)P(h)/P(e)$ and $P(h|\neg e) = P(\neg e|h)P(h)/P(\neg e)$. In frequentist terms, $P(h|e)$ is the proportion of people with positive results that actually have the disease, and $P(\neg h|\neg e)$ is the proportion of people with negative test results who do not have the disease. In the medical literature, $P(h|e)$ is often called the *positive predictive value*, and $P(\neg h|\neg e)$ the *negative predictive value* of the test.

Utilities

In many real-world medical situations, in addition to administering a test and then updating the probability of the patient having the disease or not (i.e., computing the posterior probabilities of interest), a decision of whether or not to treat the patient has to be made. Such a decision should not depend on the probabilities alone but also on the benefits and costs of the possible decisions, the outcome-based utilities. If we combine the two possible decisions, "treat" and "do not treat," with the two possible states, "disease present" and "disease absent," we have four different cases (Table 2). The utility of a true positive outcome, u_{tp}, is the utility of treating a patient who has the disease. The utility of a false-positive outcome, u_{fp}, refers to the consequences of treating a patient who does not have a disease. The utility of a false-negative outcome, u_{fn}, refers to the consequences of not treating a patient who has the disease. Finally, the utility of a true negative outcome, u_{tn}, refers to the consequences of not treating a patient who does not have the disease.

In actual medical decision-making, the numeric utilities should be based on things like the discomfort and harms caused by a disease, its individual and social consequences, costs of treatments, and distress associated with possible side effects (Djulbegovic et al., 2015). Thus, utilities can be positive or negative. For instance, Christensen-Szalanski and Bushyhead (1981) asked physicians to estimate the utilities of each possible decision-outcome combination for the case of pneumonia, using a rating scale ranging from −50 (worst thing

Table 2

Outcome-Based Utilities Assigned by a Sample of Physicians for Treating Pneumonia (Christensen-Szalanski & Bushyhead, 1981)

Decision	Disease present	Disease absent
Treat	$u_{tp} = 40$	$u_{fp} = -26$
Don't treat	$u_{fn} = -22$	$u_{tn} = 41$

Note. The numbers represent the assigned mean value for each combination of pneumonia diagnosis and true state of nature, on a scale from −50 ("worst thing I could do") to +50 ("best thing I could do"). These utilities give a therapeutic threshold $t_x = (41 - (-26))/[(41 - (-26)) + (40 - (-22))] \approx .52$.

I could do) to +50 (best thing I could do). Table 2 gives the mean estimates.

These estimates illustrate specific relations among utility functions in medical contexts. Typically, if a patient does have the disease, treating the patient has a higher utility than not treating the patient ($u_{tp} > u_{fn}$, e.g., 40 > −22 in Table 2). Similarly, if a patient does not have the disease, not treating the patient typically has higher utility than treating the patient ($u_{tn} > u_{fp}$, e.g., 41 > −26 in Table 2).

Iff (if and only if) both these conditions ($u_{tp} > u_{fn}$ and $u_{tn} > u_{fp}$) hold, we will call a utility function *proper*. If these conditions do not both hold, we will call a utility function *improper*. We will denote the utility function with a payoff matrix of the form $u = \begin{bmatrix} u_{tp} & u_{fp} \\ u_{fn} & u_{tn} \end{bmatrix}$, where the positioning of the utilities corresponds to the combination between the decision taken (treat or not treat) and the true state (disease or no disease) in Table 2. These ideas are graphically illustrated in Figure 1.

Utility Functions and the Therapeutic Threshold

The way in which utilities should affect behavior from a normative perspective has been studied in Bayesian decision theory (Savage, 1954), medical decision-making (Djulbegovic et al., 2015; Pauker & Kassirer, 1975, 1980), and Signal Detection Theory (Green & Swets, 1966; Stanislaw & Todorov, 1999; Swets, 1992). The best decision is typically defined as the decision that has the highest expected utility, given the applicable utility function and the probability model (including the prior probabilities and any evidence that has been obtained). The expected utility of treating a patient is given by $P(h)u_{tp} + P(\neg h)u_{fp}$, while the expected utility of not treating a patient is given by $P(\neg h)u_{tn} + P(h)u_{fn}$. Note that they are both functions of $P(h)$, which is the probability of the disease given everything known to date, including any prior test results. Figure 1 illustrates how various sets of underlying utility values lead to lines with the expected utility of treating or not treating, as a function of $P(h)$.

The therapeutic threshold (Pauker & Kassirer, 1975), which we denote with t_x, will be an important concept in subsequent analyses. It is the probability of disease for which treating the patient has the same expected utility as not treating the patient. If no further information could be obtained before deciding whether or not to treat the patient, the patient should be treated just in case the probability of disease is greater than t_x. If the utility function is proper, then the therapeutic threshold t_x can be identified by solving for $P(h)$ in Equation 1:

$$P(h)u_{tp} + P(\neg h)u_{fp} = P(\neg h)u_{tn} + P(h)u_{fn}, \quad (1)$$

and setting the threshold t_x to this value, namely to

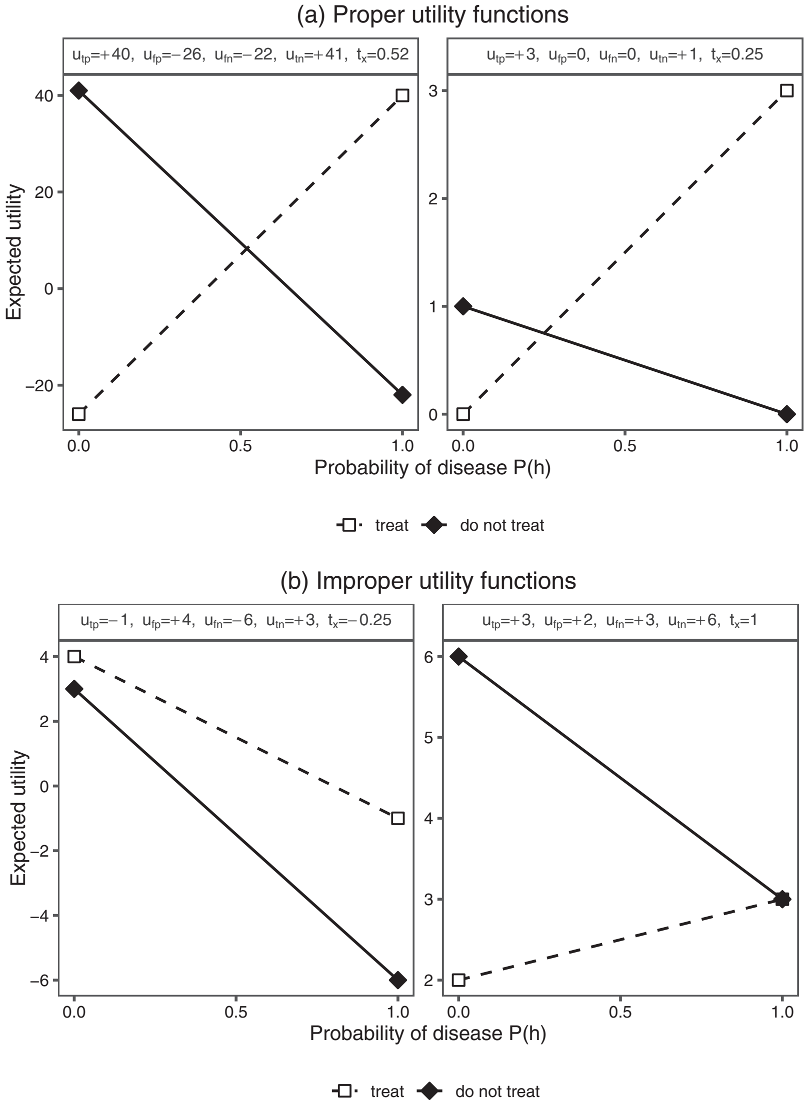
$$t_x = \frac{u_{tn} - u_{fp}}{(u_{tn} - u_{fp}) + (u_{tp} - u_{fn})}. \quad (2)$$

But what if it is not known whether a given set of utility values constitutes a proper utility function? To check whether a utility function is proper, the utility values can be plugged into the right side of Equation 2. If $(u_{tn} - u_{fp})/[(u_{tn} - u_{fp}) + (u_{tp} - u_{fn})] \geq 1$, as in the bottom-left panel of Figure 1, the utility function is improper and it is always (up to a possible tie) better to not treat, no matter what the probability of disease is. If $(u_{tn} - u_{fp})/[(u_{tn} - u_{fp}) + (u_{tp} - u_{fn})] \leq 0$, as in the bottom-right panel of Figure 1, then the utility function is improper and it is always better to treat. Although improper utility functions may seem trivial mathematically, they could be relevant in real diagnostic contexts. Importantly, if the utility function is improper, not only no single test, but no possible combination of tests, could have positive expected utility gain.

Expected Utility Gain of Tests

If the probabilities and utilities are given, then it is possible to calculate the *expected utility gain* of a Test E, denoted $eu(E)$. This is the expected

Figure 1
Examples of Utility Functions



Note. The top row illustrates two proper utility functions. The left panel is based on the pneumonia-related utilities reported in Christensen-Szalanski and Bushyhead (1981), which entail a therapeutic threshold of $t_x = .52$. The right panel is based on arbitrary values entailing a therapeutic threshold of $t_x = .25$. These are proper utility functions because the slope of the expected utility of the "treat patient" line is greater than the slope of the "do not treat" line, and those lines intersect where $0 < P(h) < 1$ on the x axis. The bottom row illustrates two arbitrary improper utility functions, for which the best decision (up to a possible tie) does not depend on the probability of the disease. In the bottom-left panel, treating the patient has higher utility than not treating, irrespective of the probability that the patient has the disease. The "treat" expected utility line has a higher slope than the "do not treat" line, illustrating that this is a necessary, but not sufficient, condition for having a proper utility function. The bottom-right panel illustrates another improper utility function, according to which it is always better to not treat the patient, up to a tie if $P(h) = 1$.

$$u_{\begin{bmatrix} u_{tp} & u_{fp} \\ u_{fn} & u_{tn} \end{bmatrix}}(e) = \max[P(h|e)u_{tp} + P(\neg h|e)u_{fp}, P(\neg h|e)u_{tn} + P(h|e)u_{fn}] - \max[P(h)u_{tp} + P(\neg h)u_{fp}, P(\neg h)u_{tn} + P(h)u_{fn}]. \quad (3)$$

utility associated with making the utility-maximizing decision after the test result is known, minus the prior utility that could be achieved on the basis of the prior probabilities and applicable utilities alone (Meder & Nelson, 2012; Schlaifer & Raiffa, 1961). Calculating $eu(E)$ requires consideration of the utilities associated with the two possible test results. The *utility of an individual test result*, $u(e)$ or $u(\neg e)$, is defined as the difference between the utility given the test result and the utility without the test result, assuming that the utility-maximizing decision (to treat or not to treat the patient) is taken in each case:

(See above)

The utility of a negative test result $\neg e$ can be defined in the same way by replacing e with $\neg e$ in Equation 3.

Accordingly, we can define the *expected utility gain of Test E* as the sum of the utilities of the possible test results e and $\neg e$, weighted by the probability of each respective test result:

$$eu_{\begin{bmatrix} u_{tp} & u_{fp} \\ u_{fn} & u_{tn} \end{bmatrix}}(E) = P(e)u_{\begin{bmatrix} u_{tp} & u_{fp} \\ u_{fn} & u_{tn} \end{bmatrix}}(e) + P(\neg e)u_{\begin{bmatrix} u_{tp} & u_{fp} \\ u_{fn} & u_{tn} \end{bmatrix}}(\neg e). \quad (4)$$

Note that not every diagnostic test increases expected utility. Intuitively, a test that does not change the probability of disease irrespective of its result (i.e., $P(h|e) = P(h|\neg e) = P(h)$), has no informational value and zero utility gain. However, even if a test does provide information about the probability of disease (i.e., $P(h|e) \neq P(h)$ and $P(h|\neg e) \neq P(h)$), it does not necessarily have positive expected utility gain. If Test E has positive expected utility gain, that is, if $eu(E) > 0$, we will call it a *useful test*.

Utilities Influence Tests' Relative Value Only via the Therapeutic Threshold

We next give three results that provide the foundation for our subsequent derivations.

Result 1. Let $H = \{h, \neg h\}$ be a binary hypothesis space associated with a disease and let $E = \{e, \neg e\}$ be a diagnostic test for H. Let $u = \begin{bmatrix} u_{tp} & u_{fp} \\ u_{fn} & u_{tn} \end{bmatrix}$ be a utility function for H. Then the expected utility of E with respect to the utility function u, $eu(E)$, is equal to the expected utility of E with respect to the utility function $u^* = \begin{bmatrix} u_{tp} - u_{fn} & 0 \\ 0 & u_{tn} - u_{fp} \end{bmatrix}$, which we denote by $eu^*(E)$; that is, $eu(E) = eu^*(E)$.

The proof is given in the Appendix. The result above essentially says that the expected utility of a Test E depends only on the differences between the utilities for correct and incorrect classifications and is independent of the particular points of reference relative to which these payoffs and costs are calculated.

Suppose we want to measure payoffs and costs in expected years of lifetime given that someone has received a specific test result. Then we could choose the amount of time someone is expected to live if they do not receive treatment for the disease as a reference point. Equally, we could choose the amount of time they are expected to live given that they do receive treatment for the disease, which they might or might not have, as our reference point. These modifications could change the utility of a positive or negative test result, $u(e)$ and $u(\neg e)$, but would not change the expected utility of a test, $eu(E)$.

Thus, to make things easier, we can always translate any set of utilities such that $u_{fp} = u_{fn} = 0$, as follows: For any utility function $u = \begin{bmatrix} u_{tp} & u_{fp} \\ u_{fn} & u_{tn} \end{bmatrix}$, we can define a new utility function $u^* = \begin{bmatrix} u^*_{tp} & 0 \\ 0 & u^*_{tn} \end{bmatrix}$ by setting $u^*_{tp} = u_{tp} - u_{fn}$ and $u^*_{tn} = u_{tn} - u_{fp}$. While this can change the utility of an individual test result, the expected utility of the test as a whole stays the same. With respect to levels of measurement as traditionally defined, what this means is that utilities need only be measured on an interval scale; utilities do not need to be measured on a ratio scale for our subsequent results to hold. Furthermore, subsequent results will make clear that so long as utilities were measured at least on an interval scale, the expected utility gain of tests—up to a constant positive multiple—exists on a ratio scale, with a true zero point.

We next analyze the expected utility gain of tests. Consider a diagnostic Test E, that is, a test for which a positive result increases the probability of disease and a negative result decreases the probability of disease. What are possible values for $eu(E)$?

