Why clinicians are natural bayesians

Christopher J Gill, Lora Sabin, Christopher H Schmid

Thought you didn't understand bayesian statistics? Read on and find out why doctors are expert in applying the theory, whether they realise it or not

Center for International Health and Development, Department of International Health, Boston University School of Public Health, Boston, MA 02118, USA

Christopher J Gill
assistant professor

Lora Sabin
assistant professor

Biostatistics Research Center, Division of Clinical Care Research, Department of Medicine, Tufts University—New England Medical Center, Boston, MA 02111, USA

Christopher H Schmid
associate professor

Correspondence to:
C J Gill
cgill@bu.edu

BMJ 2005;330:1080–3

Two main approaches are used to draw statistical inferences: frequentist and bayesian. Both are valid, although they differ methodologically and perhaps philosophically. However, the frequentist approach dominates the medical literature and is increasingly applied in clinical settings. This is ironic given that clinicians apply bayesian reasoning in framing and revising differential diagnoses without necessarily undergoing, or requiring, any formal training in bayesian statistics. To justify this assertion, this article will explain how bayesian reasoning is a natural part of clinical decision making, particularly as it pertains to the clinical history and physical examination, and how bayesian approaches are a powerful and intuitive approach to the differential diagnosis.

A sick child in Ethiopia

On a recent trip to southern Ethiopia, my colleagues and I encountered a severely ill child at a rural health clinic. The child's palms, soles, tongue, and conjunctivae were all white from severe anaemia and his spleen was swollen and firm; he was breathing rapidly, had bilateral pulmonary rales, and was semiconscious. It looked like severe malaria. The clinic's health officer evaluated the child using the integrated management of childhood illness algorithm. The algorithm uses cardinal symptoms such as rapid respiratory rate or fever to classify children as having pneumonia or malaria, or possibly both.

In this case, the child's rapid respiratory rate and absence of fever generated a diagnosis of pneumonia with advice to immediately start antibiotics. Our presence was fortuitous. We were able to give the child antimalarial drugs and transport him to the nearest hospital, where blood smear examination confirmed that his blood was teeming with malaria parasites. How did clinical judgments prove superior to the algorithm, a diagnostic tool carefully developed over two decades of research? Was it just a lucky guess?

ROYAL ASIATIC SOCIETY/BAL

Every part of clinical history and examination can be viewed as a diagnostic test

Interpreting diagnostic test results from the bayesian perspective

Clinical diagnosis ultimately rests on the ability to interpret diagnostic test results. But what is a diagnostic test? Clearly blood tests, radiography, biopsies, and other technology based evaluations qualify. However, this view is far too restrictive. In truth, any patient feature that varies in a given disease also qualifies. This definition would include each step in the clinical algorithm above, and, importantly, virtually all elements of the clinical history and physical examination.

Bayesians interpret the test result not as a categorical probability of a false positive but as the degree to which a positive or negative result adjusts the probability of a given disease. In this way, the test acts as an opinion modifier, updating a prior probability of disease to generate a posterior probability. In a sense, the bayesian approach asks, "What is the probability that this patient has the disease, given this test result?" This question proves to be an accurate encapsulation of Bayes's theorem.[1]

Bayes's theorem and its application to clinical diagnosis

Thomas Bayes was an 18th century British vicar and amateur mathematician. Bayes's theorem states that the pre-test odds of a hypothesis being true multiplied by the weight of new evidence (likelihood ratio) generates post-test odds of the hypothesis being true.[2] If used for diagnosis of disease, this refers to the odds of having a certain disease versus not having that disease.

The likelihood ratio summarises the operating characteristics of a diagnostic test as the ratio of patients with the disease to those without disease among those with either a positive or negative test result, and is derived directly from the test's sensitivity and specificity according to the following two formulas:

For a positive test result: likelihood ratio = sensitivity/(1 – specificity)

For a negative test result: likelihood ratio = (1 – sensitivity)/specificity

The following example shows how Bayes's theory of conditional probability is relevant to clinical decision making. The figure shows an electrocardiogram with an abnormal pattern of ST segment and T wave changes. Because the test provides an answer, this process must start with a question, such as, "Is this patient having a heart attack?" The bayesian approach does not yield a categorical yes or no answer but a con-

ditional probability reflecting the context in which the test is applied. This context emerges from what is generally known about heart attacks and electrocardiograms and the characteristics of the patient—for example, "Who is this patient?" "Does this patient have symptoms?" and "What was this patient doing at the time the test was done?" To illustrate this, assume this electrocardiogram was obtained from either of the following two hypothetical patients:

- Patient 1 is an obese 72 year old man with long standing type 2 diabetes, poorly controlled hypertension, 50 years of heavy cigarette smoking, and a family history of early death from cardiac disease (father and two siblings). He came to the emergency room complaining of crushing substernal chest pressure, nausea, difficulty breathing, and a numb pain radiating down his left arm.
- Patient 2 is a 28 year old, 44 kg, non-smoking, vegan woman who competes regularly in triathlons and cares for a sprightly 97 year old grandmother. She is being evaluated in the emergency department with symptoms of dizziness after running 20 km in hot weather but denies any chest discomfort or shortness of breath.

Logically, our opinion of heart attack before seeing the electrocardiogram should have differed greatly between these two patients. Since patient 1 sounds like exactly the kind of person prone to heart attacks, we might estimate his pre-test odds to be high, perhaps 5:1 (prior probability = 83%). If we assume that this electrocardiogram has a 90% sensitivity and 90% specificity for heart attacks,[3] the positive likelihood ratio would be 9 (0.9/(1 – 0.9)) and the negative likelihood ratio 0.11 ((1 – 0.9)/0.9). With this electrocardiogram patient 1's odds of heart attack increase ninefold from 5:1 to 45:1 (posterior probability = 98%). Note, our suspicion of heart attack was so high that even normal electrocardiographic appearances would be insufficient to erase all concern: multiplying the negative likelihood ratio (0.11) by the pre-test odds of 5:1 gives a posterior probability of 0.55:1 (38%).

By contrast, our suspicion of heart attack for patient 2 was very low based on her context, perhaps 1:1000 (prior probability = 0.1%). This electrocardiogram also increased patient 2's odds of heart attack ninefold to reach 9:1000 (posterior probability = 0.89%), leaving the diagnosis still very unlikely.

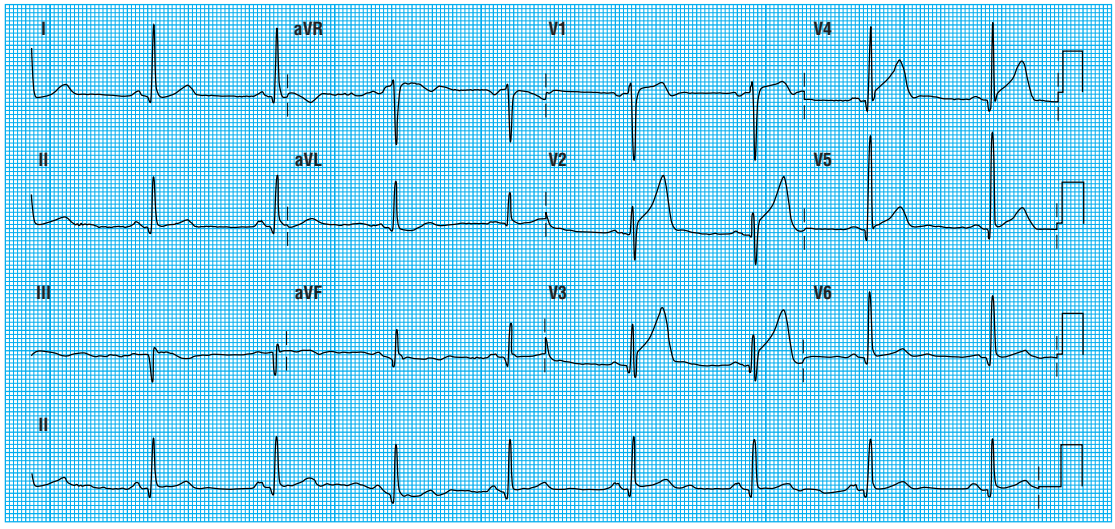
The electrocardiogram modified the prior odds by the same degree in both cases. This does not suggest that both patients would be equally likely to have this electrocardiographic result—in reality they would be unlikely to do so. The purpose of this example was to emphasise the conundrum that often arises in clinical medicine when faced with a truly unexpected test result. Diagnostic tests are mainly used in clinical medicine to answer the bayesian question, "What is the probability that the patient has the disease given an abnormal test?" not, "What is the probability of an abnormal result given that the patient has a disease?" Thus, the response to an unexpected result should be to carefully consider how it modifies the prior probability of that disease, not to second guess your original estimate of that probability.

