Reasoning about uncertainty and learning strategies in medicine **MVST IB 2004–5** (4 Feb 2005)

Rudolf N. Cardinal
Department of Experimental Psychology

Neurobiology & Human Behaviour
Psychology component: self-directed exercise

Background

You may have noticed that there are three elements of the NHB lecture course concerned with psychology — *Mental Illness* (Everitt & Baron-Cohen), *Psychology of Cognition and Memory* (McLaren), and *Neurobiology of Learning, Memory and Cognition* (Ridley). This is a fourth: a “self-directed exercise”. Although this exercise is not assessed, we hope you will find it useful and interesting, and that it helps you a little to function as a better doctor. It is intended to demonstrate the application of psychological issues to medicine, relating to issues in the preclinical and clinical courses and to the day-to-day practice of medicine and surgery, and it provides practice in thinking about medical issues using psychological concepts.

Overview

There are some problems for you to think about, to do with learning and retrieval strategies, that follow from Ian McLaren’s course. I’d also like to convey one particular application of psychology to medicine: the best and clearest way to think about uncertainty, risk, and probabilities in clinical practice, and to convey that information to other people.

1. Learning and retrieval strategies

In Ian McLaren’s first lecture, you learned about encoding, how information is stored, and the factors that influence this process — including the use of imagery, organization, spacing, rehearsal, etc. So we suggest you answer this question and discuss it with your supervisor:

- **Design an optimal strategy for learning this course. Justify the procedures used.**

The second lecture was about retrieval (and, indirectly, forgetting) as well as the control of memory — including direct and indirect access, and the role of context, schemas, and scripts. So we suggest you answer this question and discuss it with your supervisor:

- **Specify the optimal interview strategy for gathering evidence from eyewitnesses at the scene of an accident. Justify your approach.**

2. Reasoning about risk

Importance

A large proportion of doctors reason poorly and make mistakes when it comes to information about risks and probabilities in diagnosis. Mistakes like this can hurt or kill patients. It is quite likely that you will have no further formal teaching on this subject in your preclinical or clinical course. It is very easy to learn how to improve your own reasoning, and how to explain risks to your patients in ways that they too are more likely to understand. They will appreciate this.

People do not deal well with uncertainty or probabilities

Most of the examples that follow are taken from Gigerenzer (2003); he is a psychologist whose research deals with how and why most people are confused by information about risks, why we frequently make mistakes when reasoning with probabilistic information, and what we can do about it.

- *The illusion of certainty.* A woman with few risk factors for HIV infection is screened for it, and tests positive. She is shocked and upset; her colleagues find out and ostracize her; she loses her job. She moves into a halfway house for HIV-infected patients and has unprotected sex with an HIV-infected resident, reasoning that she is already infected. Her relationship with her son suffers (she is a single parent). She falls ill; her physician retests her for HIV; she is found to be HIV-negative. She never had HIV.

HIV tests have **false positives** (disease-free people sometimes test positive) — all tests do. This may be for technical reasons, or laboratory error — in this case, a lab inadvertently exchanged two patients' results. Yet the popular and incorrect image is that diagnostic tests are infallible. Patients frequently think this, and so do many health professionals. In one study, a majority of HIV counsellors stated that false positives do not occur, and half said that if a patient tests positive, it is 100% certain that he is infected with the virus (Gigerenzer *et al.*, 1998). This is simply wrong.

- *Risk communication.* A psychiatrist tells his patients that they have a 30–50% chance of developing a sexual problem, such as impotence or loss of libido, from a selective serotonin reuptake inhibitor (SSRI). Many patients become anxious upon hearing this. It turns out that many of them thought that "a 30–50% chance of developing a sexual problem" meant "something will go wrong in 30–50% of my sexual encounters".

The psychiatrist rephrases his advice, saying that for every 10 people who take the SSRI, 3 to 5 experience a sexual problem. This is mathematically identical, but clearer to the patients. *Frequencies* are clearer than probabilities. The patients were previously confused about the *reference class*: 30–50% of what? People who take the drug, or sexual encounters?