First, we know from Good (1967) that no test has negative *expected* utility, assuming (as we do here) that tests have no intrinsic cost. But under which circumstances is the expected utility of a diagnostic test strictly positive?

Result 2. Let $H = \{h, \neg h\}$ be the hypothesis space associated with a disease and let u be a proper utility function with corresponding therapeutic threshold t_x. Let E be a diagnostic test for H. Then the test has positive expected utility, that is, $eu(E) > 0$, if and only if both $P(h|\neg e) < t_x$ and $P(h|e) > t_x$.

The proof is given in the Appendix. This result may be intuitive, yet it has important implications. It follows from Result 2 that the expected utility of a Test E can be zero in exactly one of two possible ways. One way is if the probability of disease is less than the treatment threshold irrespective of the test result, that is, $P(h|\neg e) \leq t_x$ and $P(h|e) \leq t_x$. The other way is if the probability of disease is greater than the treatment threshold irrespective of the test result, that is, $P(h|\neg e) \geq t_x$ and $P(h|e) \geq t_x$. This means that, while a test is *informative* if either test result changes the probability of a person having the disease, that is, if $P(h|\neg e) < P(h) < P(h|e)$, the test is *useful* only if its outcome is able to change the utility-maximizing course of action. More precisely, a test is useful if a positive test outcome results in the utility-maximizing decision changing from not treating to treating the patient or if a negative test result results in the utility-maximizing decision changing from treating to not treating the patient. Further note that for $P(h) = t_x$, any informative test also has positive expected utility, since a change in beliefs necessarily results in $P(h|\neg e) < t_x$ and $P(h|e) > t_x$. Finally, the expected utility gain of a Test E is always either zero or of the form

$$eu(E) = P(e)P(h|e)u_{tp} + P(\neg e)P(\neg h|\neg e)u_{tn} - \max[P(h)u_{tp}, P(\neg h)u_{tn}], \quad (5)$$

given that we can set $u_{fp} = u_{fn} = 0$.

The relevant factors for computing the utility of any individual state-action combination (e.g., "the utility of treating the patient is 0.6") include the prior probability of disease, the likelihoods of obtaining a positive test given that the disease is present or absent, and the four specific utilities for correct and incorrect decisions with the resulting therapeutic threshold (Table 2 and Figure 1). Must each of the four individual utilities also be considered for determining tests' *relative* expected utility, that is, the ratio of one test's utility to another test's utility, $\frac{eu(E)}{eu(F)}$? In fact, for assessing the relative utility of tests, we do not need to know all four items from the utility function: All that matters is the therapeutic threshold t_x that results from those four utility values. To show that this is the case, we need both Result 1, above, and Result 3, which is presented below.

Consider a situation in which the probabilities are known, but in which there are two possible sets of utilities, u and u^*, where u^* is formed by multiplying all the utilities in u by a positive constant. These two utility functions entail the same therapeutic threshold t_x, and therefore the utility-maximizing decision under the two utility functions is identical. However, is it possible that one test has higher expected utility under u and the other test has higher expected utility under u^*? No: If the individual utility values are multiplied by a constant, the resulting expected utility gain of the test is multiplied by that same constant:

Result 3. Let $H = \{h, \neg h\}$ be a binary hypothesis space associated with a disease and let $E = \{e, \neg e\}$ be a diagnostic test for H. Let $u = \begin{bmatrix} u_{tp} & u_{fp} \\ u_{fn} & u_{tn} \end{bmatrix}$ and $u' = \begin{bmatrix} \alpha u_{tp} & \alpha u_{fp} \\ \alpha u_{fn} & \alpha u_{tn} \end{bmatrix}$, where $\alpha > 0$, be two utility functions for H. Then the expected utility of Test E under u' is α times the expected utility of E under u; that is, $eu'(E) = \alpha\, eu(E)$.

The proof is given in the Appendix. Although this result may be intuitive, when combined with Result 1, its implications are far-reaching.

First, we can change the unit of measurement of the utility function, and the expected utility of the test will change accordingly. For example, if the expected utility of a Test E is 0.10 expected life years, then the expected utility of Test E is also 5.2 expected life weeks, assuming 52 weeks in a year.

Second, we can combine Result 1 with Result 3. Suppose we have a proper utility function u. According to Result 1, we can replace u with a utility function u^* with the same therapeutic threshold t_x and $u^*_{fp} = u^*_{fn} = 0$, such that $eu(E) = eu^*(E)$ by setting $u^*_{tp} = u_{tp} - u_{fn}$ and $u^*_{tn} = u_{tn} - u_{fp}$. For $\alpha = u^*_{tn} + u^*_{tp} = u_{tn} - u_{fp} + u_{tp} - u_{fn}$, we can rewrite u^* as $\alpha \cdot \begin{bmatrix} (1-t_x) & 0 \\ 0 & t_x \end{bmatrix}$. If we denote the expected utility of E relative to $\begin{bmatrix} (1-t_x) & 0 \\ 0 & t_x \end{bmatrix}$ by $eu_{t_x}(E)$, then, according to result Result 3, $eu(E) = eu^*(E) = \alpha \cdot eu_{t_x}(E)$.

For the present analyses, the *relative* expected utility of tests is especially important because we are concerned with identifying which test out of two (or more) has the highest expected utility gain. Suppose we have another Test F with $eu(F) > 0$. Then the ratio $\frac{eu(E)}{eu(F)}$ for the two tests E and F relative to u is equal to $\frac{\alpha \cdot eu_{t_x}(E)}{\alpha \cdot eu_{t_x}(F)} = \frac{eu_{t_x}(E)}{eu_{t_x}(F)}$ and is thus independent of α. Now suppose we have another payoff matrix $u' = \begin{bmatrix} u'_{tp} & u'_{fp} \\ u'_{fn} & u'_{tn} \end{bmatrix}$ with the same therapeutic threshold t_x as u. Then we can replace u' with a payoff matrix $\alpha' \cdot \begin{bmatrix} (1-t_x) & 0 \\ 0 & t_x \end{bmatrix}$, with $\alpha' = u'_{tn} - u'_{fp} + u'_{tp} - u'_{fn}$, $\alpha' \neq 0$ and again, the ratio $\frac{eu(E)}{eu(F)}$ for the two tests E and F relative to u' is equal to $\frac{\alpha' \cdot eu_{t_x}(E)}{\alpha' \cdot eu_{t_x}(F)} = \frac{eu_{t_x}(E)}{eu_{t_x}(F)}$. Thus, it does not matter whether the relative utility of tests E and F is calculated using u or u'. As long as the two utility functions entail the same threshold t_x, the ratio $\frac{eu(E)}{eu(F)}$ will be the same.

The key insight from these analyses is that if a particular test-selection strategy is optimal in a situation with normalized utilities of the form $\begin{bmatrix} (1-t_x) & 0 \\ 0 & t_x \end{bmatrix}$, which entails a therapeutic threshold of t_x, then such a test-selection strategy is also optimal in any other situation whose utilities are associated with the same therapeutic threshold t_x. As a consequence, it is possible to use such normalized utility functions without loss of generality in both analytical and simulation-based investigations. An important psychological implication is that if a physician has an idea of the appropriate therapeutic threshold, for example, that it would make sense to treat a particular disease if the probability of that disease is at least 0.1, then the precise numerical ingredients of the utility function do not need to be known in order to assess the relative expected utility of the possible tests.[1]

Analytic Results on the Likelihood Difference Heuristic

The previous sections have provided a formal analysis of how to determine the expected utility of a test, and the relation between tests' utility and the therapeutic threshold implied by the utilities. In this section, we are concerned with test *selection*: Given two (or more) available tests for diagnosing a patient, how should one determine which test to conduct? Our focus here is on a very simple strategy that does not rely on calculating each test's expected utility gain.

The likelihood difference of Test E is the difference between the likelihood of obtaining a positive test result given the disease is present (the true positive rate) and the likelihood of obtaining a positive test result given the disease is absent (the false positive rate), formally $\lambda(E) = P(e|h) - P(e|\neg h)$. We call the strategy of conducting the test with the highest likelihood difference the Likelihood Difference Heuristic (LDH). Suppose there are two tests, E, such that $P(e|h) = 1$ and $P(e|\neg h) = 0.2$, and F, such that $P(f|h) = 0.9$ and $P(f|\neg h) = 0.05$. Test E has $\lambda(E) = 0.80$, whereas Test F has $\lambda(F) = 0.85$, so the LDH would select Test F to conduct. (If two or more tests tie for highest likelihood difference, the LDH selects among them randomly.)

[1] Consider some arbitrary statistical environment, for instance where $P(h) = 0.3$, and $P(e|h)$ and $P(e|\neg h)$ are generated from a uniform distribution between 0 and 1. Suppose the applicable utility function is $u = \begin{bmatrix} u_{tp} & u_{fp} \\ u_{fn} & u_{tn} \end{bmatrix} = \begin{bmatrix} 0 & -9 \\ -1 & 0 \end{bmatrix}$ which corresponds to a therapeutic threshold of $t_x = 0.1$. Further assume that a particular test-selection strategy identifies the objectively more useful test out of a pair of random tests in 70% of cases. Given these results we know that the performance of the heuristic would be the same in an environment where $u = \begin{bmatrix} 0.9 & 0 \\ 0 & 0.1 \end{bmatrix}$, or any other proper set of utilities for which $t_x = 0.1$, and no matter if the utilities measure extended life expectancy, dollars saved, or any other units.

The LDH, like information gain and other OED models, ignores outcome utilities and does not in general choose the highest expected utility gain test (Meder & Nelson, 2012). However, whereas most pure-information utilities take the prior probability of disease into account, the LDH also ignores the prior probability. In situations with outcome-based utilities, would it ever be sensible to use such a simplistic strategy to select tests? In the following, we prove that given maximal need for information, that is, if the prior probability of the disease $P(h)$ equals the therapeutic threshold t_x, the LDH is guaranteed to identify the higher-expected-utility-gain test.

Result 4. Let $H = \{h, \neg h\}$ be a hypothesis space associated with a disease and let u be a proper utility function for H with corresponding therapeutic threshold t_x. If $P(h) = t_x$, then there exists a positive β, such that for any diagnostic Test E, we have $eu(E) = \beta\lambda(E)$.

The proof is given in the Appendix. This result shows not only that the LDH will select the highest expected utility gain test but that a test's expected utility gain is a constant multiple of its likelihood difference.

What should that constant multiple, β, be? If standardized utilities of the form $\begin{bmatrix} (1-t_x) & 0 \\ 0 & t_x \end{bmatrix}$ are used, then, according to the proof of Result 4, $\beta = P(h)\ (1 - P(h)) = t_x\ (1 - t_x)$. Thus, for any diagnostic Test E and given that the decision maker is subject to maximal need of information and therefore decision indifference, the expected utility gain is highest if $P(h) = t_x = 0.5$. However, from Results 1 and 3, it follows that the ratio of two tests' expected utility depends only on the therapeutic threshold t_x entailed by the applicable utility function and is otherwise independent of the particular values in $\begin{bmatrix} u_{tp} & u_{fp} \\ u_{fn} & u_{tn} \end{bmatrix}$. Accordingly, in a situation of maximal need for information, the LDH invariably chooses a (not necessarily unique) utility-maximizing test.

What if a decision maker is not at the point of decision indifference, for example, if $P(h) \neq t_x$? In this case, the LDH is not in general optimal.

Result 5. Let $H = \{h, \neg h\}$ be a hypothesis space associated with a disease and let u be a proper set of utility values for H with corresponding therapeutic threshold t_x. Then if $P(h) \neq t_x$, there exist tests E and F such that $\lambda(E) > \lambda\ (F)$ while $eu(E) < eu(F)$.

The proof in the Appendix gives a procedure that is applicable in any situation where $P(h) \neq t_x$, for finding a pair of binary tests E and F such that one test has higher expected utility, but the other test has a higher likelihood difference.

Importantly, while this result demonstrates that the LDH is not exactly optimal if $P(h) \neq t_x$, we do not know whether the LDH's performance degrades gracefully as we move away from its optimality condition. Nor do we know whether other pure-information strategies (such as information gain or probability gain) would also be optimal if $P(h) = t_x$. We also do not knowm whether other purely informational OED strategies would perform better or worse than the LDH under particular combinations of $P(h)$ and t_x values. We address these issues in a simulation experiment in a subsequent section on "The Bigger Picture: Simulation Results and Test Selection Based on OED Models".

What Likelihoods Matter Where: Sensitivity and Specificity

Result 4 implies that if an agent is at a point of decision indifference because $P(h) = t_x$, and two tests have the same likelihood difference, then those tests also have the same expected utility. Are there some test characteristics that reveal which of two tests with the same likelihood difference has the highest expected utility if $P(h) \neq t_x$?

It turns out that there are. Specifically, the relevant test characteristics are the likelihood that the test is positive, given that the patient has the disease, $P(e|h)$, and the likelihood that the test is negative, given that the disease is absent $P(\neg e|\neg h)$. In the medical literature, $P(e|h)$ is referred to as the *sensitivity* of a test, and $P(\neg e|\neg h)$ is referred to as the *specificity* of a test. Which of these two likelihoods is more important for a test's expected utility gain depends on the relation between the prior probability of disease $P(h)$ and the therapeutic threshold t_x: If the prior probability of disease is *below* the threshold (i.e., $P(h) < t_x$), the relevant likelihood is $P(\neg e|\neg h)$, the test's specificity. Conversely, if the prior probability of disease is *above* the threshold (i.e., $P(h) > t_x$), the relevant likelihood is $P(e|h)$, the test's sensitivity.

Before we prove this, we will show that there is an interesting relationship that establishes which test likelihood is most important for the posterior probability of disease given a positive test result, $P(h|e)$, and the posterior probability of the disease being absent given a negative test result, $P(\neg h|\neg e)$. (In medicine, $P(h|e)$ is referred to the as *positive predictive value* of a test, and $P(\neg h|\neg e)$ is referred to as the *negative predictive value* of a test.)

Result 6. Let $H = (h, \neg h)$ be a hypothesis space associated with a disease and let E and F be two diagnostic tests for H with equal likelihood difference, $\lambda(E) = \lambda(F)$. Then the following holds:

i. If $P(\neg e|\neg h) > P(\neg f|\neg h)$, then $P(h|e) > P(h|f)$. Thus, if E has higher specificity than F, then the posterior probability of disease is higher given a positive result from E than given a positive result from F.
ii. If $P(e|h) > P(f|h)$, then $P(\neg h|\neg e) > P(\neg h|\neg f)$. Thus, if E has higher sensitivity than F, then the posterior probability of disease is lower given a negative result from E than given a negative result from F.