Conditional probability of tests in series

An important attribute of Bayes's theorem is that the post-test odds for a disease after conducting one test become pre-test odds for the next test, provided that the tests are different not just permutations of the same test.

To extend the previous example, a recent study found the sensitivity and specificity of a stress thallium scan for cardiac ischaemia were 83% and 94% respectively.[4] These figures give a positive likelihood ratio of 13.8 and negative likelihood ratio of 0.18. Thus, if patient 2's post-test odds for myocardial infarction after electrocardiography are 9:1000 and she subsequently had a positive stress thallium test result, her new post-odds would increase to 124:1000 (11%). The odds are still against patient 2 having a heart attack, but given the serious implications of a heart attack, she may now merit further and more accurate invasive testing. Conversely, a negative stress thallium result would decrease her odds from 9:1000 to 1.6:1000 (0.15%), leaving little justification for pursuing the heart attack question further.

This example makes clear that the only truly useless test result is one with a likelihood ratio of 1.0 (sensitivity and specificity both 50%) since multiplying



Electrocardiogram of hypothetical patient

pre-test odds by 1.0 changes nothing. By contrast, a test with 70% sensitivity and 70% specificity is imprecise but not useless since its result still modifies the odds slightly (positive likelihood ratio = 0.7/(1 – 0.7) = 2.3). Arbitrarily, 2.0 and 0.5 have been suggested as the minimally useful values for positive and negative likelihood ratios.[5]

The ability to combine test results in series achieves greater importance once we accept that each question and physical examination during a clinical encounter constitutes a diagnostic test with an attached likelihood ratio. What mainly distinguishes these from formal diagnostic tests (scans, blood tests, biopsies, etc) is that we rarely know what the sensitivities, specificities, or likelihood ratios are for these tests. At best, clinicians carry a general impression about their usefulness and, if quantified, it would not be surprising if most proved to have a positive likelihood ratio of around 2.0 or a negative likelihood ratio of around 0.5—that is, minimally useful.

It is straightforward to measure the operating characteristics of a question or examination, just as with any other diagnostic test. *JAMA*'s rational clinical examination series has measured the accuracy of physical examination for diagnosing breast cancer,[6] digital clubbing,[7] abdominal aneurysms,[8] streptococcal pharyngitis,[9] and others. Not surprisingly, the accuracy of these tests often proved unimpressive. For example, one study determined the accuracy of eliciting "shifting dullness" for identifying ascites.[10] This test operates under the assumption that gas filled intestines should float when surrounded by fluid. Accordingly, when percussing a patient's abdomen in the presence of ascites, areas of dullness and tympany should shift depending on whether the patient is lying supine or on his or her side. Using ultrasonography as the reference standard for ascites, the researchers found that shifting dullness had a sensitivity of 77% and specificity of 72%, leading to an uninspiring positive likelihood ratio of 2.75.

It would be tempting to dismiss shifting dullness as having little use, since a low likelihood ratio is virtually synonymous with a high false positive rate at a population level. However, this reasoning is flawed when applying the test to an individual patient. As with the electrocardiography example, bayesian reasoning demands that shifting dullness be interpreted in some prior context. Here, the context reflects the patient's presenting complaint "My belly has been swelling up lately," and the doctor's knowledge of things that cause bellies to swell. Considering the possibility of ascites, the doctor then refines this context further by conducting the following series of bedside diagnostic tests:

- Has this patient ever abused alcohol (test 1) or injected drugs (test 2)?
- Has the patient ever been jaundiced (test 3)?
- Has the patient ever had a blood transfusion (test 4)?
- Does the patient bruise easily (test 5)?
- Has the patient ever vomited blood (test 6)?
- Is the patient from a country with schistosomiasis (test 7)?
- Does the patient have a history of viral hepatitis (test 8)?
- Is there scleral icterus (test 9)?
- Are there spider angiomata (test 10)?
- Does the patient have a small, firm liver (test 11) or palpable splenomegaly (test 12)?
- Is there lower extremity oedema (test 13)?
- Is there gynaecomastia (test 14)?
- Does the patient's breath have an unusual fishy smell (test 15)?
- And, finally, is there shifting dullness (test 16)?