- *Drawing conclusions from numbers.* A 40-year-old woman has a screening mammogram (breast X-ray). It is positive. What is the chance that she has breast cancer? Here are four ways of expressing the same information:

    1. *Probabilities, written in English.* The probability that a 40-year-old woman has breast cancer is about 1%. If she has breast cancer, the probability that she tests positive on a screening mammogram is 90%. If she does not have breast cancer, the probability that she tests positive anyway is 9%. What are the chances that a woman who tests positive has breast cancer?

    2. *Probabilities, written in mathematical notation.*

$$P(\text{has breast cancer}) = 0.01$$
$$P(\text{tests positive} \mid \text{has breast cancer}) = 0.90$$
$$P(\text{tests positive} \mid \text{doesn't have breast cancer}) = 0.09$$

$$\text{What is } P(\text{has breast cancer} \mid \text{tests positive})?$$

    3. *Medical jargon.* In 40-year-old women, screening mammography has a sensitivity of 90% and a specificity of 91%. The prevalence of breast cancer in 40-year-old women is 1%. What is the positive predictive value?

You probably find that it's not easy to work out any of the above in your head, and the jargon doesn't help either (it's explained later, though it should also be familiar from your IA epidemiology lectures). So try this:

4. *Natural frequencies.* Think of 100 women, aged forty. One has breast cancer, and she will probably test positive. Of the 99 who do not have breast cancer, about 9 will also test positive. Thus, a total of 10 women will test positive. How many of those who test positive actually have breast cancer?

Natural frequencies are good, because people reason with natural frequencies much better than with probabilities. Being literate and numerate people, you may underestimate just how poor most people are at handling probabilities. You should know that 40% is the same as 0.4. Yet in one survey, 1000 Germans were asked what “40 percent” means — one quarter, 4 out of 10, or every 40th person? About a third of those asked got this wrong (Gigerenzer, 2003). Also common are mistaken inferences: e.g. thinking that because most heroin addicts have used marijuana, most marijuana users will become heroin addicts.

- *Court cases.* Always fun, these. You are accused of murder and tried. The only piece of evidence is a DNA match between you and a trace found on the victim. The prosecution calls an expert witness who testifies that the probability that this match occurred by chance is 1 in 100,000. The defence insists that this be rephrased: out of every 100,000 people, 1 will show a match. You live in London (census population ~7 million). Which statement do you prefer?

Uncertainty is everywhere in medicine

Uncertainty, and the need to reason with probabilities and other kinds of uncertain information, is prevalent in clinical medicine. Patients want certainty in diagnosis and prognosis, and so do doctors.

- *A common clinical scenario: when should you tolerate diagnostic uncertainty?* A very frail 89-year-old man with substantial ischaemic heart disease who takes aspirin regularly has just had a stroke. On admission, he is also found to be anaemic and his blood tests suggest that this is a result of bleeding (he has a microcytic, hypochromic anaemia, suggesting iron deficiency, with a raised platelet count, suggesting blood loss as the cause). A faecal occult blood test confirms your suspicion that the bleeding is gastrointestinal. Common causes would be bleeding from a stomach ulcer (perhaps exacerbated by his aspirin) or lower gut bleeding, e.g. from colorectal cancer. You are uncertain. Will you investigate further? The usual starting point might be upper gastrointestinal endoscopy (oesophagogastroduodenoscopy or OGD), which involves sedating the patient and inserting a large endoscopic camera down his throat. Follow-up investigations might include a barium enema, and perhaps a colonoscopy, looking for lower gut problems. He may not tolerate these investigations well. And if you find a large ulcer or cancer requiring surgery, will he survive the surgery? Alternatively, you could start a proton pump inhibitor (PPI) to suppress gastric acid production (and check his aspirin dose, though he may well need to stay on low-dose aspirin); PPIs are usually tolerated well. The diagnosis will remain uncertain, although if the signs of bleeding disappear after the PPI is started, you might conclude that it was stomach bleeding after all. Would you tolerate the same uncertainty in a young patient who would be fit for a hemicolectomy if cancer were to be discovered?