The proof is given in the Appendix. We now give a more general result on the relationship among the sensitivity $P(e|h)$, the specificity $P(\neg e|\neg h)$, the prior probability $P(h)$, and the therapeutic threshold t_x.

Result 7. Let $H = (h, \neg h)$ be the hypothesis space associated with a disease and let u be a proper utility function for H. Let $E = (e, \neg e)$ and $F = (f, \neg f)$ be two diagnostic tests for H, with $\lambda(E) = \lambda(F)$. Then the following holds:

i. If $P(h) < t_x$ and $P(\neg e|\neg h) \geq P(\neg f|\neg h)$, we have $eu(E) \geq eu(F)$.
ii. If $P(h) > t_x$ and $P(e|h) \geq P(f|h)$, we have $eu(E) \geq eu(F)$.

The proof is given in the Appendix. One interesting insight from the proof is that we can calculate the expected utility of a Test E as a function of the specificity of E if $P(h) < t_x$ or as a function of the sensitivity of E if $P(h) > t_x$. More precisely, we can show that:

i. If $P(h) < t_x$, then the expected utility of a test is either zero or of the form

$$eu(E) = u_{tp}P(h)\lambda(E) + [spec(E) - 1][u_{tn}P(\neg h) - u_{tp}P(h)]. \quad (6)$$

ii. If $P(h) > t_x$, then the expected utility of a test is either zero or of the form

$$eu(E) = u_{tn}P(\neg h)\lambda(E) + [sens(E) - 1][u_{tp}P(h) - u_{tn}P(\neg h)]. \quad (7)$$

This means that for $p(h) < t_x$, $eu(E)$ will be smaller than $u_{tp}P(h)\lambda(E)$, as $[u_{tn}P(\neg h) - u_{tp}P(h)]$ is positive in that case. Similarly, for $p(h) > t_x$, we can see that $eu(E)$ will be smaller than $u_{tp}P(\neg h)\lambda(E)$, as $[u_{tp}P(h) - u_{tn}P(\neg h)]$ is positive in that case. Recall that for $P(h) = t_x$, we have $eu(E) = u_{tn}P(\neg h)\lambda(E) = u_{tp}P(h)\lambda(E)$.

Is this result also true if the tests have different likelihood differences? It is not, as the following counterexample shows. Suppose the prior probability of disease $P(h) = 0.50$ is larger than the therapeutic threshold $t_x = 0.15$. Suppose further that we have two tests E and F, with $P(e|h) = 0.90$, $P(e|\neg h) = 0.45$, $P(f|h) = 0.84$, and $P(f|\neg h) = 0$. Because positive test results increase the probability of disease for both tests, both tests are informative. Test E has higher sensitivity (0.90) than Test F (0.84). But whereas Test E has zero expected utility gain, Test F has positive expected utility gain.

What is happening in this scenario? Because the prior probability of disease is greater than the therapeutic threshold, if no test could be conducted, the best decision would be to treat the patient. In this type of situation, a test has positive utility if the posterior probability of disease given a negative test outcome would be less than the therapeutic threshold. In this case, the posterior probability of disease given a negative outcome to Test E is about 0.154, so the best decision would still be to not treat the patient. In the case of Test F, however, the probability of disease given a negative test result is about 0.138, which is below the therapeutic threshold. Since a negative test result changes the best course of action to take (i.e., the utility-maximizing treatment decision), Test F has positive expected utility gain.

The Bigger Picture: Simulation Results and Test Selection Based on OED Models

If the therapeutic threshold t_x equals the prior probability of disease $P(h)$, then the LDH invariably identifies the test with highest expected utility gain. Do better-known OED models, which select tests based on epistemic utility functions, also do so? What if the precise values of $P(h)$ and t_x are not known? If the prior probability of disease is only approximately equal to the treatment criterion, $P(h) \approx t_x$, is it still a good bet to use the LDH? How does the LDH compare with prominent OED models if the prior probability of disease and the therapeutic threshold are not even approximately equal, $P(h) \neq t_x$? Here, we address these questions via simulations and a numeric example.

Other OED Models Are Suboptimal Where the LDH Is Optimal

The LDH invariably identifies the best test if $P(h) = t_x$. But if $P(h) = t_x$, and $0 < t_x < 1$, then any informative test is useful. Do other models of the value of information also identify the most useful test in this case? Here, we consider some OED models that are prominent in the medical decision-making and psychological literature. (For equations and example calculations with these OED models, see Nelson, 2005.) The models we consider are as follows:

- expected information gain, based on expected reduction in Shannon's (1948) entropy (Benish, 1999, 2003; Lindley, 1956; Oaksford & Chater, 1994);
- probability gain, which quantifies improvement in classification accuracy (Baron, 1985; Nelson, 2005); and
- Bayesian diagnosticity and log diagnosticity, which are based on likelihood ratios or log likelihood ratios (Good, 1950; Good & Card, 1971).

Do these OED test-selection methods, like the LDH, also correctly identify the best test when $P(h) = t_x$? Consider a case where $P(h) = t_x = 0.25$, $P(e|h) = 0.81$, $P(e|\neg h) = 0.27$, $P(f|h) = 0.43$, and $P(f|\neg h) = 0$. The likelihood difference of Test E, $\lambda(E) = 0.81 - 0.27 = 0.54$, which is greater than the likelihood difference of Test F, $\lambda(E) = 0.43 - 0.00 = 0.43$. Because $P(h) = t_x$, we know from Result 4 that the ratio in the expected utility gain of Test E to Test F is $\lambda(E)/\lambda(F) = 0.54/0.43 \approx 1.26$, and Test E is more useful. Do the other OED models also correctly identify Test E as more useful? In contrast to the LDH, they do not. Test F has higher information gain (0.246 bit vs. 0.167 bit for Test E), probability gain (0.108 vs. null), Bayesian diagnosticity (infinite vs. 3.501), and $\log_{10}$ diagnosticity (infinite vs. 0.541). Thus, it is not the case that just any pure-information test-selection strategy, even if $P(h) = t_x$, can be counted on to identify the highest utility gain test to conduct.

Information gain and Kullback–Leibler (KL) divergence, which have been suggested in the medical literature (Benish, 1999, 2003), pick the lower-expected-utility-gain test. KL divergence and information gain are identical for purposes of evaluating tests' expected usefulness (Oaksford & Chater, 1994). Probability gain, which accounts well for human test selection in probabilistic classification tasks when probabilities are learned experientially (Nelson et al., 2010), also picks the lower-expected-utility-gain test in this example. It is worth noting that information gain and probability gain are both special cases of the Sharma–Mittal family of generalized information gain measures (Crupi et al., 2018; Sharma & Mittal, 1975), and other information-theoretic models could also be used. The likelihood ratio-based methods, Bayesian diagnosticity and log diagnosticity, are not part of the Sharma–Mittal family of information-theoretic models. These models actually deem the lower-expected-utility-gain test to be *infinitely* useful. The fact that the likelihood ratio-based methods fail here may be surprising to readers familiar with Signal Detection Theory (Green & Swets, 1966; Stanislaw & Todorov, 1999), where it is a convention to state decision thresholds in terms of likelihood ratios. These likelihood ratios take prior probabilities and payoff utilities into account, in effect quantifying how much evidence would be needed to change the best *decision*. However, this example shows that it does not follow that likelihood ratio-based *test-selection* methods are optimal. Nor is this example unique, or dependent on having a likelihood value that is either 0 or 1. For instance, suppose we were to change $P(f|\neg h)$ from 0 to 0.001, while keeping the therapeutic threshold, prior probabilities and other likelihoods unchanged. Probability gain and information gain would not change appreciably in this case. The likelihood ratio-based methods, though no longer deeming Test E to have infinite expected utility, would still deem Test E to be more useful than Test F. Further discussion of limitations of the likelihood ratio-based OED models, for instance, how in some circumstances they are insensitive to prior probabilities, can be found in Nelson (2005). Minka (2001) gives a broader discussion of limitations of likelihood-based methods in statistics.

Where Does Choice of Test Matter?

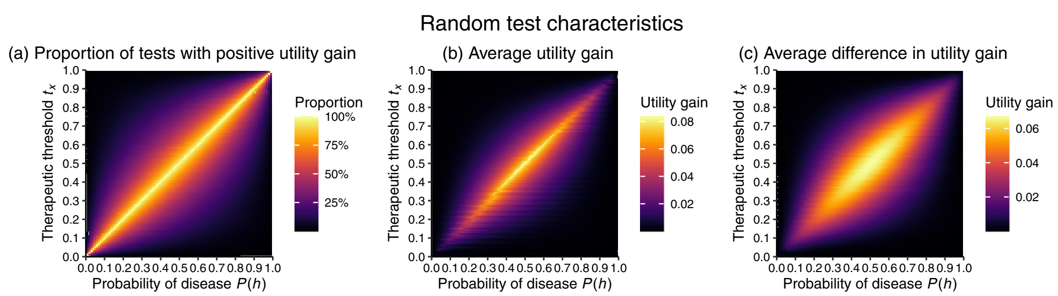
Our analytic results on the LDH, and the example above, show that the LDH, but not prominent OED models from the literature, is optimal if the prior probability of disease exactly equals the therapeutic threshold. But even if one has a good estimate of those quantities, they seldom are exactly known. What happens if the prior probability of disease is close to, but not exactly, the therapeutic threshold? What if the prior probability $P(h)$ and the therapeutic threshold t_x are very different from each other?

Before simulating the performance of various strategies, we first explore the implications of various $P(h)$ and t_x values. This is important because it may be the case that for some combinations of $P(h)$ and t_x, few tests actually have positive expected utility gain, and that for other combinations of $P(h)$ and t_x, randomly drawn tests may differ substantially in their expected utility gain. Figure 2 shows the results of these simulations. Figure 2a gives the empirical frequency that a randomly generated test has positive (greater than 10^{-10}) expected utility gain, as a function of $P(h)$ and t_x, each incremented from 0.01, 0.02, … 0.99. Every test was generated by independently sampling two numbers from a uniform distribution on the interval [0,1], assigning the larger probability to $P(e|h)$ (the true positive rate) and the smaller probability to $P(e|\neg h)$ (the false positive rate). Figure 2a and b are based on 10,000 random tests for each combination of $P(h)$ and t_x; Figure 2c is based on 10,000 pairs of random tests for each combination of $P(h)$ and t_x. To avoid floating-point errors, only tests with expected utility gain greater than 10^{-10} normalized utility units of the form $\begin{bmatrix} (1-t_x) & 0 \\ 0 & t_x \end{bmatrix}$ were deemed to have positive expected utility gain.

Results show that the combination of $P(h)$ and t_x strongly influences the tendency of a random test to have positive expected utility gain (Figure 2a). For instance, if $P(h)$ is very low and t_x is very high, it is extremely rare for a random test to have positive expected utility gain. This makes sense: If the prior probability of disease is low, and a high threshold t_x applies, then only a test with very high positive predictive value $P(h|e)$ can potentially change the best decision from don't treat to treat. Conversely, if $P(h)$ is very high and t_x is very low, then only a test with very high negative predictive value $P(\neg h|\neg e)$ can have positive utility gain. Figure 2b shows that this general pattern is preserved if we consider the mean expected utility gain of randomly generated tests. Figure 2c shows that a similar pattern applies if we consider the mean expected utility gain difference of randomly generated *pairs* of tests.

The upshot of all this is that, from the perspective of making the right test-selection decision or of capturing as much expected utility gain as possible, not all combinations of $P(h)$ and t_x

Figure 2
Characteristics of Randomly Generated Tests



Note. (a) Proportion of randomly generated tests that have positive expected utility gain as a function of the probability of disease $P(h)$ (x axis) and therapeutic threshold t_x (y axis). (b) Average utility gained from randomly generated tests. (c) Average difference in expected utility gain in pairs of randomly generated tests. See the online article for the color version of this figure.

values are equally important. Rather, situations in which $P(h)$ and t_x have similar values, especially if these values are not too extreme, are by far the most important. In the next section, we address how the LDH and other strategies perform as a function of $P(h)$ and t_x.

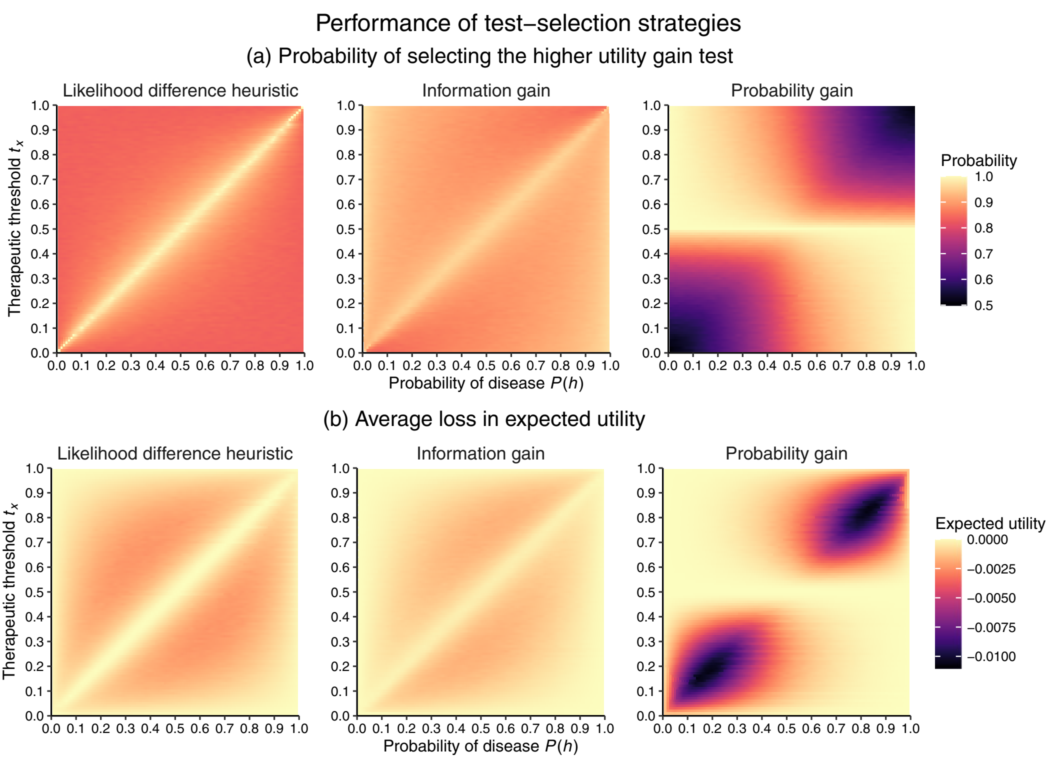
Simulations of the LDH, Information Gain, and Probability Gain

We now know that prominent OED models in medical decision-making and psychological literature, unlike the LDH, are not in general optimal when $P(h) = t_x$. We further know that for purposes of making a right choice where it matters, situations where $P(h) \approx t_x$, especially if the values are not too extreme, are especially important. How do the LDH and other pure-information-based OED models perform for various $P(h)$ and t_x values?

