The "Last Mile" Problem in Personalized Medicine: a Dynamic Interactive Graphical Software Solution

Authors: Dr John Fountain[*] and Dr Philip Gunby[+]

[*]Dr John Fountain (corresponding author)
PhD (Stanford) , Senior Lecturer
Economics Department
University of Canterbury
Christchurch, New Zealand

[+]Dr Philip Gunby
PhD (Western Ontario), Senior Lecturer
Economics Department
University of Canterbury
Christchurch, New Zealand

Contact information for corresponding author:
Dr John Fountain
Economics Department
University of Canterbury
Private Bag 4800
Christchurch , New Zealand, 8140
Phone: 643 3642849
Fax: 643 3642635
Email: john.fountain@canterbury.ac.nz ; jflbnz@me.com

Originality Declaration:

I, as corresponding author, promise that I and all persons listed as coauthors on this submitted work have read and understand the "Originality of Manuscripts statement regarding submissions to JAMIA (available at http://jamia.bmj.com/site/about/originalityofmanuscripts.xhtml and confirm that this submission is a new, original work that has not been previously submitted, published in whole or in part, or simultaneously submitted for publication in another journal. Also, in accordance with the aforementioned policy, we have included as part of the submission any previously published materials that overlap in content with this new original manuscript.

Abstract

Clinicians and patients typically experience difficulty with the conditional probability reasoning (Bayes Theorem) required to make inferences about health states on the basis of diagnostic test results. This problem will grow in importance as we move into the era of personalized medicine where an increasing supply of imprecise diagnostic tests meets an increasing demand to use such tests on the part of intelligent but statistically innumerate clinicians and patients. We describe a user friendly, interactive, graphical software interface for calculating, visualizing, and communicating accurate inferences about uncertain health states when diagnostic information (test sensitivity and specificity, and health state prevalence) is relatively imprecise and ambiguous in its application to a specific patient. The software is free, open-source, and runs on all popular PC operating systems (Windows, Mac, Linux)

Acknowledgements
We the authors acknowledge that we have no financial , commercial or other conflicts of interest in the software discussed in this paper. Our source code is public domain and open source. The Wolfram Demonstrations Project is an open-code , public domain resource that uses dynamic computation to illuminate concepts in science, technology, mathematics, art, finance, and… other fields (http://demonstrations.wolfram.com/about.html) and Mathematica Player is owned wholly by Wolfram Research and is freely available for download at Wolfram Research (http://www.wolfram.com/products/player/)

Introduction

The directors of the National Institutes of Health (NIH) and the Food and Drug Administration (FDA) recently spelled out a set of pathways under development to regulate, but also facilitate, an expected explosive growth in personalized medicine[1]. Conspicuously absent from their plans are behavioral methods to improve the way clinicians (counselors, nurses, doctors) and patients, understand and communicate diagnostic information as it applies to personal health problems.

Any complex system will only be as good as its' weakest link. In this case clinicians and patients are the weakest link because they typically[2,3] experience serious difficulties reasoning with diagnostic information, especially when an understanding and calculation of inverse conditional probabilities (Bayes' theorem) is required. We call this the "last mile" problem for personal medicine because it is directly analogous to the last mile problem in modern telecommunication systems: how to deliver enhanced end user services from modern, technologically sophisticated, high volume networks of a providers to older, noisy, low capacity copper wire systems at the end user's place of residence.

The problem will only grow worse as we enter the era of personalized medicine. The fundamental idea behind personalized medicine[1] involves widespread use of diagnostic tests based on newly discovered genes and proteins to better predict individual patients' clinical responses to specific drug therapies. The combination of a rapid increase in the supply of new and untried diagnostic tests and increased demand for the new diagnostic tests from intelligent but statistically innumerate clinicians and patients is a recipe for trouble. Improved methods for calculating, using and communicating health risk information based on diagnostic tests at the level of clinicians and patients are sorely needed.

We have developed one such method — an interactive, visual software tool that:

- eliminates calculation errors in inference tasks based on diagnostic test outcomes;
- facilitates interactive robustness checks (what-if reasoning) on inferences
- builds on the demonstrated[3,4,5,6] ability of manual, text based natural frequency representations of inference tasks to improve comprehension and communication of information about health risks ;

- has a user-friendly graphical interface involving only the manipulation of sliders and menu buttons;
- is based on freely available, open source software capable of running on all widely used operating systems (Linux, Mac, Windows).