- *CT screening: saving life, or exploiting our desire for certainty?* You can now have private whole-body computer tomography (CT) scanning. This is becoming a hotly-debated issue in medicine. The UK company, LifeScan[1], offers a choice of a heart scan (detecting calcification indicative of coronary artery atherosclerosis), a lung scan (for lung cancer), a heart and lung scan, a bone density scan (for osteoporosis), a virtual colonoscopy (for colon cancer), a body scan with virtual colonoscopy, or the whole works — heart, lung, abdomen, and bone density. The web site offers a risk assessment. I filled it out; I’m 29 and think I’m reasonably healthy, but my grandfather died in his 80s of a heart attack, so I have a family history of heart disease. That was the only risk factor I ticked, and without asking my age the web site told me: “You are in an at risk group and may benefit from one of the LifeScan services, in particular a Heart Scan.” It added: “Anybody over the age of 50 although considered at average risk should consider colon screening.” Fortunately, the web site also notes elsewhere that “for people under 40… it will be necessary to obtain a letter from their GP or consultant.” Undoubtedly, whole-body CT scanning can pick up undiagnosed cancers or other diseases — in one series, cancers were found in about 1% of self-referred adults over forty (19 out of 1777 people) — and it has its advocates (see Brant-Zawadzki, 2002a; 2002b). Sometimes these cancers

[1] http://www.lifescanuk.org/

> would be picked up later anyway, some might never have progressed enough during the patient's lifetime to be a problem, and some might be untreatable; at present we don't know if CT screening actually helps people to survive longer. Like any test, CT screening will cause false positives, meaning that healthy people will be investigated further, unnecessarily. Whole-body CT scanning may also cause cancer, because it gives a fair dose of X-rays — an abdominal CT scan typically gives 10 millisieverts (mSv), equivalent to 500 chest X-rays or 3.3 years' worth of background radiation, and it is estimated that of every 2000 people receiving this dose, 1 extra person would get cancer[2]. And it's been estimated that if 2000 forty-five-year-olds each had an annual screening CT for thirty years, about 38 would develop a fatal cancer as a result (Brenner & Elliston, 2004). LifeScan say "our… protocols minimize risk to non-calculable levels", which is patently untrue, and "our low dose total body technique produces $^1/_2$ to $^1/_3$ of the exposure of a conventional CT study, more than 2 million of which are ordered yearly by doctors in the UK for known medical problems", which neatly sidesteps some of the issues. Furthermore, a normal CT scan doesn't guarantee that you haven't got cancer. The US Food and Drug Administration comment as follows[3]: "At this time the FDA knows of no data demonstrating that whole-body CT screening is effective in detecting any particular disease early enough for the disease to be managed, treated, or cured and advantageously spare a person at least some of the detriment associated with serious illness or premature death. Any such presumed benefit of whole-body CT screening is currently uncertain, and such benefit may not be great enough to offset the potential harms such screening could cause."

Doctors reason poorly with probabilities, like most people

Doctors are poor at interpreting probabilistic information. Hoffrage & Gigerenzer (1998) asked 48 physicians this question:

> To facilitate early detection of breast cancer, starting at a particular age, women are encouraged to participate at regular intervals in routine screening, even if they have no obvious symptoms. Imagine that you conduct such breast cancer screening using mammography in a particular region of the country. The following information is available about asymptomatic women aged 40 to 50 in such a region who participate in mammography screening:
>
> *The probability that one of these women has breast cancer is 0.8%. If a woman has breast cancer, the probability is 90% that she will have a positive mammogram. If a woman does* not *have breast cancer, the probability is 7% that she will still have a positive mammogram. Imagine a woman who has a positive mammogram. What is the probability that she actually has breast cancer?*

The doctors — some of them very senior clinicians — thought hard about the problem. Their estimates ranged from 1% (the **base rate,** or **prevalence** of breast cancer) to 90% ("the probability of a positive result given that a woman has breast cancer", which they confused with what they were after, namely "the probability of breast cancer given that a woman has a positive test result"). The most popular answer was 90%; the median answer was 70%; only 10% of the physicians gave the correct answer (which is 9.4%) or something very close to it, and even some of those gave the correct answer only for the wrong reasons (such as confusing "the probability of a positive result given that the patient has *not* got cancer" with "the probability of cancer given a positive result"). Other studies have found physicians to be even worse (e.g. Eddy, 1982). Next, the question was re-expressed in natural frequencies, like this:

[2] http://www.fda.gov/cdrh/ct/risks.html

[3] http://www.fda.gov/cdrh/ct/ and http://www.fda.gov/cdrh/ct/screening.html

> *Eight out of every 1,000 women have breast cancer. Of these 8 women with breast cancer, 7 will have a positive mammogram. Of the remaining 992 women who don't have breast cancer, some 70 will still have a positive mammogram. Imagine a sample of women who have positive mammograms in screening. How many of these women actually have breast cancer?*