To address this question, we simulated the performance of the LDH, information gain, and probability gain, as shown in Figure 3. Similar to Figure 2, we considered every combination of $P(h)$ and t_x, each ranging from 0.01, 0.02, … 0.99. The values plotted are the means from simulations of 10,000 random pairs of tests at each point. In the top row of Figure 3, we plot the probability that using a particular strategy leads to choosing the higher-expected-utility-gain test out of a pair of tests. To be conservative, in the case where the difference in the normalized utility values is less than 10^{-10}, we counted either test choice as a success. From left to right, results are plotted for the LDH, information gain, and probability gain. The LDH (top left) is optimal if $P(h) = t_x$, consistent with our analytical results, with its performance decreasing quickly but smoothly as we move away from $P(h) = t_x$. The LDH's performance is quite a bit better than chance

Figure 3

Simulation of Various Pure-Information Test-Selection Strategies' Performance



Note. (a) Probability that the best test is selected from a pair of tests if the LDH, information gain, or probability gain model are used. (b) The expected utility gain loss that would be incurred by using the LDH, information gain, or probability gain model, respectively, to select among pairs of tests. LDH = likelihood difference heuristic. See the online article for the color version of this figure.

throughout the range of $P(h)$ and t_x values considered. Information gain (top middle) is not optimal anywhere among the range of $P(h)$ and t_x values considered. However, information gain appears fairly robust across all combinations of $P(h)$ and t_x values and performs better than the LDH for many combinations of $P(h)$ and t_x. Probability gain is optimal if $t_x = 0.5$, which is the case where the utility-maximizing decision corresponds to the most probable state, namely to treat the patient iff $P(h) > 0.5$. However, unlike the LDH and information gain, if $P(h)$ and t_x are both either very low or very high, probability gain performs very poorly, approaching chance performance.

In the bottom row of Figure 3, we plot the mean loss in expected utility resulting from using the LDH, information gain, and probability gain, to select among pairs of tests. In some respects, these results parallel the results for probability of selecting the best test. However, some key differences should be pointed out. One key difference is that if $P(h) \approx 0$ and $t_x \approx 1$, there is next to no cost of whichever test-selection strategy is used because (as shown in Figure 2) almost no possible decision has positive expected utility gain. The other key difference is in the comparatively broad and smooth-boundaried regions bordering the boundary cases in which the LDH and probability gain are demonstrably optimal. If $P(h) \approx t_x$, even if $P(h)$ and t_x are not exactly equal, using the LDH is a solid bet. A similar statement can be made for probability gain: if $t_x \approx 0.5$, using probability gain is a solid bet.

To summarize, not all combinations of $P(h)$ and t_x are equally important from the perspective of maximizing outcome-based utilities; rather, cases where $P(h) \approx t_x$, and the values are not too extreme, are especially important. Interestingly, these conditions highly overlap with where the LDH is either exactly ($P(h) = t_x$) or approximately ($P(h) \approx t_x$) optimal. If the optimality conditions of the LDH or probability gain are approximately met, then there is effectively no loss in utility from using the appropriate pure-information strategy, rather than explicitly calculating the actual expected utility gain of each test. Information gain, though not optimal for any combination of $P(h)$ and t_x, gives robust performance throughout and is in fact better than the LDH for many combinations of $P(h)$ and t_x. Probability gain is optimal if $t_x = 0.5$, and robust in cases where $t_x \approx 0.5$. However, it is not as though any pure-information strategy would be sensible to use. In particular, if $P(h)$ and t_x are both fairly low, or if $P(h)$ and t_x are both fairly high, then using probability gain could be quite costly in terms of expected utility gain.

A Real-World Example: Latent Tuberculosis Testing

Newell and Card (1985) noted that "Nothing drives basic science better than a good applied problem." Are there real-world situations in which the LDH could be applied? Indeed, there are many situations in medicine (Alberg et al., 2004) and other domains in which it is necessary to choose among binary (or binarized) tests. Here, we consider the case of latent tuberculosis screening tests. According to the World Health Organization, active tuberculosis (TB) is one of the top 10 causes of death worldwide (World Health Organization [WHO], 2019, 2020). TB is estimated to have killed 1.5 million people in 2018 and 1.4 million people in 2019 (WHO, 2019, 2020). An important strategy to prevent spread of the disease is to identify individuals with latent tuberculosis infection (LTBI), that is, individuals who have been infected with *Mycobacterium tuberculosis* but do not present any radiographic or bacteriologic evidence or symptoms of TB. These individuals are expected to develop active TB at a later stage of their life in approximately 10%–12% of cases and might benefit from preventive treatments (Esmail et al., 2014).

LTBI screening represents an interesting case study for several reasons. To begin with, there is no "gold standard" test for LTBI. The existing tests can be grouped into two classes, the tuberculin skin test (TST) and the interferon gamma release assays (IGRA), each of which has advantages and limitations. For simplicity, we will refer to the TST and to the IGRA as if each were a single test, rather than classes of tests. In particular, TST has been widely used for a century, with well-studied clinical applications and cutoff points for therapeutic indications for different ages and risk groups. However, its measurement is subject to interobserver variability, a positive test result does not distinguish between recent and earlier infections (which have a lower risk of progression to disease), and repetitions of the test might generate a booster phenomenon. IGRA tests are not subject to reader bias and do not tend to boost responses when repeated.

However, accuracy can be decreased by problems in collecting or transporting blood specimens. Additionally, there is limited data on the use in some groups, such as young children and immunocompromised persons. Some studies (e.g., Al-Orainey, 2009; Mazurek et al., 2010; Smith et al., 2011) have suggested a slightly higher sensitivity for the TST and a higher specificity for the IGRA.

What likelihood values apply for the IGRA and TST tests? Based on our reading of the literature, reasonable estimates for each test's sensitivity and specificity, values are shown in Table 3. The point for the present article is not whether these numbers are exactly right. Indeed, every study has a different estimate, also depending on which specific IGRA or TST test is considered. Rather, our motivation here is to use plausible numbers from a real-world example to try to connect to our theoretical analyses. Recall that a test's sensitivity is $P(e|h)$, and that 1 − specificity = $1 - P(\neg e|\neg h) = P(e|\neg h)$. If we use these values to compute the likelihood difference λ for each test, we get λ(TST) = .78 and λ(IGRA) = .83 (Table 3). According to these numbers, then, the likelihood difference is greater for the IGRA test. Thus, if a doctor would use the LDH, they would choose the IGRA rather than the TST test.

Under Which Circumstances Would Each Test Be Most Useful?

Does the TST or IGRA test have higher expected utility gain? More precisely, what are the conditions in terms of prior probability of disease $p(h)$ and therapeutic threshold t_x under which the TST or IGRA test would have higher expected utility gain? Before asking what prior probability or therapeutic threshold is most plausible, we consider every possible combination of prior probabilities and therapeutic thresholds. In Figure 4a, we plot the usefulness of the TST test as a function of the prior probability of disease $P(h)$ and the therapeutic threshold t_x, with both $P(h)$ and t_x ranging from 0.01, 0.02, ... 0.99. Figure 4b shows the same analysis for the IGRA test. The color of each point denotes the expected utility gain of a test for a particular combination of $p(h)$ and t_x. As the legends to the right of Figure 4a and 4b note, black denotes zero utility; dark purple denotes low utility; red denotes greater utility; orange and yellow denote highest utility. Broadly speaking, we see that both tests tend to have higher utility where $P(h)$ and t_x are similar in value to each other, and moderate (neither close to 1 nor close to 0). There is also a broad range of circumstances, for example, if the prior probability of disease is high and the therapeutic threshold is low (or vice versa), in which both of the tests have zero utility. Given these two analyses, we can determine the circumstances under which each test has higher expected utility gain. Figure 4c plots this, with red denoting where the IGRA test has higher utility and blue denoting where the TST test has higher utility. Black denotes where the tests are both useless because no test result could change which course of action (treat or don't treat) has higher expected utility gain. This type of analysis, paralleling results from the simulations in Figure 3, shows that even approximate knowledge of the therapeutic threshold and prior probability of disease could be enough to enable selection of the higher utility test.

What Prior Probabilities and Therapeutic Threshold Values Are Plausible?

What is the true probability of having LTBI? One reason for our interest in LTBI screening is the enormous variation of LTBI prevalence across countries and, within the same country, across specific subpopulations. For example, it has been reported that around one third of the world's population is affected by LTBI (Lamberti et al., 2014), including more than 40% in Uganda or 50% in the Ivory Coast, but less than 5% in the U.S. and most European countries (Bentley et al., 2012; Mazurek et al., 2010). However, even in countries in which the general prevalence of LTBI

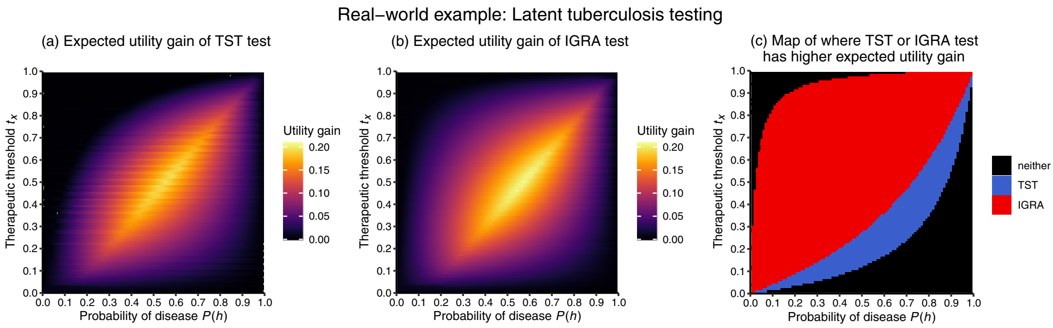
Table 3
Approximate Likelihoods of LTBI Screening Tests

Test	Sensitivity (%)	Specificity (%)	$P(e\|h)$	$P(e\|\neg h)$	LikDiff λ
TST	88	90	.88	.10	.78
IGRA	85	98	.85	.02	.83

Note. LTBI = latent tuberculosis infection; TST = tuberculin skin test; IGRA = interferon gamma release assays.

Figure 4

Real-World Example: Usefulness of IGRA and TST Tests in Latent Tuberculosis Testing



Note. (a) Expected utility gain of the TST test. (b) Expected utility gain of the IGRA test. (c) Map showing where the TST (blue) or IGRA test (red) has higher expected utility gain; with black denoting areas where neither test has expected utility gain. The tests tied exactly in 3 of the $99^2 = 9{,}801$ points plotted. All of these points were on the boundary between the regions where the TST and IGRA tests have higher utility. For purposes of plotting the figure, these three points were arbitrarily assigned to the TST test. IGRA = interferon gamma release assays; TST = tuberculin skin test. See the online article for the color version of this figure.

is low, people working in hospitals, homeless shelters, correctional facilities, or residential facilities for patients with HIV infection are considered at higher risk (Nienhaus et al., 2014; Zhang et al., 2013). For example, the prevalence of LTBI within German health care workers has been reported to be around 9.9% (Schablon et al., 2010).

What therapeutic threshold is appropriate for treating LTBI? We informally asked a small sample of experienced physicians from the Negrar Center for Tropical Diseases (Verona, Italy) to provide an estimate of the therapeutic threshold for LTBI. When asked for the probability of LTBI above which it would make sense to treat a young adult in good health, they gave an estimate of about $t_x = 10\%$. Interestingly, this therapeutic threshold is essentially identical to the estimated LTBI prevalence of 9.9% within German health care workers. Thus, if a screening LTBI test was to be performed on people from a similar reference class (e.g., health care professionals living somewhere else in Europe), IGRA should be preferred over TST as it is likely to provide a higher expected gain in utility, given that $P(h) \approx t_x$.

Limitations and Implications of the Latent Tuberculosis Infection Analysis

A number of simplifications are required for the above (or any) analysis, some of which might limit its usefulness; we therefore consider some limitations here. To start with, probabilities we have estimated from some articles in the literature, and the therapeutic threshold t_x estimate obtained from a small sample of physicians, should not be taken as authoritative. However, our formal analyses do not change according to where the payoffs come from, and they can easily be applied not only to the patient's utilities but also to public health or environmental objectives, and ideally to a combination of them. We also did not consider intrinsic test costs. TST is inexpensive, although it does require trained health care professionals and two visits (the first for administration, the second for interpretation). By contrast, IGRA tests can be done on a single visit but require laboratory facilities. Such costs could be included in the analyses and could change the results in Figure 4 as to the utility of each test or the circumstances (combinations of $P(h)$ and t_x) under which each test is more useful. Other simplifications are that people vary in their tendency for LTBI to progress to full TB or in what developing TB implies for them. A more sophisticated model would also include the risk of progressing to TB if not treated (which is higher for individuals with significant comorbidities) as well as the personal or social consequences of developing TB (which are more serious for individuals who are in contact with children or immunocompromised people). Information gain or other general-purpose OED models can be used in situations with multiple possible states of the world, for example, if there were three or more possibilities in terms of the likelihood of LTBI to develop into active TB. However, the LDH, or other procedures that are based on binary tests and states of the world, would no longer be applicable.

Arguing in favor of the applicability of the LDH, although quantities in nature are often continuously variable, states of nature and test results are often treated as binary. Moreover, the LDH (in contrast to an equation that takes all payoffs into account) is easy to use and easy to explain and thus could be a helpful tool in training practitioners about test selection. Another motivation for using the LDH could be its robustness: It makes sense to use the LDH if $P(h) \approx t_x$, even if neither $P(h)$ nor t_x are known exactly. Our LTBI analysis suggests that there may be relevant situations in which the LDH identifies the higher-utility-gain test.

General Discussion

Actively acquiring information through test selection is important for reducing uncertainty and identifying the best course of action to take. However, tests are often time-consuming and costly, both in terms of human and monetary resources. What is a sensible way to select a test if situation-specific utilities apply, and what are the conditions under which different test-selection strategies perform well? If exact probabilities and the applicable outcome-based utilities are known, then from a normative perspective one should use this information and conduct a Bayesian decision-theoretic analysis to identify the utility-maximizing test. In this case, neither the LDH nor other general-purpose OED strategies based on informational utility functions are needed. But seldom are $P(h)$, or the therapeutic threshold t_x that is entailed by the outcome-based utilities, known exactly. Our results show that if the prior probability of disease $P(h)$ is even approximately equal to the therapeutic threshold t_x, it is either optimal or close to optimal to use the LDH to select tests. Importantly, these circumstances overlap highly with the situations in which one's test-selection strategy has the biggest impact in terms of expected utility gain (Figure 2). Not just any pure-information OED model should be used, however (Figure 3). The LDH and information gain have reasonable performance across a wide range of $P(h)$ and t_x values. By contrast, for some cases—in particular where $P(h)$ and t_x are both low or high—probability gain approximates chance performance.