By now it should be obvious that the decision to test for shifting dullness last or to conclude the bedside evaluation at this stage was arbitrary, and the decision to test for shifting dullness at all emerged as a logical consequence of the doctor's cumulative degree of suspicion to that point. Note, in this hypothetical example, the physician violates the statistical requirement that the tests operate independently, since scleral icterus and jaundice are usually manifestations of the same thing (raised bilirubin concentration). However, this reflects the reality that there is some redundancy in our clinical evaluations.

Dismissing shifting dullness for its low likelihood ratio risks setting clinicians on a slippery slope towards clinical impotence. If we pursued this reasoning to its logical conclusion, many (perhaps most) other questions or examinations might also prove minimally useful. But this conclusion follows only by considering each test in isolation. Instead, suppose we applied the arbitrary minimally useful positive likelihood ratio of 2 to each of the above 16 tests. If all returned positive, the aggregate likelihood ratio could reach 65 356 (2 to the power 16). For comparison, a current generation rapid HIV antibody test carries a sensitivity and specificity of 99.6% and 99.4% respectively.[11] This would be considered an excellent test, but it has a positive likelihood ratio of only 166 (0.996/(1 – 0.994)).

In reality, clinicians don't calculate a running tally of likelihood ratios as they evaluate patients. Rather, they interpret each positive result as "somewhat more suggestive" of the disease and each negative test as "somewhat less suggestive" and conceptualise the pretest and post-test odds in qualitative rather than quantitative terms. Nevertheless, somewhat suggestive to the power x may reach critical mass. This process was what allowed our diagnosis of malaria in the Ethiopian child to be far more than just a lucky guess.

This is not to say that clinical impressions are always better than formal diagnostic tests, particularly as clinical evaluations are rarely as definitive as in this purposefully contrived example. Moreover, single findings can occasionally be very powerful and definitive, just as the results of certain formal diagnostic tests (such as a spinal fluid Gram stain showing bacteria). Nevertheless, the history and physical examination are immensely powerful tools, potentially more powerful than many other formal tests in which clinicians place great faith.

Bayesian reasoning in the pursuit of esoteric diagnoses

The bayesian approach is useful for formulating and revising differential diagnoses, particularly for rare diseases. Consider a patient presenting with fever. Literally thousands of conditions cause fever, many of

them common, others unusual, and some extraordinarily uncommon. The clinical challenge is to prioritise these myriad potential causes of fever and generate a short list of plausible explanations, and to update that list as new information becomes available. This starts with the interview:

Doctor: How long have you had a fever?

Patient: Three days.

Doctor thinks, "Sounds like an acute infection, probably just a cold."

Doctor: Where have you been recently?

Patient: Libreville, in Gabon.

Doctor thinks, "Well now, this might be a tropical infection, perhaps malaria, typhoid, tuberculosis, some kind of parasite . . . or possibly one of those esoteric viruses we learned about in medical school.

Doctor: What did you do there?

Patient: I was part of a compassionate relief team helping rural villagers, many of whom were dying with bleeding gums, high fever, cough, and skin rash.

Doctor thinks, "Hmm, esoteric virus quite plausible."

Doctor: Do you have these symptoms too?

Patient: Yes, my gums bleed when I brush, I have a painful skin rash, and I'm coughing blood (cough, cough)."

Doctor thinks, "Nasty esoteric virus very likely. Need to get this patient isolated and call Centers for Disease Control and Prevention and Department of Homeland Security. Have I just been exposed to Ebola virus?"