This conveys the same information as before (numbers rounded). Now it's easy to work out the right answer; it's $\frac{7}{7+70}=0.09$, or 9%. This time, the majority of the physicians gave the right answer or something very close to it. All the information was available the first time round, the physicians took many minutes over their answers, and they could have converted the probabilities into frequencies if they had thought of this approach — yet few got it right. The problem represents a simple and common clinical situation about which they would be expected to advise patients.

Bayes' rule — correct, but hard to understand and apply

The formal way to solve the probabilistic problem is by using Bayes' Theorem (Bayes, 1763). The point of this section is not to confuse you, but to show you that the mathematically correct way to solve these sorts of problems is quite hard and long-winded! Skip it if you want.

We write $P(A)$ to denote the probability that event A happens. Similarly, we write $P(B)$ to denote the probability that event A happens. We write $P(A \mid B)$ to denote the probability that event A happens, *given that* event B has already happened. Similarly, we write $P(B \mid A)$ to denote the probability that event B happens, *given that* event A has already happened.

Bayes' rule is this:

$$P(B \mid A) = \frac{P(B) \times P(A \mid B)}{P(A)}$$

So let's return to our original breast cancer screening problem (we'll use some slightly different numbers just for variety) and restate it in mathematical terms:

> *"The probability that a 40-year-old woman has breast cancer is about 1%.*
>
> $P(\text{has breast cancer}) = 0.01$
>
> *If she has breast cancer, the probability that she tests positive on a screening mammogram is 90%.*
>
> $P(\text{tests positive} \mid \text{has breast cancer}) = 0.90$
>
> *If she does not have breast cancer, the probability that she tests positive anyway is 9%.*
>
> $P(\text{tests positive} \mid \text{doesn't have breast cancer}) = 0.09$
>
> *What are the chances that a woman who tests positive has breast cancer?"*
>
> What is $P(\text{has breast cancer} \mid \text{tests positive})$?

By Bayes' rule, substituting "breast cancer" for B and "testing positive" for A,

$$\begin{aligned} P(\text{breast cancer} \mid \text{positive}) &= \frac{P(\text{breast cancer}) \times P(\text{positive} \mid \text{breast cancer})}{P(\text{positive})} \\ &= \frac{0.01 \times 0.90}{P(\text{positive})} \end{aligned}$$

So now we need to know P(positive) — the chance of testing positive, regardless of anything else. A woman can test positive if she has breast cancer (a true positive) or if she doesn't (a false positive) — and these are the only two ways to test positive. And we know that the probability of one of our women having breast cancer is 0.01, so the probability of not having breast cancer must be 0.99. So we know that:

$$\begin{aligned} P(\text{positive}) &= P(\text{cancer and positive}) + P(\text{no cancer and positive}) \\ &= P(\text{cancer})P(\text{positive} \mid \text{cancer}) + P(\text{no cancer})P(\text{positive} \mid \text{no cancer}) \\ &= (0.01 \times 0.90) + (0.99 \times 0.09) \\ &= 0.0981 \end{aligned}$$

Now we know that, we can calculate P(breast cancer | positive) from our previous result:

$$\begin{aligned} P(\text{breast cancer} \mid \text{positive}) &= \frac{0.01 \times 0.90}{P(\text{positive})} \\ &= \frac{0.01 \times 0.90}{0.0981} \\ &= 0.0917 \end{aligned}$$

So if a 40-year-old woman tests positive in a screening mammogram, there is about a 9% chance that she has cancer — or, alternatively, of 100 women that test positive, 9 will have cancer and 91 will not.

Natural frequencies are easy to understand and explain

It's clear that while we *could* use Bayes' formula to work out the likelihood that someone testing positive has the disease, it's difficult, and we're likely to make mistakes. We're certainly not likely to be able to do it successfully in our heads. We reason better with *frequencies* or *concrete cases* than abstract probabilities. Probabilities are all "normalized" to be between 0 and 1; this makes them mathematically useful, but hard for us to compare and reason with. Natural frequencies represent the real world more clearly. People will understand your explanations about uncertainty better if you use natural frequencies.