We have analyzed the situation of selecting a single test before making a treatment decision. However, in many situations, more than one test can be conducted. This is a critical issue because it is not in general the case that the best standalone test is the best first test to conduct in a sequence of tests (Geman & Jedynak, 1996, 2001; Hyafil & Rivest, 1976; Meder et al., 2019; Nelson et al., 2018). Planning an optimal sequential question strategy requires knowledge of how the test results relate to each other, which cannot in general be derived from the individual test likelihoods. For instance, in the LTBI example, it is unclear what the probability of disease given any particular set of results for the two tests is. People (Jarecki et al., 2017), popular machine learning methods (Domingos & Pazzani, 1997), and medical decision-making models (Hamm et al., 2014) tend to presume a priori that tests are class-conditionally independent (i.e., test outcomes are conditionally independent given the true category). This assumption, even if seldom strictly correct, can lead to high classification performance under a wide range of circumstances (Domingos & Pazzani, 1997); however, its implications for test selection need to be studied.

In addition, a number of empirical questions follow from the theoretical analyses of this article. One is the relationship between direct estimates of t_x (e.g., by doctors or other health care professionals) and corresponding t_x values computed from elicited situation-specific utilities. The possible match between these values would provide an important justification for the use of t_x in clinical practice to compute the value of tests, even in case the underlying utilities are unknown. Moreover, while use of likelihood differences has been studied empirically in domains such as causal reasoning (Cheng & Novick, 1990) and belief updating (Gigerenzer & Hoffrage, 1995), people's use of the LDH has not yet been tested in an explicitly medical diagnostic context or in related situations with external payoffs. Another question is how people select tests, as a function of how probability values and outcome utilities are communicated. The implications of outcome utilities for human test selection have been studied in only a very small number of articles (e.g., Baron & Hershey, 1988; Markant & Gureckis, 2012; Meder & Nelson, 2012). These studies suggest that it is difficult to make asymmetric payoffs intuitive in human test selection. It is possible that people follow different strategies according to whether the relevant probability values are learned through experience (Nelson et al., 2010) or presented using the standard probability format with explicit priors and likelihoods (in which case, the LDH is easy to apply) or via other numeric or graphical methods (Wu et al., 2017). One objective for future work should be to identify how to make utilities, probability values, and test characteristics intuitive, such that people take them into account appropriately when searching for information.

References

Alberg, A. J., Park, J. W., Hager, B. W., Brock, M. V., & Diener-West, M. (2004). The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. *Journal of General Internal Medicine*, *19*(5p1), 460–465. https://doi.org/10.1111/j.1525-1497.2004.30091.x

Al-Orainey, I. O. (2009). Diagnosis of latent tuberculosis: Can we do better? *Annals of Thoracic Medicine*, *4*(1), 5–9. https://doi.org/10.4103/1817-1737.44778

Austerweil, J. L., & Griffiths, T. L. (2011). Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science*, *35*(3), 499–526. https://doi.org/10.1111/j.1551-6709.2010.01161.x

Barnard, G. A., & Bayes, T. (1958). Studies in the history of probability and statistics: IX. Thomas Bayes's essay towards solving a problem in the doctrine of chances. *Biometrika*, *45*(3–4), 293–315. https://doi.org/10.2307/2333180 (Original work published 1763)

Baron, J. (1985). *Rationality and intelligence*. Cambridge University Press.

Baron, J., Beattie, J., & Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning: II. Congruence, information, and certainty. *Organizational Behavior and Human Decision Processes*, *42*(1), 88–110. https://doi.org/10.1016/0749-5978(88)90021-0

Baron, J., & Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning: I. Priors, error costs, and test accuracy. *Organizational Behavior and Human Decision Processes*, *41*(2), 259–279. https://doi.org/10.1016/0749-5978(88)90030-1

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chance. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418 (Reprinted from 1958, Biometrika, 45, 296–315). https://doi.org/10.1098/rstl.1763.0053

Benish, W. A. (1999). Relative entropy as a measure of diagnostic information. *Medical Decision Making*, *19*(2), 202–206. https://doi.org/10.1177/0272989X9901900211

Benish, W. A. (2003). Mutual information as an index of diagnostic test performance. *Methods of Information in Medicine*, *3*(03), 260–264.

Bentley T. G., Catanzaro A., & Ganiats T. G. (2012) Implications of the impact of prevalence on test thresholds and outcomes: Lessons from tuberculosis. *BMC Research Notes*, *5*(1), Article 563. https://doi.org/10.1186/1756-0500-5-563

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708–731. https://doi.org/10.1037/xlm0000061

Chamberlin, T. C. (1890). The method of multiple working hypotheses. *Science*, *15*, 92–96.

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, *58*(4), 545–567. https://doi.org/10.1037/0022-3514.58.4.545

Christensen-Szalanski, J. J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(4), 928–935. https://doi.org/10.1037//0096-1523.7.4.928

Coenen, A., Nelson, J. D., & Gureckis, T. (2019). Asking the right questions about human inquiry: Nine open challenges. *Psychonomic Bulletin and Review*, *26*(5), 1548–1587. https://doi.org/10.3758/s13423-018-1470-5

Crupi, V., Nelson, J. D., Meder, B., Cevolani, G. & Tentori, K. (2018). Generalized information theory meets human cognition: Introducing a unified framework to model uncertainty and information search. *Cognitive Science*, *42*(5), 1410–1456. https://doi.org/10.1111/cogs.12613

Crupi, V., Tentori, K., & Lombardi, L. (2009). Pseudodiagnosticity revisited. *Psychological Review*, *116*(4), 971–985. https://doi.org/10.1037/a0017050

Djulbegovic, B., van den Ende, J., Hamm, R. M., Mayrhofer, T., Hozo, I., Pauker, S. G., & for the International Threshold Working Group (ITWG). (2015). When is rational to order a diagnostic test, or prescribe treatment: The threshold model as an explanation of practice variation. *European Journal of Clinical Investigation*, *45*(5), 485–493. https://doi.org/10.1111/eci.12421

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *29*(2/3), 103–130. https://doi.org/10.1023/A:1007413511361

Domurat, A., Kowalczuk, O., Idzikowska, K., Borzymowska, Z., & Nowak-Przygodzka, M. (2015). Bayesian probability estimates are not necessary to make choices satisfying Bayes' rule in elementary situations. *Frontiers in Psychology*, *6*(340), Article 1194. https://doi.org/10.3389/fpsyg.2015.01194

Esmail, H., Barry, C. E., 3rd, Young, D. B., & Wilkinson, R. J. (2014). The ongoing challenge of latent tuberculosis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1645), Article 20130437. https://doi.org/10.1098/rstb.2013.0437

Filimon, F., Nelson, J. D., Sejnowski, T. J., Sereno, M. I., & Cottrell, G. W. (2020). The ventral striatum dissociates information expectation, reward anticipation, and reward receipt. *Proceedings of the National Academy of Sciences of the USA*, *117*(26), 15200–15208. https://doi.org/10.1073/pnas.1911778117

Geman, D., & Jedynak, B. (1996). An active testing model for tracking roads in satellite images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*(1), 1–14. https://doi.org/10.1109/34.476006

Geman, D., & Jedynak, B. (2001). Model-Based classification trees. *IEEE Transactions on Information Theory*, *47*(3), 1075–1082. https://doi.org/10.1109/18.915664

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704.

Good, I. J. (1950). *Probability and the weight of evidence*. Griffin.

Good, I. J. (1967). On the principle of total evidence. *The British Journal for the Philosophy of Science*, *17*(4), 319–321. http://www.jstor.org/stable/686773

Good, I. J., & Card, W. I. (1971). The diagnostic process with special reference to errors. *Methods of Information in Medicine*, *10*(03), 176–188. https://doi.org/10.1055/s-0038-1636045

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Gureckis, T. M., & Markant, D. B. (2012). Self-Directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464–481. https://doi.org/10.1177/1745691612454304

Hamm, R. M., Beasley, W. H., Johnson, W. J. (2014). A balance beam aid for instruction in clinical diagnostic reasoning. *Medical Decision Making*, *34*(7), 854–862. https://doi.org/10.1177/0272989x14529623

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, *73*(5), 538–540. https://doi.org/10.1097/00001888-199805000-00024

Hyafil, L., & Rivest, R. L. (1976). Constructing optimal binary decision trees is NP complete. *Information Processing Letters*, *5*(1), 15–17. https://doi.org/10.1016/0020-0190(76)90095-8

Jarecki, J. B., Meder, B., & Nelson, J. D. (2017). Naive and robust: Class-conditional independence in human classification learning. *Cognitive Science*, *42*(1), 4–42. https://doi.org/10.1111/cogs.12496

Lamberti, M., Muoio, M., Monaco, M. G. L., Uccello, R., Sannolo, N., Mazzarella, G., Garzillo, E. M., Arnese, A., La Cerra, G., & Coppola, N. (2014). Prevalence of latent tuberculosis infection and associated risk factors among 3,374 healthcare students in Italy. *Journal of Occupational Medicine and Toxicology*, *9*(1), Article 488. https://doi.org/10.1186/s12995-014-0034-5

Legge, G. E., Klitz, T. S., & Tjan, B. S. (1997). Mr. Chips: An ideal-observer model of reading. *Psychological Review*, *104*(3), 524–553. https://doi.org/10.1037/0033-295x.104.3.524

Liefgreen, A., Pilditch, T., & Lagnado, D. (2020). Strategies for selecting and evaluating information. *Cognitive Psychology*, *123*(3), Article 101332. https://doi.org/10.1016/j.cogpsych.2020.101332

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, *27*(4), 986–1005. https://doi.org/10.1214/aoms/1177728069

Maddox, W. T. (2002). Toward a unified theory of decision criterion learning in perceptual categorization. *Journal of the Experimental Analysis of Behavior*, *78*(3), 567–595. https://doi.org/10.1901/jeab.2002.78-567

Maddox, W. T., & Bohil, C. J. (1998). Base-Rate and payoff effects in multidimensional perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1459–1482. https://doi.org/10.1037/0278-7393.24.6.1459

Maddox, W. T., & Bohil, C. J. (2003). A theoretical framework for understanding the effects of simultaneous base-rate and payoff manipulations on decision criterion learning in perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 307–320. https://doi.org/10.1037/0278-7393.29.2.307

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94–122. https://doi.org/10.1037/a0032108

Markant, D. M., & Gureckis, T. M. (2012). Does the utility of information influence sampling behavior? In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 719–724). Cognitive Science Society.

Mazurek G. H., Jereb J., Vernon A., LoBue, P., Goldberg, S., & Castro, K. (2010). Updated guidelines for using interferon gamma release assays to detect Mycobacterium tuberculosis infection—United States, 2010. *MMWR Recommendations and Reports*, *59*(RR-5), 1–25. https://npin.cdc.gov/publication/updated-guidelines-using-interferon-gamma-release-assays-detect-mycobacterium

McDowell, M., Galesic, M., & Gigerenzer, G. (2018). Natural frequencies do foster public understanding of medical tests: Comment on Pighin, Gonzalez, Savadori, and Girotto (2016). *Medical Decision Making*, *38*(3), 390–399. https://doi.org/10.1177/0272989x18754508

Meder, B. M., Crupi, V., & Nelson, J. D. (in press). What makes a good question? Prospects for a comprehensive theory of human information acquisition. In I. Cogliati-Dezza, C. Wu & E. Schulz (Eds.), *The drive for knowledge: The science of human information-Seeking*. Cambridge University Press.

Meder, B., & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making*, *7*(2), 119–148. http://journal.sjdm.org/12/12314/jdm12314.html

Meder, B., Nelson, J. D., Jones, M. C., & Ruggeri, A. (2019). Stepwise versus globally optimal search in children and adults. *Cognition*, *191*, Article 103965. https://doi.org/10.1016/j.cognition.2019.05.002

Meier, K. M., & Blair, M. B. (2013). Waiting and weighting: Information sampling is a balance between efficiency and error-reduction. *Cognition*, *126*(2), 319–325. https://doi.org/10.1016/j.cognition.2012.09.014

Minka, T. P. (2001). *On the pathologies of orthodox statistics*. Retrieved October 6, 2021, from https://www.microsoft.com/en-us/research/publication/pathologies-orthodox-statistics/

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, *116*(3), 499–518. https://doi.org/10.1037/a0016104

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*, 387–391. https://doi.org/10.1038/nature03390

Nakamura, K. (2006). Neural representation of information measure in the primate premotor cortex. *Journal of Neurophysiology*, *96*, 478–485. https://doi.org/10.1152/jn.01326.2005

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*(4), 979–999. https://doi.org/10.1037/0033-295x.112.4.979

Nelson, J. D., & Cottrell, G. W. (2007). Probabilistic model of eye movements in concept formation. *Neurocomputing: An International Journal*, *70*(13–15), 2256–2272. https://doi.org/10.1016/j.neucom.2006.02.026

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. *Cognition*, *130*(1), 74–80. https://doi.org/10.1016/j.cognition.2013.09.007

Nelson, J. D., McKenzie, C. R. M., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science*, *21*(7), 960–969. https://doi.org/10.1177/0956797610372637

Nelson, J. D., Meder, B. & Jones, M. (2018). *Towards a theory of heuristic and optimal planning for sequential information search*. PsyArXiv. https://doi.org/10.31234/osf.io/bxdf4

Newell, A., & Card, S. K. (1985). The prospects for psychological science in human-computer interaction. *Human-Computer Interaction*, *1*(3), 209–242. https://doi.org/10.1207/s15327051hci0103_1

Nienhaus, A., Schablon, A., Preisser, A. M., Ringshausen, F. C., & Diel, R. (2014). Tuberculosis in healthcare workers—a narrative review from a German perspective. *Journal of Occupational Medicine and Toxicology*, *9*(1), Article 9. https://doi.org/10.1186/1745-6673-9-9

Nozick, R. (1981). *Philosophical explanations*. Harvard University Press.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631. https://doi.org/10.1037/0033-295X.101.4.608

Pauker, S. G., & Kassirer, J. P. (1975). Therapeutic decision making: A cost-benefit analysis. *New England Journal of Medicine*, *293*(5), 229–234. https://doi.org/10.1056/nejm197507312930505

Pauker, S. G., & Kassirer, J. P. (1980). The threshold approach to clinical decision making. *New England Journal of Medicine*, *302*(20), 1109–1117. https://doi.org/10.1056/nejm198005153022003

Popper, K. R. (1959). *The logic of scientific discovery*. Basic books.