Case Description

Just how difficult is this problem of statistical innumeracy, in particular the difficulties clinicians and patients have when calculating and interpreting the results of diagnostic test information for individual patients? Theoretical and empirical answers to this question are an active research area [2-10]. A recent survey[8] of 9 studies in this involving over 600 subjects from a variety of professional and socio-economic backgrounds, including studies on physicians, reports that only around 5-50% of subjects are able to make accurate inferences when basic information relevant to a typical inference problem (sensitivity, specificity, and prevalence) is presented to them. These results confirm earlier research[10] on the statistical innumeracy of physicians showing that an astonishing 70-75% of medical students, house physicians and practicing physicians cannot correctly calculate the inverse probabilities required to generate post-test predictive probabilities.

Natural frequency representations of inference tasks do help[3,4,6,7] these research subject improve their one-shot inductive calculating and reasoning with "given" precise information in questions that would not look out of place in an undergraduate statistics course. But being a static text and manual calculation based format, natural frequency representations inhibit the ability to make repeated and comparative inductive inferences, quickly, transparently, and without error. Because even the best medical evidence applied to an individual case is almost always imprecise or incomplete to some extent (sampling errors, measurement errors, sample selection biases, confounding variables not controlled for or measured), a useful representation of an inference task should facilitate robustness checks and tentative explorations of what-if counterfactual questions to discover the implications of ambiguities and imprecision in estimates for an individual patient. Trying to express, calculate, re-calculate , record and compare text based manual natural frequency representations in a clinical consultation situation with intelligent but statistically innumerate clinicians and patients would be error ridden and confusing – which is why it is seldom, if ever, done[11].

There must be – and is – a better way using modern computer technology.

Method of Implementation

The software tool implementing the dynamic graphical interface we have developed runs on all popular PC operating systems (Windows, Mac, Linux). Original (but more abstract) source code on an earlier version was initially distributed at the open source Wolfram Demonstration Project[12] . The current version of this code with enhancements that improve it's use on a PC and an introductory/instructional webcast audio/video showing the interface in action is available in the public domain at the University of Canterbury's UCTV website[13]. To actually run and interact with software it is necessary to download and install Mathematica Player . Mathematica Player is made freely available for download over the web by the developers of Mathematica[14] in order to permit users to run the thousands of open source programs on the Wolfram Demonstration Project. Absolutely no experience with Mathematica or any programming language is required to use this software. User controls are in the form of sliders and menu buttons set in a user friendly interface with familiar sliders, menu buttons, text input fields. Being open source, users with experience in Mathematica can access and modify the code as they wish.

The proof of the pudding is in the eating so we turn now to some examples of how to use the software to make inferences about health risks on the basis of diagnostic test information.

Examples and Observations

Figure 1 below is a screen shot of the interface. It has four interconnected parts: two sets of user controlled input variables on the left (one a benchmark for comparison purposes) and two outputs on the right, (one tabular, the other graphical). The table in the top right displays the possible combinations of the health conditions/diseases and diagnostic test outcomes as a logician's truth table, augmented by natural frequency information. D is the logical truth value, 1 or 0, of a proposition "a patient has a disease" and T is the logical truth value for the proposition that "a patient has a positive diagnostic test result for the disease". The four

columns of 1's (true) and 0's (false) in the table identify the four logically possible combinations of disease D and test outcome T for the patient, each labeled with their conventional epidemiological and clinical names.

Prior to any diagnostic testing, clinicians and patients are uncertain about values of D and T for their individual cases. There are several ways of quantitatively expressing this uncertainty. The sliders in the left panel that specify numerical inputs for test *sensitivity*, P(T=1|D=1), test *specificity*, P(T=0|D=0), and *pre-test base rate* or prevalence of the disease ,P(D), are one way , using conditional and marginal probabilities. Another logically equivalent way is the frequency information in the bottom rows of the natural frequency table which uses whole numbers in the form of counts of cases in a hypothetical population supposedly "just like" the patient whose health outcomes we are trying to predict. It may seem like only a cosmetic change, but "natural frequency" representations of uncertainty using whole number arithmetic have been shown[2-6] to dramatically improve understanding of and communications about health risks. The top row of frequencies (in black) in the table are derived from the selected input values for the sliders on the top left panel; the bottom row (in grey) corresponds to the benchmark slider values in the bottom left panel.