Whenever faced with a problem involving probabilities, try to create a natural frequency representation. You will find the problem easier, and you will explain it more clearly to other people.

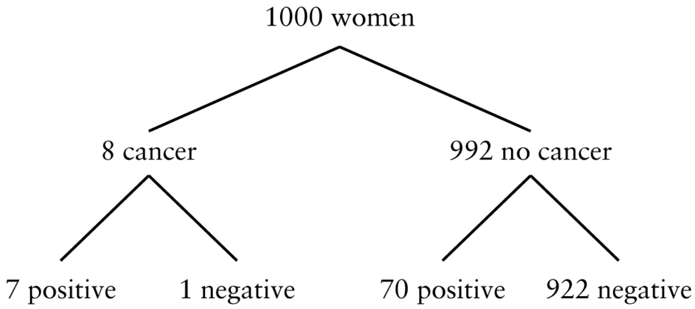
Why call them "natural" frequencies? There are ways of making frequencies "unnatural", by normalizing them like probabilities — and this makes them difficult to work with again. In our breast cancer example, we could use natural frequencies:

Eight out of every 1,000 women have breast cancer. Of these 8 women with breast cancer, 7 will have a positive mammogram. Of the remaining 992 women who don't have breast cancer, some 70 will still have a positive mammogram. Imagine a sample of women who have positive mammograms in screening. How many of these women actually have breast cancer? √

… or unnatural (normalized) frequencies:

Eight out of every 1,000 women have breast cancer. Of every 1,000 women with breast cancer, 900 will have a positive mammogram. Of 1,000 women without breast cancer, 70 will still have a positive mammogram. Imagine a sample of women who have positive mammograms in screening. How many of these women actually have breast cancer? ×

You can see which is easiest to think about, and best to use when explaining to others. An excellent and fast way to sketch any situation like this is with a **frequency tree:**



Why do we reason poorly with probabilities?

In the study by Hoffrage & Gigerenzer (1998), the doctors gave a wide range of answers to the probabilistic question, most of them wrong. It turned out that they had used a wide range of incorrect strategies. One was to confuse *p*(positive | disease) with *p*(disease | positive), but there were many others. Some individuals used two different incorrect strategies when asked two consecutive questions of the same style. Many were aware that they were performing poorly, but they just didn't know the correct way to proceed. In contrast, although the physicians were not perfect when given natural frequency information, they were much better — for one thing, this form of information made them more likely to use information about the prevalence of the disease. So-called *base rate neglect* (ignoring the base rate, or prevalence, of a disease) is common with probabilistic information, and it seems less so when natural frequencies are used. This is not so surprising, because when natural frequencies ("7 positive tests out of 8 with breast cancer out of 1000 women") are converted to probabilities ("*p*[positive | disease] = 7 out of 8 = 0.9"), information about the prevalence is removed. Substantial improvements were seen in some cases when the doctors were taught an explicit *strategy* — the use of frequency trees.

Since this strategy helps doctors to reason correctly, Gigerenzer (2003) also advocates it as a way of communicating risk to patients, which seems eminently sensible. When asked about risks, professional HIV counsellors (including physicians) frequently give numbers that are confusing, incorrect, or impossible (Gigerenzer *et al.*, 1998), so it may be that the enforced clarity of natural frequencies improves the situation. For example, when counselling a low-risk heterosexual man about HIV testing, Gigerenzer suggests the following kind of phrasing:

(Sensitivity?) "The test will be positive for about 998 of 1,000 people infected with HIV. Depending on circumstances, such as the specific tests used, this estimate can vary."

(False positives?) "About 1 in 10,000. False positives can be reduced by repeated testing (ELISA and Western blot), but not completely eliminated. They are caused by certain medical conditions as well as by laboratory errors."

(Prevalence?) "About 1 in 10,000 heterosexual men with low-risk behaviour is infected with HIV."

(Positive predictive value?) "Think of 10,000 low-risk men like you. One is infected and will test positive with practical certainty. Of the 9,999 noninfected men, 1 will also test positive. Thus we expect that out of 2 men who test positive, only 1 has HIV. This is the situation you would be in if you were to test positive; your chance of having the virus would be about 1 in 2."