Ruggeri, A., & Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. *Cognition*, *143*, 203–216. https://doi.org/10.1016/j.cognition.2015.07.004

Savage, L. J. (1954). *The foundations of statistics*. Wiley.

Schablon, A., Harling, M., Diel, R., & Nienhaus, A. (2010). Risk of latent TB infection in individuals employed in the healthcare sector in Germany: A multicentre prevalence study. *BMC Infectious Diseases*, *10*(1), Article 1. https://doi.org/10.1186/1471-2334-10-107

Schlaifer, R., & Raiffa, H. (1961). *Applied statistical decision theory*. Harvard University.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423, 623–656. https://doi.org/10.1145/584091.584093

Sharma, B., & Mittal, D. (1975). New non–additive measures of entropy for discrete probability distributions. *Journal of Mathematical Sciences (Delhi)*, *10*, 28–40.

Skov, R. B., & Sherman, S. J. (1986). Information-Gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology*, *22*(2), 93–121. https://doi.org/10.1016/0022-1031(86)90031-4

Slowiaczek, L. M., Klayman, J., Sherman, S. J., & Skov, R. B. (1992). Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory and Cognition*, *20*, 392–405. https://doi.org/10.3758/BF03210923

Smith, R., Cattamanchi, A., Steingart, K. R., Denkinger, C., Dheda, K., Winthrop, K. L., & Pai, M. (2011). Interferon-Gamma release assays for diagnosis of latent tuberculosis infection: Evidence in immune-mediated inflammatory disorders. *Current Opinion in Rheumatology*, *23*(4), 377–384. https://doi.org/10.1097/bor.0b013e3283474d62

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments and Computers*, *31*, 137–149. https://doi.org/10.3758/BF03207704

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489. https://doi.org/10.1207/s15516709cog2703_6

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist*, *47*(4), 522–532. https://doi.org/10.1037//0003-066x.47.4.522

Trommershäuser, J., Maloney, L. T., & Landy M. S. (2003a). Statistical decision theory and trade-offs in the control of motor response. *Spatial Vision*, *16*(3–4), 255–275. https://doi.org/10.1163/156856803322467527

Trommershäuser, J., Maloney, L. T., & Landy M. S. (2003b). Statistical decision theory and rapid, goal-directed movements. *Journal of the Optical Society of America*, *20*(7), 1419–1433. https://doi.org/10.1364/josaa.20.001419

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New Horizons in Psychology* (pp. 135–151). Penguin.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*, 273–281. https://doi.org/10.1080/14640746808400161

Weinstein, M. C., Torrance, G., & McGuire, A. (2009). QALYs: The basics. *Value in Health*, *12*(S1), S5–S9. https://doi.org/10.1111/j.1524-4733.2009.00515.x

Wells, G. L., & Lindsay, R. C. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin*, *88*(3), 776–784. https://doi.org/10.1037/0033-2909.88.3.776

World Health Organization (WHO). (2019). *Global tuberculosis report 2019*. Retrieved February 4, 2020, from https://apps.who.int/iris/bitstream/handle/10665/329368/9789241565714-eng.pdf

World Health Organization (WHO). (2020). *Global tuberculosis report 2020*. Retrieved August 4, 2021, from https://apps.who.int/iris/bitstream/handle/10665/336069/9789240013131-eng.pdf

Wu, C. M., Meder, B., Filimon, F., & Nelson, J. D. (2017). Asking better questions: How presentation formats influence information search. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(8), 1274–1297. https://doi.org/10.1037/xlm0000374

Zhang, X., Jia, H., Liu, F., Pan, L., Xing, A., Gu, S., Du, B., Sun, Q., Wei, R. & Zhang, Z. (2013). Prevalence and risk factors for latent tuberculosis infection among health care workers in China: A cross-sectional study. *PLOS ONE*, *8*(6), Article e66412. https://doi.org/10.1371/journal.pone.0066412

(*Appendix follows*)

Appendix

Proofs of Mathematical Results

Proof of Result 1. Let $H = \{h, \neg h\}$ be a binary hypothesis space associated with a disease and let $E = \{e, \neg e\}$ be a diagnostic test for H. Let $u = \begin{bmatrix} u_{tp} & u_{fp} \\ u_{fn} & u_{tn} \end{bmatrix}$ be a utility function for H. Then the expected utility of E with respect to the utility function u, which we denote by $eu(E)$, is equal to the expected utility of E with respect to the utility function $u^* = \begin{bmatrix} u_{tp} - u_{fn} & 0 \\ 0 & u_{tn} - u_{fp} \end{bmatrix}$, which we denote by $eu^*(E)$; that is, $eu(E) = eu^*(E)$.

Proof. Let us denote the utilities of e and $\neg e$ with respect to u by $u(e)$ and $u(\neg e)$ and the utilities of e and $\neg e$ with respect to u^* by $u^*(e)$ and $u^*(\neg e)$. We first calculate the utility of a positive test result for the payoff function u^*:

$$\begin{aligned} u^*(e) &= \max[P(h|e)(u_{tp} - u_{fn}), P(\neg h|e)(u_{tn} - u_{fp})] - \max[P(h)(u_{tp} - u_{fn}), P(\neg h)(u_{tn} - u_{fp})] \\ &= \max[P(h|e)u_{tp} + P(\neg h|e)u_{fp}, P(\neg h|e)u_{tn} + P(h|e)u_{fn}] \\ &\quad - \max[P(h)u_{tp} + P(\neg h)u_{fp}, P(\neg h)u_{tn} + P(h)u_{fn}] \\ &\quad - P(\neg h|e)u_{fp} - P(h|e)u_{fn} + P(\neg h)u_{fp} + P(h)u_{fn} \\ &= u(e) + (P(\neg h) - P(\neg h|e))u_{fp} + (P(h) - P(h|e))u_{fn} \\ &= u(e) + (P(h) - P(h|e))(u_{fn} - u_{fp}). \end{aligned} \tag{A1}$$

In the same way, we calculate the utility of a negative test result and get:

$$u^*(\neg e) = u(\neg e) + (P(h) - P(h|\neg e))(u_{fn} - u_{fp}). \tag{A2}$$

To calculate the expected utility $eu^*(E)$ of Test E, we weigh the utilities of the test results with the probabilities $P(e)$ and $P(\neg e)$ of the test results and use the above relationship between the utilities with respect to the two payoff functions:

$$\begin{aligned} eu^*(E) &= P(e)u^*(e) + P(\neg e)u^*(\neg e) \\ &= P(e)[u(e) + (P(h) - P(h|e))(u_{fn} - u_{fp})] + P(\neg e)[u(\neg e) \\ &\quad + (P(h) - P(h|\neg e))(u_{fn} - u_{fp})] \\ &= eu(E) + P(e)(P(h) - P(h|e))(u_{fn} - u_{fp}) + P(\neg e)(P(h) - P(h|\neg e))(u_{fn} - u_{fp}) \\ &= eu(E) + [P(e)P(h) - P(e \wedge h) + P(\neg e)P(h) - P(\neg e \wedge h)](u_{fn} - u_{fp}) \\ &= eu(E) + [(P(e) + P(\neg e))P(h) - P(e \wedge h) - P(\neg e \wedge h)](u_{fn} - u_{fp}) \\ &= eu(E) + (P(h) - P(h))(u_{fn} - u_{fp}) \\ &= eu(E). \end{aligned} \tag{A3}$$

□

Proof of Result 2. Let $H = \{h, \neg h\}$ be the hypothesis space associated with a disease and let u be a proper utility function with corresponding therapeutic threshold t_x. Let E be a diagnostic test for H. Then the test is useful, that is, $eu(E) > 0$, iff both $P(h|\neg e) < t_x$ and $P(h|e) > t_x$.

(*Appendix continues*)

Proof. Without loss of generality (see Result 1) we can assume that $u_{fn} = u_{fp} = 0$.

We will first show that if $eu(E) > 0$, it follows that $P(h|\neg e) < t_x$ and $P(h|e) > t_x$. If under both test results the same option as before is at least as useful as the other, then $eu(E) = 0$. This can easily be seen by calculating $eu(E)$ and choosing the same alternative (to treat or not to treat) both before the test result is known and afterward, independent of the particular test result. If $eu(E) > 0$, then the decisions under the two test results have to be different, which is only the case if a negative test result leads to an updated probability of the person having the disease strictly below the therapeutic threshold, that is, $P(h|\neg e) < t_x$, and a positive test result leads to an updated probability of the person having the disease strictly above the therapeutic threshold, that is, $P(h|e) > t_x$.

To show that from $P(h|\neg e) < t_x$ and $P(h|e) > t_x$, it follows that $eu(E) > 0$, consider the following: For $P(h) \le t_x$, the expected utility is given by

$$\begin{aligned} eu(E) &= P(e)P(h|e)u_{tp} + P(\neg e)P(\neg h|\neg e)u_{tn} - P(\neg h)u_{tn} \\ &> P(e)P(\neg h|e)u_{tn} + P(\neg e)P(\neg h|\neg e)u_{tn} - P(\neg h)u_{tn} = 0. \end{aligned} \tag{A4}$$

The expected utility is thus strictly larger than 0. For $P(h) \ge t_x$, the expected utility can be calculated as

$$\begin{aligned} eu(E) &= P(e)P(h|e)u_{tp} + P(\neg e)P(\neg h|\neg e)u_{tn} - P(h)u_{tp} \\ &> P(e)P(h|e)u_{tp} + P(\neg e)P(h|\neg e)u_{tp} - P(h)u_{tp} = 0. \end{aligned} \tag{A5}$$

Again, $eu(E)$ is strictly larger than 0. □

Proof of Result 3. Let $H = \{h, \neg h\}$ be a binary hypothesis space associated with a disease and let $E = \{e, \neg e\}$ be a diagnostic test for H. Let $u = \begin{bmatrix} u_{tp} & u_{fp} \\ u_{fn} & u_{tn} \end{bmatrix}$ and $u' = \begin{bmatrix} \alpha u_{tp} & \alpha u_{fp} \\ \alpha u_{fn} & \alpha u_{tn} \end{bmatrix}$, where $\alpha > 0$, be two utility functions for H. Then the expected utility of Test E under u' is α times the expected utility of E under u; that is, $eu'(E) = \alpha\, eu(E)$.

Proof. Since

$$\begin{aligned} u_{\begin{bmatrix} \alpha\cdot u_{tp} & \alpha\cdot u_{fp} \\ \alpha\cdot u_{fn} & \alpha\cdot u_{tn} \end{bmatrix}}(e) &= \max[P(h|e)\alpha u_{tp} + P(\neg h|e)\alpha u_{fp}, P(\neg h|e)\alpha u_{tn} + P(h|e)\alpha u_{fn}] \\ &\quad - \max[P(h)\alpha u_{tp} + P(\neg h)\alpha u_{fp}, P(\neg h)\alpha u_{tn} + P(h)\alpha u_{fn}] \\ &= \max[\alpha(P(h|e)u_{tp} + P(\neg h|e)u_{fp}), \alpha(P(\neg h|e)u_{tn} + P(h|e)u_{fn})] \\ &\quad - \max[\alpha(P(h)u_{tp} + P(\neg h)u_{fp}), \alpha(P(\neg h)u_{tn} + P(h)u_{fn})] \\ &= \alpha \cdot \max[P(h|e)u_{tp} + P(\neg h|e)u_{fp}, P(\neg h|e)u_{tn} + P(h|e)u_{fn}] \\ &\quad - \alpha \cdot \max[P(h)u_{tp} + P(\neg h)u_{fp}, P(\neg h)u_{tn} + P(h)u_{fn})] \\ &= \alpha \cdot u_{\begin{bmatrix} u_{tp} & u_{fp} \\ u_{fn} & u_{tn} \end{bmatrix}}(e), \end{aligned} \tag{A6}$$

and equally,

$$u_{\begin{bmatrix} \alpha\cdot u_{tp} & \alpha\cdot u_{fp} \\ \alpha\cdot u_{fn} & \alpha\cdot u_{tn} \end{bmatrix}}(\neg e) = \alpha \cdot u_{\begin{bmatrix} u_{tp} & u_{fp} \\ u_{fn} & u_{tn} \end{bmatrix}}(\neg e), \tag{A7}$$

(*Appendix continues*)

thus

$$
\begin{aligned}
eu_{\begin{bmatrix}\alpha\cdot u_{tp} & \alpha\cdot u_{fp}\\ \alpha\cdot u_{fn} & \alpha\cdot u_{tn}\end{bmatrix}}(E) &= P(e)u_{\begin{bmatrix}\alpha\cdot u_{tp} & \alpha\cdot u_{fp}\\ \alpha\cdot u_{fn} & \alpha\cdot u_{tn}\end{bmatrix}}(e) + P(\neg e)u_{\begin{bmatrix}\alpha\cdot u_{tp} & \alpha\cdot u_{fp}\\ \alpha\cdot u_{fn} & \alpha\cdot u_{tn}\end{bmatrix}}(\neg e)\\
&= \alpha \cdot P(e)u_{\begin{bmatrix}u_{tp} & u_{fp}\\ u_{fn} & u_{tn}\end{bmatrix}}(e) + \alpha \cdot P(\neg e)u_{\begin{bmatrix}u_{tp} & u_{fp}\\ u_{fn} & u_{tn}\end{bmatrix}}(\neg e)\\
&= \alpha \cdot eu_{\begin{bmatrix}u_{tp} & u_{fp}\\ u_{fn} & u_{tn}\end{bmatrix}}(E). \qquad \text{(A8)}
\end{aligned}
$$

□

Proof of Result 4. Let $H = \{h, \neg h\}$ be a hypothesis space associated with a disease and let u be a proper utility function for H with corresponding therapeutic threshold t_x. If $P(h) = t_x$, then there exists a positive β, such that for any diagnostic Test E for H, we have $eu(E) = \beta\lambda(E)$.