The graph in Figure 1 displays *positive post-test predictive probabilities* (solid red line, P(D=1|T=1)) and *negative post-test predictive probabilities* (dashed blue line, P(D=1|T=0)) for the patient for every possible pre-test disease prevalence rate P(D=1) from 0 to 1 based on the test-sensitivity and test specificity determined by the slider settings for these variables in the top left panel (arbitrary default values are 80% and 70% respectively). It also shows the specific levels of each predictive probability as conditional probabilities (the labeled squares at 40% and 7% on the respective curves) at the selected pre-test base rate P(D=1) (20% in Figure 1, marked by a vertical black line).

Figure1 here

Figure 2 shows the impact on post-test predictive probabilities of decreasing the base rate from 20% to 5% (using the slider in the top left) when test sensitivity and specificity remain unchanged. As the base rate slider is manipulated, the position of the vertical line as the base rate indicator changes (with an opaque line marking the original base rate at a benchmark

level for comparison purposes) and predictive probabilities and natural frequencies instantly updated. While the predictive probabilities are calculated by the program rather than manually by the user, the interface is not a "black box" calculator. These levels can be checked against and explained by the corresponding natural frequency representations. For example in Figures 1 the positive predictive probability of 0.4 corresponds to 16 true positive cases in (column 1) of the table out of 16+4=20 positive cases (columns 1 and 2) with positive test results.

Figure2 here

The *differences* between positive and negative predictive probabilities at base rates for the disease relevant to the user are also important. Often in a clinician-patient consultation the issue isn't just how to interpret a single *ex post* test result, but whether or not to take the test in the first place. One key factor in that decision is how much can be learned about the chances of having the disease by taking the test. If the difference between positive and negative predictive probabilities (the vertical gap above and below the main diagonal between corresponding curves) is small, then the test will not reduce uncertainty much compared to what already is assessed pre-test.

The initial sensitivity and specificity values used for the calculations in Figures 1 and 2 are low, 80% and 70% respectively. What if the test was more (or less) sensitive or more (or less) specific, or more (or less) on both counts, or perhaps a tradeoff exists between sensitivity and specificity? Figure 3 shows the impact of one of those changes. Test sensitivity and test specificity are now each close to 95%. Notice that the positive predictive probability now increases above 80% for all but the lowest base rates for diseases. Similarly the negative predictive probability decreases to less than 20% for all but the highest prevalence rates. Overall the gap between two post-test predictive probability curves has increased dramatically (in comparison to the lower sensitivity and specificity in the benchmark case, illustrated by the opaque curves in Figure 3 and controlled by the benchmark sliders in the left panel). This reveals that the diagnostic test has more discriminatory power when test sensitivity and specificity are improved, and therefore testing may be more worthwhile performing. Many other combinations of sensitivity and specificity and pre-test base rates can easily be expressed and explored with this software tool.

Figure 3 here

Discussion

We have augmented the standard manual natural frequency representation of inference task problems in three main ways. First, our interface integrates natural frequency representations with standard clinical ways of representing uncertainties about health risks and diagnostic tests (sensitivity, specificity, and base rate). Second, the interface provides visually clear, dynamically updated representations of both inputs to and outputs of an inference task based on medical diagnostics. Third, the many calculations and re calculations necessary when undertaking robustness checks and exploring the implications (for post-test health risks) of imprecision and ambiguities in underlying information sources can be performed and visualized simply, quickly, flexibly and correctly.

As Edward Tufte[15] says: "…clarity and excellence in thinking is very much like the clarity and excellence in the display of data. When principles of design replicate the principles of thought, the act of arranging information becomes an act of insight". Our interface implements good behavioral[3,5] and Bayesian[16,17] statistical principles of inductive inference. It's use in clinical settings and in pre-clinical medical teaching institutions may help to resolve the last mile problem of personalized medicine.

References

1. Hamburg, Margaret A., and Francis S. Collins, (2010) “The Path to Personalized Medicine, New England Medical Journal 363;4 , nejm.org, july 22, 2010

2 Gigerenzer, Gerd, and Ulrich Hoffrage (1995) “How to Improve Bayesian Reasoning Without Instruction:Frequency Formats” *Psychological Review*, 102 (4), 1995, 684–704.