We have several kinds of innate psychological bias towards certain kinds of information that mean that we can reason perfectly well in some situations, but very poorly in other situations that are logically identical. Some of these are applicable to reasoning in medicine. Such reasoning biases are covered further by the Experimental Psychology Special Option, for those of you that choose to take it.

Interpreting risk information: medical jargon, and how to mislead

Suppose you wish to understand and explain the benefits and risks of mammography screening in asymptomatic women. First, the generalities: women benefit if cancers are detected earlier than they would otherwise be, and those cancers would have gone on to cause the patient harm, and those cancers can be treated successfully, such that morbidity (illness) or mortality (death) can be prevented or delayed. Women may be harmed by false positives (unnecessary worry, investigations, and perhaps surgery), detection and treatment of cancers that would not have progressed within their lifetime, radiation-induced cancers, and false negatives (false reassurance). A very good way of measuring the overall benefit is to conduct randomized controlled trials comparing death rates in women who are screened or not screened. The results of one set of Swedish trials (see Nystrom *et al.*, 2002) are as follows:

	Number of women (aged 40–74)	Deaths from breast cancer	Deaths from breast cancer, per 1000 women[4], rounded
Screened	129,750	511	4
Control	117,260	584	5

Fewer women died from breast cancer in the screened group. How would we express this? Here are several ways of expressing the same information:

- **Natural frequencies — actual proportions in each group.** In the control group, 5 per 1000 women died of breast cancer; in the screened group, 4 per 1000 women died of breast cancer.
- **Absolute risk reduction (ARR).** The chance of dying from breast cancer is reduced by 1 in 1000, or 0.1% (that is, 5 per 1000 minus 4 per 1000).
- **Relative risk (RR),** sometimes called the **probability ratio.** The relative risk of dying if you are screened is 0.8, compared to if you are not screened (that is, 4 per 1000 divided by 5 per 1000). A relative risk of 1 indicates an ineffective treatment; a relative risk of <1 is good, and a relative risk of >1 indicates that the “treatment” is actually bad for you.
- **Relative risk reduction (RRR).** The chance of dying from breast cancer is reduced by 20% (a decrease from 5 to 4 per 1000 is a decrease of 20%).
- **Odds ratio (OR).** The odds of dying in the screened group are 4:996 = $^4/_{996}$. The odds of dying in the control group are 5:995 = $^5/_{995}$. The odds ratio is $^4/_{996} \div {}^5/_{995} = 0.8$.[5]
- **Event-free patients (EFP).** In the control group, 99.5% of women did not die from breast cancer. In the screened group, 99.6% of women did not die from breast cancer.
- **Number needed to treat (NNT).** To save the life of one woman, you would need to screen 1000.
- **Increase in life expectancy.** You can’t calculate the increase in life expectancy directly from the information we have here — but out of interest, women who participate in screening from the age of 50 to 69 increase their life expectancy by an average of 12 days (Salzmann *et al.*, 1997). (It’s easy to see how this sort of figure is arrived at: if 1 in 1000 women is saved by the screening and lives an extra 33 years, this would be an average of 12 days per woman screened.) This can be compared with other common scenarios — for example, this is roughly similar to the effect of driving 300 miles fewer per year (see Gigerenzer, 2003,

[4] An better measure would be deaths from breast cancer per woman–year, not per woman, but the conclusions do not alter substantially here.

[5] The odds of an event A happening are $P(A) \div P(\neg A)$, where $\neg A$ means “not A”. You could also write odds = $P(A) \div [1 - P(A)]$. Given odds, you can also work out probabilities: $P(A)$ = odds / (1 + odds). Bayes’ rule can also be written in terms of odds. The odds *against* something are the reciprocal of the odds *for* something. Odds ratios and relative risks can both be confusing; some more examples and explanations are given in the NST IB statistics booklet at http://pobox.com/~rudolf/psychology (pages 18–20, 22, and 76–77 of the 2004–5 handout).

p. 60). This form of analysis is sometimes used to compare the cost-effectiveness of treatments; for example, such screening costs $21,400 per year of life saved (Salzmann *et al.*, 1997).