Proof. Without loss of generality we can assume that $u_{fn} = u_{fp} = 0$ by Result 1. Since $P(h) = t_x$ by assumption, we have $P(h)u_{tp} = P(\neg h)u_{tn}$, i.e., the utilities associated with choosing h and choosing $\neg h$ are equal. We posit $\beta = P(h)u_{tp} = P(\neg h)u_{tn}$. Note that β is strictly positive since $P(h) > 0$ and $u_{tp} > u_{fn} = 0$. Also, β is independent from Test E as it is only affected by the prior probability and the utilities. Let us first look at the utilities of test results e and $\neg e$. The test results are labeled such that $P(h|e) \geq P(h)$ and thus $P(\neg h|e) \leq P(\neg h)$, by which

$$P(h|e)u_{tp} \geq P(h)u_{tp} = P(\neg h)u_{tn} \geq P(\neg h|e)u_{tn}. \qquad \text{(A9)}$$

This, in turn, implies

$$u(e) = \max[P(h|e)u_{tp}, P(\neg h|e)u_{tn}] - \max[P(h)u_{tp}, P(\neg h)u_{tn}] = P(h|e)u_{tp} - \beta. \qquad \text{(A10)}$$

Regarding negative test result $\neg e$ we have $P(\neg h|\neg e) \geq P(\neg h)$, and thus $P(h|\neg e) \leq P(h)$, by which

$$P(\neg h|\neg e)u_{tn} \geq P(\neg h)u_{tn} = P(h)u_{tp} \geq P(h|\neg e)u_{tp}, \qquad \text{(A11)}$$

which in turn implies

$$u(\neg e) = \max[P(\neg h|\neg e)u_{tn}, P(h|\neg e)u_{tp}] - \max[P(h)u_{tp}, P(\neg h)u_{tn}] = P(\neg h|\neg e)u_{tn} - \beta. \qquad \text{(A12)}$$

Putting all the foregoing together, we compute:

$$
\begin{aligned}
eu(E) &= u(e)P(e) + u(\neg e)P(\neg e)\\
&= [P(h|e)u_{tp} - \beta]P(e) + [P(\neg h|\neg e)u_{tn} - \beta]P(\neg e)\\
&= P(h|e)P(e)u_{tp} + P(\neg h|\neg e)P(\neg e)u_{tn} - \beta[P(e) + P(\neg e)]\\
&= P(e|h)P(h)u_{tp} + P(\neg e|\neg h)P(\neg h)u_{tn} - \beta\\
&= \beta P(e|h) + \beta P(\neg e|\neg h) - \beta\\
&= \beta P(e|h) + \beta - \beta P(e|\neg h) - \beta\\
&= \beta\lambda(E). \qquad \text{(A13)}
\end{aligned}
$$

□

(*Appendix continues*)

Proof of Result 5. Let $H = \{h, \neg h\}$ be a hypothesis space associated with a disease and let u be a proper set of utility values for H with corresponding therapeutic threshold t_x. Then if $P(h) \neq t_x$, there exist Tests E and F such that $\lambda(E) > \lambda(F)$ while $eu(E) < eu(F)$.

Proof. Without loss of generality we can assume that $u_{fn} = u_{fp} = 0$ by Result 1. We will show a general method to generate a pair of binary Tests E and F such that $\lambda(E) > \lambda(F)$. while $eu(E) < eu(F)$. More precisely, we will define an informative but useless Test E, that is, a test with $P(h|\neg e) < P(h) < P(h|e)$ and $eu(E) = 0$, and a useful Test F such that $\lambda(E)$ exceeds $\lambda(F)$ by some tiny amount $\varepsilon > 0$. We will prove this separately for the two subcases of $P(h) \neq t_x$, namely $P(h) < t_x$ and $P(h) > t_x$.

Case 1: $P(h) < t_x$. Let us posit:

$$P(e|h) = 1. \tag{A14}$$

$$P(e|\neg h) = \frac{P(h)u_{tp}}{P(\neg h)u_{tn}}. \tag{A15}$$

$$P(f|h) = 1 - \left[\frac{P(h)u_{tp}}{P(\neg h)u_{tn}} + \varepsilon\right]. \tag{A16}$$

$$P(f|\neg h) = 0. \tag{A17}$$

Note that $\frac{P(h)u_{tp}}{P(\neg h)u_{tn}} < 1$, because by assumption $P(h) < t_x$ and hence $P(h)u_{tp} < P(\neg h)u_{tn}$. To ensure that $P(f|h) > 0$, ε must be chosen so that $\varepsilon < 1 - \frac{P(h)u_{tp}}{P(\neg h)u_{tn}}$. As $1 - \frac{P(h)u_{tp}}{P(\neg h)u_{tn}}$ is strictly positive, one can always find a positive but tiny enough value of ε to satisfy this condition. Clearly, the likelihood difference heuristic favors Test E over F, because

$$\lambda(E) - \lambda(F) = 1 - \frac{P(h)u_{tp}}{P(\neg h)u_{tn}} - \left[1 - \left(\frac{P(h)u_{tp}}{P(\neg h)u_{tn}} + \varepsilon\right)\right] = \varepsilon > 0. \tag{A18}$$

However, as we will now show, expected utility computations imply the opposite ranking, that is, $eu(E) < eu(F)$. More precisely, $eu(E) = 0$, while $eu(F)$ is strictly positive. Let us first show that $eu(E) = 0$. Given that $P(h) < t_x$, we have $P(h)u_{tp} < P(\neg h)u_{tn}$, hence $\max[P(h)u_{tp}, P(\neg h)u_{tn}] = P(\neg h)u_{tn}$. Moreover, given that $P(\neg h|\neg e) > P(\neg h)$ and thus $P(h|\neg e) < P(h)$, we have:

$$P(\neg h|\neg e)u_{tn} > P(\neg h)u_{tn} > P(h)u_{tp} > P(h|\neg e)u_{tp}, \tag{A19}$$

which implies:

$$\begin{aligned} u(\neg e) &= \max[P(h|\neg e)u_{tp}, P(\neg h|\neg e)u_{tn}] - \max[P(h)u_{tp}, P(\neg h)u_{tn}] \\ &= P(\neg h|\neg e)u_{tn} - P(\neg h)u_{tn}. \end{aligned} \tag{A20}$$

As for the positive test result e, note that by Bayes's theorem and our probability assignments:

$$\begin{aligned} P(h|e) &= \frac{P(e|h)P(h)}{P(e|h)P(h) + P(e|\neg h)P(\neg h)} \\ &= \frac{1 \cdot P(h)}{1 \cdot P(h) + \frac{P(h)u_{tp}}{P(\neg h)u_{tn}}P(\neg h)} \\ &= \frac{P(h)}{P(h)(1 + \frac{u_{tp}}{u_{tn}})} \\ &= \frac{u_{tn}}{u_{tp} + u_{tn}}, \end{aligned} \tag{A21}$$

(*Appendix continues*)

which simply equals t_x (because $u_{fp} = u_{fn} = 0$). So $P(h|e) = t_x$, and thus $P(h|e)u_{tp} = P(\neg h|e)u_{tn}$, which in turn implies:

$$u(e) = \max[P(h|e)u_{tp}, P(\neg h|e)u_{tn}] - \max[P(h)u_{tp}, P(\neg h)u_{tn}]$$
$$= P(\neg h|e)u_{tn} - P(\neg h)u_{tn}. \tag{A22}$$

Putting all the foregoing together:

$$\begin{aligned} eu(E) &= u(e)P(e) + u(\neg e)P(\neg e) \\ &= [P(\neg h|e)u_{tn} - P(\neg h)u_{tn}]P(e) + [P(\neg h|\neg e)u_{tn} - P(\neg h)u_{tn}]P(\neg e) \\ &= [P(\neg h|e)P(e) - P(\neg h)P(e) + P(\neg h|\neg e)P(\neg e) - P(\neg h)P(\neg e)]u_{tn} \\ &= [P(\neg h) - P(\neg h)]u_{tn} = 0. \end{aligned} \tag{A23}$$

Regarding Test F, we have once again $P(\neg h|\neg f)u_{tn} > P(\neg h)u_{tn} > P(h)u_{tp} > P(h|\neg f)u_{tp}$, and thus:

$$u(\neg f) = \max[P(h|\neg f)u_{tp}, P(\neg h|\neg f)u_{tn}] - \max[P(h)u_{tp}, P(\neg h)u_{tn}]$$
$$= P(\neg h|\neg f)u_{tn} - P(\neg h)u_{tn}. \tag{A24}$$

But, on the other hand, because $P(f|\neg h) = 0$:

$$\begin{aligned} P(h|f) &= \frac{P(f|h)P(h)}{P(f|h)P(h) + P(f|\neg h)P(\neg h)} \\ &= 1 > \frac{u_{tn}}{u_{tp} + u_{tn}} = t_x, \end{aligned} \tag{A25}$$

so $P(h|f) > t_x$, and thus $P(h|f)u_{tp} > P(\neg h|f)u_{tn}$, by which in turn:

$$u(f) = \max[P(h|f)u_{tp}, P(\neg h|f)u_{tn}] - \max[P(h)u_{tp}, P(\neg h)u_{tn}] = P(h|f)u_{tp} - P(\neg h)u_{tn}, \tag{A26}$$

and:

$$\begin{aligned} &P(h|f)u_{tp}P(f) + P(\neg h|\neg f)u_{tn}P(\neg f) - P(\neg h)u_{tn} > P(\neg h|f)u_{tn}P(f) + P(\neg h|\neg f)u_{tn}P(\neg f) \\ &- P(\neg h)u_{tn}P(h|f)u_{tp}P(f) + P(\neg h|\neg f)u_{tn}P(\neg f) - P(\neg h)u_{tn} > 0P(h|f)u_{tp} \\ &P(f) + P(\neg h|\neg f)u_{tn}P(\neg f) - P(\neg h)u_{tn}(P(f) + P(\neg f)) > 0 \\ &u(f)P(f) + u(\neg f)P(\neg f) > 0 \\ &eu(F) > 0. \end{aligned} \tag{A27}$$

Case 2: $P(h) > t_x$. In this case, let us thus posit:

$$P(e|h) = 1 - \frac{P(\neg h)u_{tn}}{P(h)u_{tp}}. \tag{A28}$$

$$P(e|\neg h) = 0. \tag{A29}$$

$$P(f|h) = 1. \tag{A30}$$

(*Appendix continues*)

$$P(f|\neg h) = \frac{P(\neg h)u_{tn}}{P(h)u_{tp}} + \varepsilon. \tag{A31}$$

Note that $\frac{P(\neg h)u_{tn}}{P(h)u_{tp}} < 1$, because by assumption $P(h) > t_x$, so $P(h)u_{tp} > P(\neg h)u_{tn}$. To ensure that $P(f|\neg h) < 1$, ε must be chosen so that $\varepsilon < 1 - \frac{P(\neg h)u_{tn}}{P(h)u_{tp}}$. As $1 - \frac{P(\neg h)u_{tn}}{P(h)u_{tp}}$ is strictly positive, one can always find a positive but tiny enough value of ε to satisfy this condition. Clearly, the likelihood difference heuristic favors Test E over F, because

$$\lambda(E) - \lambda(F) = 1 - \frac{P(\neg h)u_{tn}}{P(h)u_{tp}} - \left[1 - \left(\frac{P(\neg h)u_{tn}}{P(h)u_{tp}} + \varepsilon\right)\right] = \varepsilon > 0. \tag{A32}$$

However, as we will now show, expected utility computations imply the opposite ranking, that is, $eu(E) < eu(F)$. More precisely, $eu(E) = 0$, while $eu(F)$ is strictly positive. Let us first show that $eu(E) = 0$. Given that $P(h) > t_x$, we have $P(h)u_{tp} > P(\neg h)u_{tn}$, so $\max[P(h)u_{tp}, P(\neg h)u_{tn}] = P(h)u_{tp}$. Moreover, given that $P(h|e) > P(h)$ and thus $P(\neg h|e) < P(\neg h)$, we also have: $P(h|e)u_{tp} > P(h)u_{tp} > P(\neg h)u_{tn} > P(\neg h|e)u_{tn}$ which implies:

$$\begin{aligned} u(e) &= max[P(h|e)u_{tp}, P(\neg h|e)u_{tn}] - max[P(h)u_{tp}, P(\neg h)u_{tn}] \\ &= P(h|e)u_{tp} - P(h)u_{tp}. \end{aligned} \tag{A33}$$

As for the negative test result $\neg e$, note that by Bayes's theorem and our probability assignments:

$$\begin{aligned} P(\neg h|\neg e) &= \frac{P(\neg e|\neg h)P(\neg h)}{P(\neg e|\neg h)P(\neg h) + P(\neg e|h)P(h)} \\ &= \frac{P(\neg h)}{P(\neg h) + \frac{P(\neg h)u_{tn}}{P(h)u_{tp}}P(h)} \\ &= \frac{P(\neg h)}{P(\neg h)(1 + \frac{u_{tn}}{u_{tp}})} \\ &= \frac{u_{tp}}{u_{tp} + u_{tn}}, \end{aligned} \tag{A34}$$

which simply equals t_x (because $u_{fn} = u_{fp} = 0$). So $P(\neg h|\neg e) = t_x$, and thus $P(\neg h|\neg e)u_{tn} = P(h|\neg e)u_{tp}$. This in turn implies:

$$\begin{aligned} u(\neg e) &= \max[P(h|\neg e)u_{tp}, P(\neg h|\neg e)u_{tn}] - \max[P(h)u_{tp}, P(\neg h)u_{tn}] \\ &= P(h|\neg e)u_{tp} - P(h)u_{tp}. \end{aligned} \tag{A35}$$

Putting all the foregoing together yields:

$$\begin{aligned} eu(E) &= u(e)P(e) + u(\neg e)P(\neg e) \\ &= (P(h|e)u_{tp} - P(h)u_{tp})P(e) + (P(h|\neg e)u_{tp} - P(h)u_{tp})P(\neg e) \\ &= (P(h|e)P(e) - P(h)P(e) + P(h|\neg e)P(\neg e) - P(h)P(\neg e))u_{tp} \\ &= (P(h) - P(h))u_{tp} = 0. \end{aligned} \tag{A36}$$

(*Appendix continues*)

Regarding Test F, we still have $P(h|f)u_{tp} > P(h)u_{tp} > P(\neg h)u_{tn} > P(\neg h|f)u_{tn}$ and therefore:

$$u(f) = \max[P(h|f)u_{tp}, P(\neg h|f)u_{tn}] - \max[P(h)u_{tp}, P(\neg h)u_{tn}] = P(h|f)u_{tp} - P(h)u_{tp}. \tag{A37}$$

But on the other hand, because $P(\neg f|h) = 0$:

$$P(h|\neg f) = \frac{P(\neg f|h)P(h)}{P(\neg e|h)P(h) + P(\neg f|\neg h)P(\neg h)} = 0 < \frac{u_{tn}}{u_{tp} + u_{tn}} = t_x. \tag{A38}$$

So $P(h|\neg f) < t_x$, and thus $P(h|\neg f)u_{tp} < P(\neg h|\neg f)u_{tn}$, therefore

$$\begin{aligned}
&P(h|f)u_{tp}P(f) + P(\neg h|\neg f)u_{tn}P(\neg f) - P(h)u_{tp} > P(h|f)u_{tp}P(f) \\
&\qquad + P(h|\neg f)u_{tp}P(\neg f) - P(h)u_{tp} \\
&P(h|f)u_{tp}P(f) + P(\neg h|\neg f)u_{tn}P(\neg f) - P(h)u_{tp} > 0 \\
&P(h|f)u_{tp}P(f) + P(\neg h|\neg f)u_{tn}P(\neg f) - P(h)u_{tp}(P(f) + P(\neg f)) > 0 \\
&u(f)P(f) + u(\neg f)P(\neg f) > 0 \\
&eu(F) > 0.
\end{aligned} \tag{A39}$$

□

Proof of Result 6. Let $H = (h, \neg h)$ be a hypothesis space associated with a disease and let E and F be two diagnostic tests for H with equal likelihood difference, $\lambda(E) = \lambda(F)$. Then the following holds:

i. If $P(\neg e|\neg h) > P(\neg f|\neg h)$, then $P(h|e) > P(h|f)$; that is, if E has higher specificity than F, then E has a higher positive predictive value than F.
ii. If $P(e|h) > P(f|h)$, then $P(\neg h|\neg e) > P(\neg h|\neg f)$; that is, if E has higher sensitivity than F, then E has a higher negative predictive value than F.