3 Gigerenzer, Gerd, Wolfgang Gaissmaier, Elke Kurz-Milcke, Lisa M. Schwartz, and StevenWoloshin (2008) "Helping Doctors and Patients Make Sense of Health Statistics" *Psychological Science in the Public Interest* 2008 Volume 8 No 2 53-98

4 Gigerenzer , Gerd and Adrian Edwards “Simple Tools For Understanding Risks: From Innumeracy To Insight” British Medical Journal, Vol. 327, No. 7417 (Sep. 27, 2003), pp. 741-744

5 Gigerenzer, G. (2003). *Calculated Risks: How to Know When Numbers Deceive You.* New York: Simon & Schuster.

6 Brase, Gary L (2002) "Which statistical formats facilitate what decisions? The perception and Influence of Different Statistical Information Formats" *Journal of Behavioural Decision Making* Vol 15 381-401

7 Hoffrage, U. and G. Gigerenzer. (1998). “Using Natural Frequencies to Improve Diagnostic Inferences.” *Academic Medicine*. 73:538-540.

8 Barbey, A. and S. Sloman. (2007). “Base-rate Respect: From Ecological Rationality to Dual Processes.” *Behavioral and Brain Sciences.* 30:241-297.

9 Brase, G. (2008). “Frequency Interpretation Of Ambiguous Statistical Information Facilitates Bayesian Reasoning.” *Psychometric Bulletin & Review.* 15(2):284-289.

10 Berwick, D.M., Fineberg, H.v., Weinstein, M.C. “When Doctors meet Numbers” The American Journal of Medicine. Volume 71(6) 1981 pp 991-998

11 Griffiths, Frances , Eileen Green, and Maria Tsouroufli (2005) “The nature of medical evidence and its inherent uncertainty for the clinical consultation: qualitative study” *British Medical Journal* Mar 2005; 330: 511

12 Fountain, John and Philip Gunby. "The Discriminatory Power of Diagnostic Information from Discrete Medical Tests", (2010)The Wolfram Demonstrations Project, stable URL http://demonstrations.wolfram.com/TheDiscriminatoryPowerOfDiagnosticInformationFromDiscreteMed/

13 Fountain, J and P Gunby "The Discriminatory Power of Diagnostic Information from Discrete Medical Tests", (2010), modified source code for use with PCs and video instructional material available at URL:http://uctv.canterbury.ac.nz/post/4/1130.

14 Wolfram Demonstration Project , Player download URL http://www.wolfram.com/products/player/.

15 Tufte, Edward (1997) *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press Cheshier Connecticut

16 Howson, Colin and Peter Urbach. *Scientific Reasoning: the Bayesian Approach* Chicago:Open Court, Second Edition (1993)