After this interview, the now masked and gowned doctor examines the patient and finds a raised temperature, haemorrhages on the patient's conjunctivae, soft palate and finger nail beds, faecal occult blood, a tender swollen liver, and mild jaundice. With each new finding, the probability of nasty esoteric virus increases further despite the fact that none of these tests is remotely specific for infection with haemorrhagic fever virus. However, their poor performance individually does not diminish their importance when combined in a logical sequence. Quite the opposite, since within the span of a few minutes, our doctor has correctly shifted the differential diagnosis from influenza, sinus infection, or possibly pneumonia, to Lassa fever, filo virus infection, or yellow fever without a single blood test, x ray examination, or biopsy and without having more than an educated guess about their associated likelihood ratios. Just as importantly, the doctor's qualitative impression of the odds of nasty esoteric viral infection evolved from "possible" to "very likely," which now dictates what formal diagnostic tests should logically follow to establish the specific diagnosis and how the patient should be managed initially.

Conclusions

We are not arguing that the bayesian approach is a perfect means of reaching a correct diagnosis. Admittedly, the definition of pre-test odds of a disease for a given patient is inherently subjective. But the alternative to subjectivity is to exclude clinical judgment (which is all about context) from patient care. Our goal was to place the clinical evaluation into its appropriate context and to buttress the primacy of history and physical examination in clinical decision making. In so doing, we pay homage to our senior clinical mentors whose probing interviews and painstaking physical examinations so often yielded the truth about their patients' illnesses. Although it is unlikely that they viewed themselves as such, they were bayesians to the core.

Summary points

Clinical decision making is fundamentally bayesian

All clinical history questions and physical examination manoeuvres constitute diagnostic tests, although their sensitivities and specificities are rarely known precisely

Clinicians apply bayesian reasoning in framing and revising differential diagnoses

A Bayesian approach is essential for interpreting surprising test results in the context of history taking and physical examination

We thank Harry Selker and Joni Beshansky for providing the sample electrocardiogram and David Hamer, William MacLeod, and Stanley Sagov for reviewing the manuscript.

Contributors and sources: CJG is a clinical infectious disease specialist who works with an applied research unit conducting clinical trials in developing nations. The impetus for this article emerged after years of clinical practice and after taking a course in applied bayesian statistics, taught by CHS. CJG conceived and primarily wrote the paper. LS helped in writing and rewriting the paper and contributed additional ideas to the manuscript. CHS was CJG's professor of bayesian statistics during his fellowship and helped ensure that the views and arguments presented were in agreement with bayesian theory. CJG is the guarantor.

Funding: CJG was supported by grant NIH/NIAID K23 AI62208 01.

Competing interests: None declared.

1 Berry D. *Statistics, a bayesian perspective.* 1st ed. Belmont, CA: Wadsworth Publishing, 1996.
2 Goodman SN. Toward evidence-based medical statistics. II. The Bayes factor. *Ann Intern Med* 1999;130:1005-13.
3 Selker HP, Griffith JL, D'Agostino RB. A tool for judging coronary care unit admission appropriateness, valid for both real-time and retrospective use. A time-insensitive predictive instrument (TIPI) for acute cardiac ischemia: a multicenter study. *Med Care* 1991;29:610-27.
4 Stolzenberg J, London R. Reliability of stress thallium-201 scanning in the clinical evaluation of coronary artery disease. *Clin Nucl Med* 1979;4:225-8.
5 Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? *JAMA* 1994;271:703-7.
6 Barton MB, Harris R, Fletcher SW. The rational clinical examination. Does this patient have breast cancer? The screening clinical breast examination: should it be done? How? *JAMA* 1999;282:1270-80.
7 Myers KA, Farquhar DR. The rational clinical examination. Does this patient have clubbing? *JAMA* 2001;286:341-7.
8 Lederle FA, Simel DL. The rational clinical examination. Does this patient have abdominal aortic aneurysm? *JAMA* 1999;281:77-82.
9 Ebell MH, Smith MA, Barry HC, Ives K, Carey M. The rational clinical examination. Does this patient have strep throat? *JAMA* 2000;284:2912-8.
10 Cattau EL Jr, Benjamin SB, Knuff TE, Castell DO. The accuracy of the physical examination in the diagnosis of suspected ascites. *JAMA* 1982;247:1164-6.
11 Van den Berk GE, Frissen PH, Regez RM, Rietra PJ. Evaluation of the rapid immunoassay determine HIV 1/2 for detection of antibodies to human immunodeficiency virus types 1 and 2. *J Clin Microbiol* 2003;41:3868-9.

(Accepted 5 February 2005)