Proof. (i) Specificity clause: To prove the first part of this result, we will calculate the posterior $P(h|e)$ as a function of λ, the specificity $P(\neg e|\neg h)$ and the base rate $P(h)$. We set $\alpha = 1 - P(\neg e|\neg h) = P(e|\neg h)$. Then the sensitivity of E is given by $P(e|h) = \lambda + \alpha$. Given this, the posterior probability $P(h|e)$ can be calculated as

$$\begin{aligned}
P(h|e) &= \frac{P(e|h)P(h)}{P(e)} \\
&= \frac{P(e|h)P(h)}{P(e|h)P(h) + P(e|\neg h)P(\neg h)} \\
&= \frac{(\lambda + \alpha)P(h)}{(\lambda + \alpha)P(h) + \alpha(1 - P(h))} \\
&= \frac{\lambda P(h) + \alpha P(h)}{\lambda P(h) + \alpha P(h) + \alpha - \alpha P(h))} \\
&= \frac{\lambda P(h) + \alpha P(h)}{\lambda P(h) + \alpha}.
\end{aligned} \tag{A40}$$

Suppose we have two tests E and F with $\alpha_1 = 1 - P(\neg e|\neg h)$ and $\alpha_2 = 1 - P(\neg f|\neg h)$. If $P(\neg e|\neg h) > P(\neg f|\neg h)$, that is, if E has higher specificity than F, then $\alpha_1 < \alpha_2$. For $\alpha_1 < \alpha_2$, we have

(*Appendix continues*)

$$\begin{aligned}
&\alpha_1\lambda - \alpha_1\lambda P(h) < \alpha_2\lambda - \alpha_2\lambda P(h)\\
&\alpha_1\lambda + \alpha_2\lambda P(h) < \alpha_2\lambda + \alpha_1\lambda P(h)\\
&\lambda^2 P(h) + \alpha_1\lambda + \alpha_2\lambda P(h) + \alpha_1\alpha_2 < \lambda^2 P(h) + \alpha_2\lambda + \alpha_1\lambda P(h) + \alpha_1\alpha_2\\
&(\lambda P(h) + \alpha_1)(\lambda + \alpha_2) < (\lambda P(h) + \alpha_2)(\lambda + \alpha_1)\\
&\frac{\lambda + \alpha_1}{\lambda P(h) + \alpha_1} > \frac{\lambda + \alpha_2}{\lambda P(h) + \alpha_2}\\
&\frac{\lambda P(h) + \alpha_1 P(h)}{\lambda P(h) + \alpha_1} > \frac{\lambda P(h) + \alpha_2 P(h)}{\lambda P(h) + \alpha_2}\\
&P(h|e) > P(h|f).
\end{aligned} \tag{A41}$$

Thus, for two Tests E and F with $spec(E) > spec(F)$, we have $P(h|e) > P(h|f)$.

(ii) Sensitivity clause: Analogous to the first part of this result, we calculate the posterior $P(h|\neg e)$ as a function of λ, the sensitivity $P(e|h)$ and the base rate $P(h)$. We set $\beta = 1 - P(e|h) = P(\neg e|h)$. Then the specificity of E is given by $P(\neg e|\neg h) = \lambda + \beta$. Calculating the posterior $P(h|\neg e)$ yields

$$\begin{aligned}
P(h|\neg e) &= \frac{P(\neg e|h)P(h)}{P(\neg e)} = \frac{P(\neg e|h)P(h)}{P(\neg e|h)P(h) + P(\neg e|\neg h)P(\neg h)} = \frac{\beta P(h)}{\beta P(h) + (\beta + \lambda)(1 - P(h))}\\
&= \frac{\beta P(h)}{\beta P(h) + \beta + \lambda - \beta P(h) - \lambda P(h)} = \frac{\beta P(h)}{\beta + \lambda(1 - P(h))}.
\end{aligned} \tag{A42}$$

Suppose we have two Tests E and F with $\beta_1 = 1 - P(e|h)$ and $\beta_2 = 1 - P(f|h)$. If $P(e|h) > P(f|h)$, that is, if E has higher sensitivity than F, then $\beta_1 < \beta_2$. For $\beta_1 < \beta_2$, we have

$$\begin{aligned}
&\beta_1\lambda(1 - P(h)) < \beta_2\lambda(1 - P(h))\\
&\beta_1\beta_2 + \beta_1\lambda(1 - P(h)) < \beta_1\beta_2 + \beta_2\lambda(1 - P(h))\\
&\beta_1[\beta_2 + \lambda(1 - P(h))] < \beta_2[\beta_1 + \lambda(1 - P(h))]\\
&\frac{\beta_1}{\beta_1 + \lambda(1 - P(h))} < \frac{\beta_2}{\beta_2 + \lambda(1 - P(h))}\\
&\frac{\beta_1 P(h)}{\beta_1 + \lambda(1 - P(h))} < \frac{\beta_2 P(h)}{\beta_2 + \lambda(1 - P(h))}\\
&P(h|\neg e) < P(h|\neg f).
\end{aligned} \tag{A43}$$

Thus, for two Tests E and F with $sens(E) > sens(F)$, we have $P(h|\neg e) < P(h|\neg f)$. □

Proof of Result 7. Let $H = (h, \neg h)$ be the hypothesis space associated with a disease and let u be a proper utility function for H. Let $E = (e, \neg e)$ and $F = (f, \neg f)$ be two diagnostic tests for H, with $\lambda(E) = \lambda(F)$. Then the following holds:

i. If $P(h) < t_x$, and $P(\neg e|\neg h) \geq P(\neg f|\neg h)$, we have $eu(E) \geq eu(F)$.
ii. If $P(h) > t_x$, and $P(e|h) \geq P(f|h)$, we have $eu(E) \geq eu(F)$.

Proof. Without loss of generality, we can assume that $u_{fp} = u_{fn} = 0$. We will prove this theorem separately for pairs of useless tests, pairs of useful tests, and pairs where one test is useful and one is not.

(*Appendix continues*)

Case 1: Suppose we have two useless tests $E = (e, \neg e)$ and $F = (f, \neg f)$. Then both tests have an expected utility of zero, so it does not matter which of the two tests we choose and therefore both (i) and (ii) hold for these two tests.

Case 2: To prove that the theorem is true for pairs of useful tests, we will show that the expected utility of a test can be calculated from $\lambda(E)$ and a positive multiple of the test's specificity for $P(h) < t_x$ and from $\lambda(E)$ and a positive multiple of the test's sensitivity for $P(h) > t_x$. Suppose we have a useful diagnostic Test $E = \{e, \neg e\}$.

i. $P(h) < t_x$: For a base rate $P(h)$ below the therapeutic threshold t_x, before knowing the test result acting on the hypothesis $\neg h$ that the disease is absent maximizes expected utility. Since Test E is useful, we have $P(h|e) > t_x > P(h|\neg e)$ and the expected utility of E is thus given by

$$eu(E) = u_{tp}P(h|e)P(e) + u_{tn}P(\neg h|\neg e)P(\neg e) - u_{tn}P(\neg h). \tag{A44}$$

By elementary probability calculations, we can rewrite $eu(E)$ as

$$\begin{aligned} eu(E) &= u_{tp}P(e|h)P(h) + u_{tn}P(\neg e|\neg h)P(\neg h) - u_{tn}[P(e|\neg h) + P(\neg e|\neg h)]P(\neg h) \\ &= u_{tp}P(e|h)P(h) - u_{tn}P(e|\neg h)P(\neg h) \\ &= u_{tp}P(h)[1 - P(\neg e|h)] - u_{tn}P(\neg h)[1 - P(\neg e|\neg h)] \\ &= u_{tp}P(h) - u_{tp}P(h)P(\neg e|h) - u_{tn}P(\neg h) + u_{tn}P(\neg h)P(\neg e|\neg h). \end{aligned} \tag{A45}$$

We can rewrite $P(\neg e|h)$ as

$$1 - P(e|h) = P(\neg e|\neg h) + P(e|\neg h) - P(e|h) = spec(E) - \lambda(E). \tag{A46}$$

This yields:

$$\begin{aligned} eu(E) &= u_{tp}P(h) - u_{tp}P(h)[spec(E) - \lambda(E)] - u_{tn}P(\neg h) + u_{tn}P(\neg h)spec(E) \\ &= u_{tp}P(h)\lambda(E) + spec(E)[u_{tn}P(\neg h) - u_{tp}P(h)] + u_{tp}P(h) - u_{tn}P(\neg h) \\ &= u_{tp}P(h)\lambda(E) + [spec(E) - 1][u_{tn}P(\neg h) - u_{tp}P(h)]. \end{aligned} \tag{A47}$$

Note that $u_{tn}P(\neg h) - u_{tp}P(h)$ is positive, as $P(h) < t_x$. Thus, the larger the specificity of E, the larger the expected utility.

ii. $P(h) > t_x$: For a base rate $P(h)$ above the therapeutic threshold t_x, before knowing the test result acting on the hypothesis h that the disease is present maximizes expected utility. Since Test E is useful, we have $P(h|e) > t_x > P(h|\neg e)$ and the expected utility of E is given by

$$eu(E) = u_{tp}P(h|e)P(e) + u_{tn}P(\neg h|\neg e)P(\neg e) - u_{tp}P(h). \tag{A48}$$

(*Appendix continues*)

Analogous to Case (i), we can rewrite $eu(E)$ as

$$\begin{aligned}
&= u_{tp}P(e|h)P(h) + u_{tn}P(\neg e|\neg h)P(\neg h) - u_{tp}[P(e|h) + P(\neg e|h)]P(h)\\
&= u_{tn}P(\neg e|\neg h)P(\neg h) - u_{tp}P(\neg e|h)P(h)\\
&= u_{tn}P(\neg h)[1 - P(e|\neg h)] - u_{tp}P(h)[1 - P(e|h)]\\
&= u_{tn}P(\neg h) - u_{tn}P(\neg h)P(e|\neg h) - u_{tp}P(h) + u_{tp}P(h)P(e|h) \qquad \text{(A49)}\\
&= u_{tn}P(\neg h) - u_{tn}P(\neg h)[sens(E) - \lambda(E)] - u_{tp}P(h) + u_{tp}P(h)sens(E)\\
&= u_{tn}P(\neg h)\lambda(E) + sens(E)[u_{tp}P(h) - u_{tn}P(\neg h)] + u_{tn}P(\neg h) - u_{tp}P(h)\\
&= u_{tn}P(\neg h)\lambda(E) + [sens(E) - 1][u_{tp}P(h) - u_{tn}P(\neg h)].
\end{aligned}$$

Note that $u_{tp}P(h) - u_{tn}P(\neg h)$ is positive, as $P(h) > t_x$. Thus, the larger the sensitivity of E, the larger the expected utility.

Case 3: To show that the theorem is also true for pairs where one test is useful and the other is not, we will prove that it is not possible to find a pair of Tests E and F with $\lambda(E) = \lambda(F)$ such that Test E is the one with higher specificity or sensitivity respectively, but Test E is useless while Test F is not.

i. $P(h) < t_x$: Suppose we have a pair of diagnostic Tests E and F with $\lambda(E) = \lambda(F)$, $eu(E) = 0$, and $eu(F) > 0$, but $spec(E) > spec(F)$, so that choosing the test with higher specificity would lead to choosing the useless Test E instead of the useful Test F. According to Result 6 we know that, if E has higher specificity than F, then $P(h|e) > P(h|f)$. According to Result 2, Test F is useful iff $P(h|\neg f) < t_x < P(h|f)$. Since $P(h|e) > P(h|f)$, $P(h|e)$ has to lie above t_x too and since E is a diagnostic test, we know that $P(h|\neg e) \leq P(h) \leq P(h|e)$. Therefore, $P(h|\neg e) < P(h) < t_x < P(h|e)$ and thus E is useful which contradicts the assumption that E is useless. Thus, such a pair of tests does not exist and for every pair E and F with $\lambda(E) = \lambda(F)$, $eu(E) = 0$, and $eu(F) > 0$, we know that $spec(E) \leq spec(F)$ and choosing the test with higher specificity entails choosing the test with higher expected utility.

ii. $P(h) > t_x$: Suppose we have a pair of diagnostic Tests E and F with $\lambda(E) = \lambda(F)$, $eu(E) = 0$, and $eu(F) > 0$, but $sens(E) > sens(F)$, so that choosing the test with higher sensitivity would lead to choosing the useless Test E instead of the useful Test F. According to Result 6 we know that, if E has higher sensitivity than F, then $P(h|\neg e) < P(h|\neg f)$. According to Result 2, Test F is useful iff $P(h|\neg f) < t_x < P(h|f)$. Since $P(h|\neg e) < P(h|\neg f)$, $P(h|\neg e)$ has to lie below t_x too and since E is a diagnostic test, we know that $P(h|\neg e) \leq P(h) \leq P(h|e)$. Therefore, $P(h|\neg e) < t_x < P(h) < P(h|e)$ and thus E is useful which contradicts the assumption that E is useless. Thus, such a pair of tests does not exist and for every pair E and F with $\lambda(E) = \lambda(F)$, $eu(E) = 0$, and $eu(F) > 0$, we know that $sens(E) \leq sens(F)$ and choosing the test with higher sensitivity entails choosing the test with higher expected utility. □

Received November 8, 2020
Revision received October 14, 2021
Accepted October 15, 2021 ■