17 Lad, F. *Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction*. New York:Wiley-Interscience (1996).

Figure 1

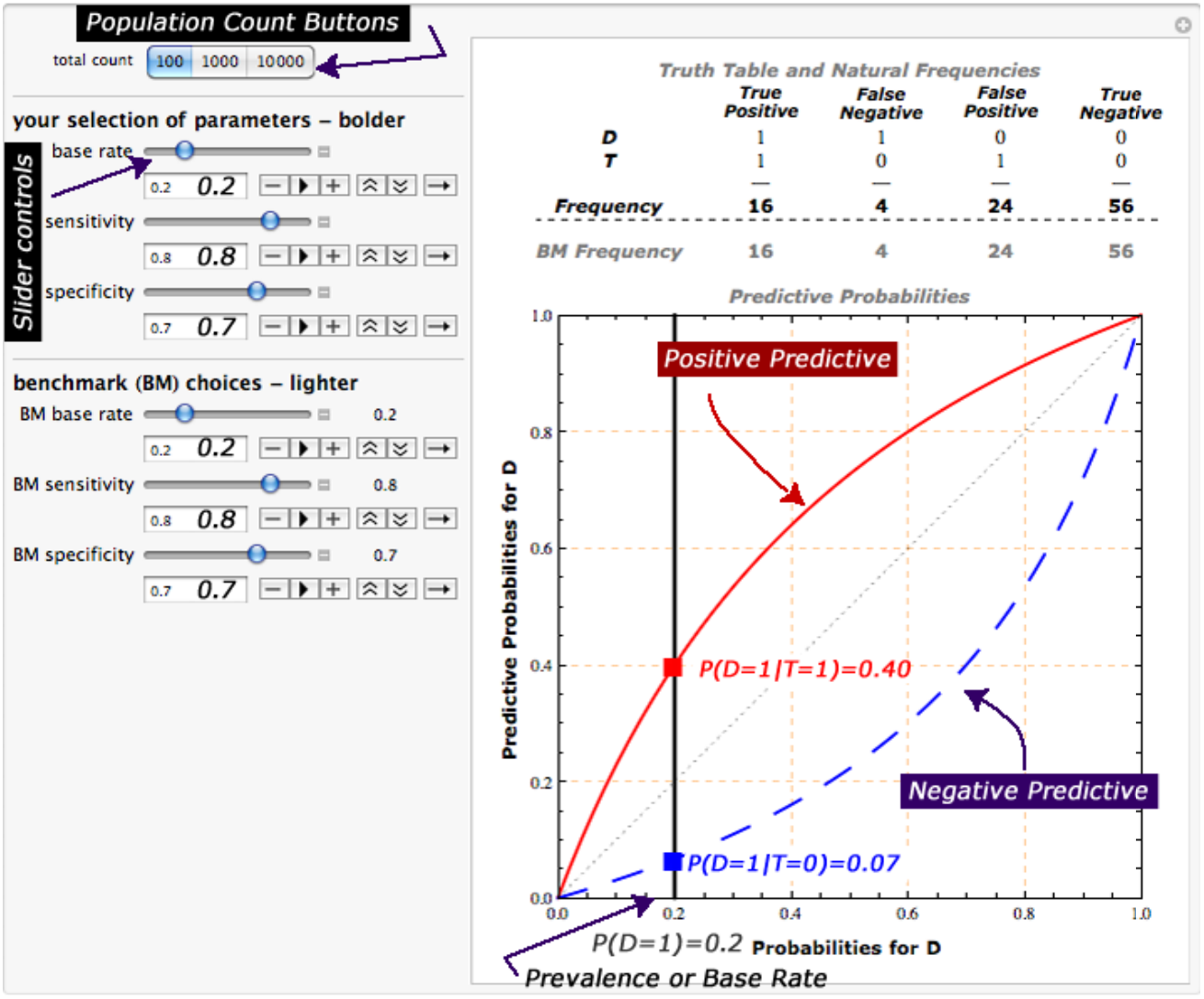

Population Count Buttons
total count
100
1000
10000
your selection of parameters – bolder
Slider controls
base rate
0.2
sensitivity
0.8
specificity
0.7
benchmark (BM) choices – lighter
BM base rate
BM sensitivity
BM specificity
Truth Table and Natural Frequencies
True Positive
False Negative
False Positive
True Negative
D
T
Frequency
16
4
24
56
BM Frequency
Predictive Probabilities
Positive Predictive
Negative Predictive
P(D=1|T=1)=0.40
P(D=1|T=0)=0.07
Predictive Probabilities for D
Probabilities for D
P(D=1)=0.2
Prevalence or Base Rate

Fig 2

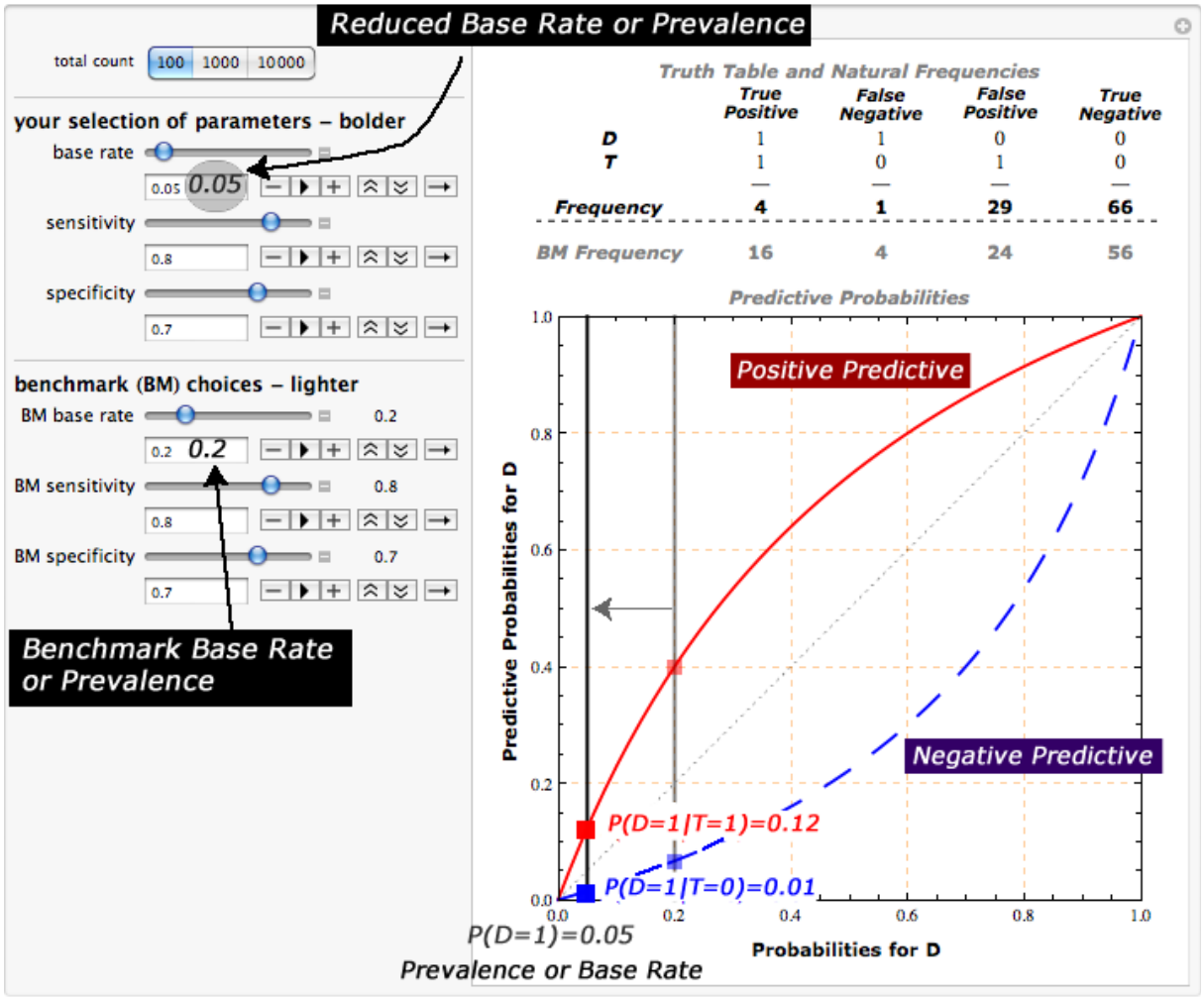
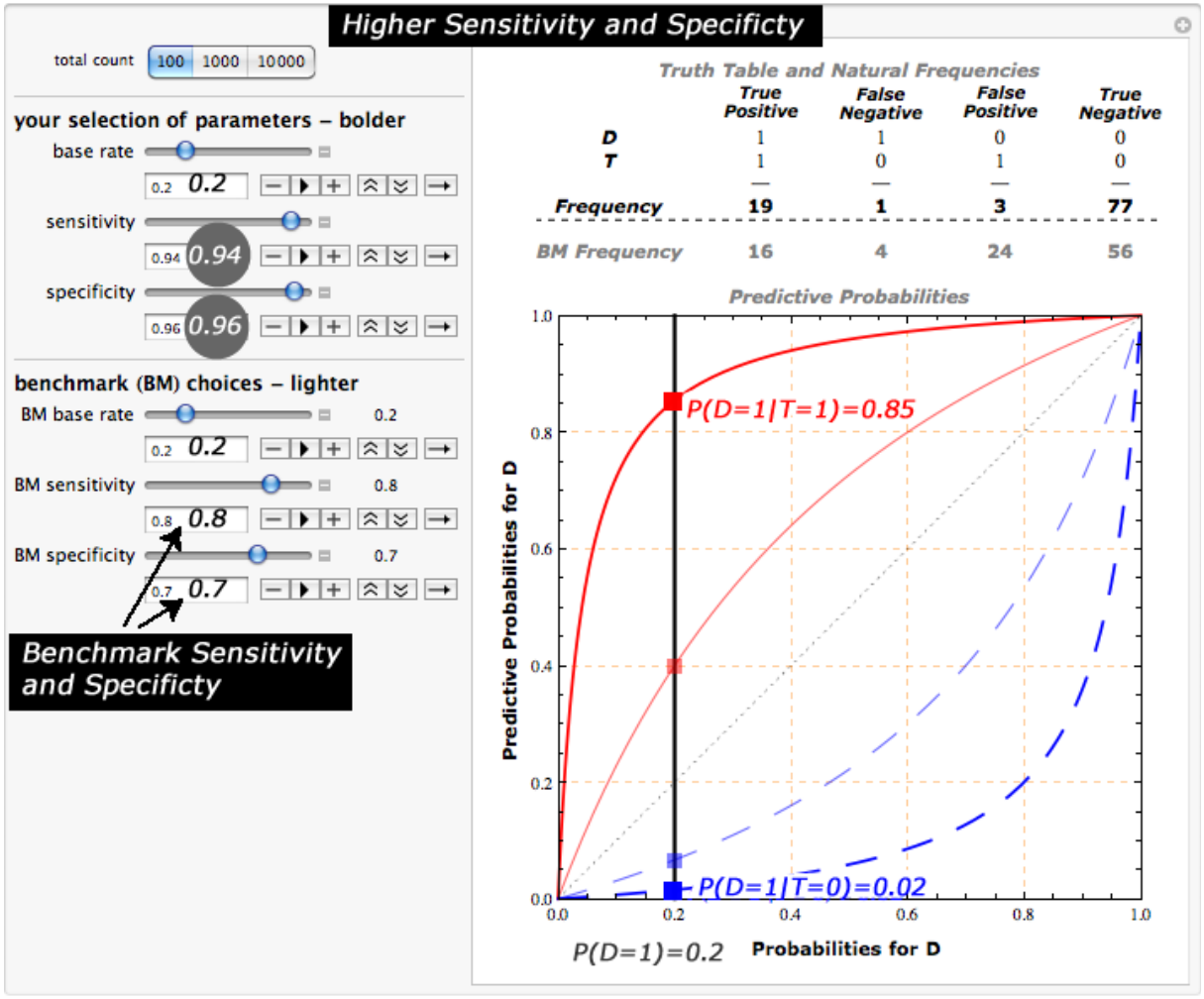

Reduced Base Rate or Prevalence
total count
100
1000
10000
your selection of parameters – bolder
base rate
0.05
0.05
sensitivity
0.8
specificity
0.7
benchmark (BM) choices – lighter
BM base rate
0.2
0.2
0.2
BM sensitivity
0.8
0.8
BM specificity
0.7
0.7
Benchmark Base Rate or Prevalence
Truth Table and Natural Frequencies
True Positive
False Negative
False Positive
True Negative
D
T
Frequency
4
1
29
66
BM Frequency
16
4
24
56
Predictive Probabilities
Positive Predictive
Negative Predictive
P(D=1|T=1)=0.12
P(D=1|T=0)=0.01
Predictive Probabilities for D
Probabilities for D
0.0
0.2
0.4
0.6
0.8
1.0
P(D=1)=0.05
Prevalence or Base Rate

Fig 3


Higher Sensitivity and Specificty
total count
100
1000
10000
your selection of parameters – bolder
base rate
0.2
sensitivity
0.94
specificity
0.96
benchmark (BM) choices – lighter
BM base rate
0.2
BM sensitivity
0.8
BM specificity
0.7
Benchmark Sensitivity and Specificty
Truth Table and Natural Frequencies
True Positive
False Negative
False Positive
True Negative
D
T
Frequency
19
1
3
77
BM Frequency
16
4
24
56
Predictive Probabilities
P(D=1|T=1)=0.85
P(D=1|T=0)=0.02
Predictive Probabilities for D
Probabilities for D
P(D=1)=0.2

Population Count Buttons

100 | 1000 | 10000

Slider controls

your selection of parameters – bolder

base rate: 0.2

sensitivity: 0.8

specificity: 0.7

benchmark (BM) choices – lighter

BM base rate: 0.2

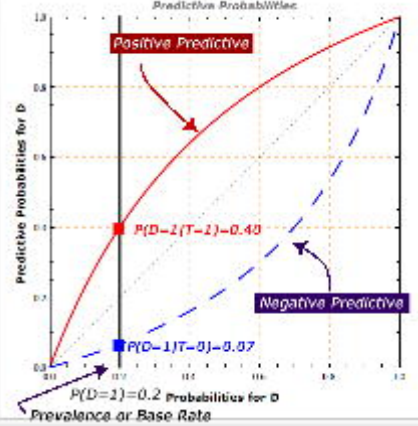
BM sensitivity: 0.8

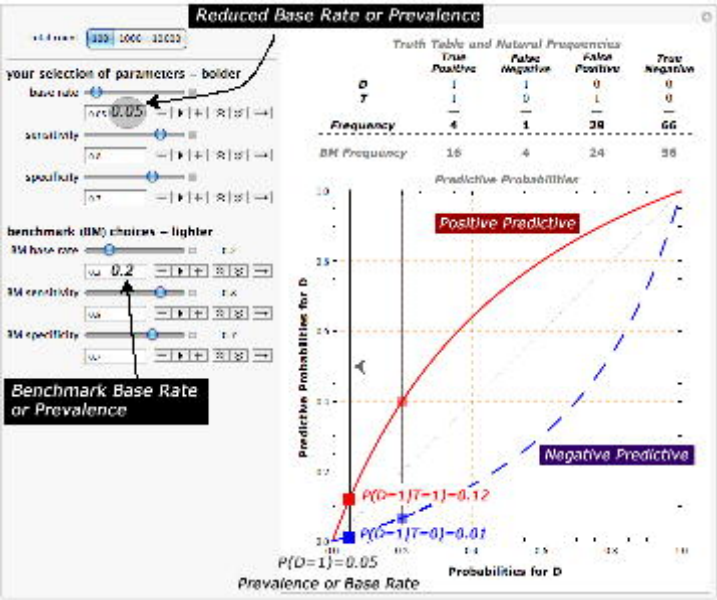
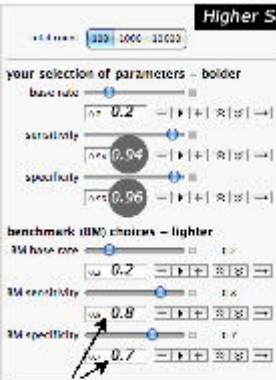
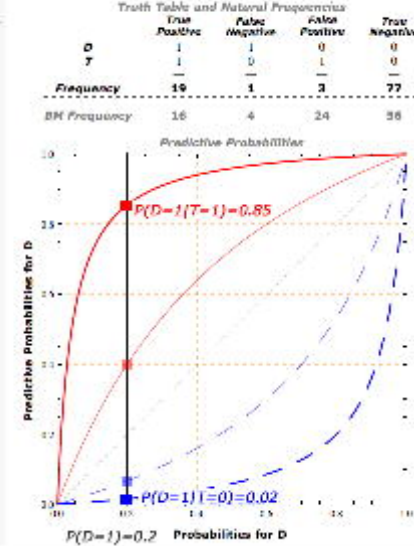
BM specificity: 0.7

Truth Table & Natural Frequencies

Truth Table and Natural Frequencies

	True Positive	False Negative	False Positive	True Negative
D	1	1	0	0
T	1	0	1	0
Frequency	16	4	24	56
BM Frequency	16	4	24	56




Reduced Base Rate or Prevalence
your selection of parameters – bolder
base rate
0.05
sensitivity
specificity
benchmark (BM) choices – lighter
BM base rate
0.2
BM sensitivity
BM specificity
Benchmark Base Rate or Prevalence
Truth Table and Natural Frequencies
True Positive
False Negative
False Positive
True Negative
D
T
Frequency
4
1
38
66
BM Frequency
16
4
24
56
Predictive Probabilities
Positive Predictive
Negative Predictive
P(D=1|T=1)=0.12
P(D=1|T=0)=0.01
P(D=1)=0.05
Prevalence or Base Rate
Predictive Probabilities for D
Probabilities for D

Higher Sensitivity and Specificty