It is very easy to mislead people with these figures. If you want to make an effect sound large, quote the relative risk reduction ("screening reduces deaths from breast cancer by 20%"). People will interpret this to mean "for every 100 people screened, the lives of 20 will be saved", but this is not true. A relative risk reduction of 20% is great news if the disease is very common, as many lives will be saved, but not so good if the disease is very rare. Even professional health funding agencies are influenced by the phrasing of funding requests — in one study, health authorities were most likely to fund a treatment if the relative risk reduction was given and least likely to fund when event-free patients or absolute risk reduction were quoted, with NNT in between (Fahey *et al.*, 1995). Conversely, to minimize the apparent risk of a dangerous procedure, you would present absolute rather than relative risks (Pitches *et al.*, 2003)! Actual frequencies, absolute risk reduction, NNT, and increase in life expectancy are all clear ways to express risk.

Interpreting the results of diagnostic tests: more medical jargon

You will come across the terminology of diagnostic tests many times — and the terminology can be confusing! For a given test, people can have the disease or not, and they can test positive or negative for it. We can express the four possible outcomes like this:

	Disease present	*No disease present*
Positive test result	*a* **true positives**	*b* **false positives**
Negative test result	*c* **false negatives**	*d* **true negatives**

Here are four terms in common use:

- **sensitivity:** if a patient has the disease, how likely is the test to be correct (positive)? It's calculated as $\frac{\text{disease and positive}}{\text{disease}} = \frac{a}{a+c}$.
- **specificity:** if a patient doesn't have the disease, how likely is the test to be correct (negative)? It's calculated as $\frac{\text{no disease and negative}}{\text{no disease}} = \frac{d}{b+d}$.
- **false negative rate (FNR):** if a patient has the disease, how likely is the test to be wrong (negative)? It's calculated as $\frac{\text{disease and negative}}{\text{disease}} = \frac{c}{a+c}$. Obviously, FNR = 1 – sensitivity.
- **false positive rate (FPR):** if a patient doesn't have the disease, how likely is the test to be wrong (positive)? It's calculated as $\frac{\text{no disease and positive}}{\text{no disease}} = \frac{b}{b+d}$. Obviously, FPR = 1 – specificity.

These four are properties of the *test*. No matter how common or rare the disease is, a given test has a given sensitivity (and therefore FNR) and specificity (and therefore FPR). For example, we might have an ELISA (enzyme-linked immunosorbent assay) for HIV detection that has a sensitivity of 99.9% and a specificity of 99.9% for` detecting antibodies to HIV, which is pretty good.[6] A Western blot test might have a different sensitivity and specificity.

We also have these:

[6] What accounts for the errors? Well, some proteins detected by HIV tests are also raised by other infectious agents (giving rise to false positives). For example, an ELISA test for HIV may produce false positives if the patient has Lyme disease, syphilis, or systemic lupus erythematosus. Western blots are more specific, but never believe that any test is perfect. HIV infection may have been acquired recently and antibodies haven't had time to develop (a false negative). And, of course, the lab may confuse two patients' blood, or perform the assay wrong, or mix up the results. It is a mistake to think that human errors do not contribute to false positives just as technical errors do!

- **pre-test probability** or **base rate** or **prevalence:** before we do the test, how likely is the patient to have the disease? It's calculated as $\frac{\text{disease}}{\text{all people}} = \frac{a+c}{a+b+c+d}$. (The term "prevalence" applies to diseases; the terms "pre-test probability" or "base rate" apply to all sorts of test situations, not just diseases.)
- **positive predictive value (PPV):** if a patient tests positive, how likely is he to have the disease? It's calculated as $\frac{\text{disease and positive}}{\text{positive}} = \frac{a}{a+b}$.
- **negative predictive value (NPV):** if a patient tests negative, how likely is he *not* to have the disease? It's calculated as $\frac{\text{no disease and negative}}{\text{negative}} = \frac{d}{c+d}$.

The PPV and the NPV depend on the properties of the test, *but also on the prevalence*. Let's illustrate this. First, we test 1,000,000 people with no known risk factors for HIV — blood donors, perhaps — in whom the prevalence of HIV is 0.1%. Assume our test's sensitivity and specificity are both 99.9%. Our table might look like this:

1,000,000 blood donors (prevalence 0.1%)

	Disease	*No disease present*	
Positive test	999	999	*Total 1,998 who test positive*
Negative test	1	998,001	*Total 998,002 who test negative*
	Total 1,000 who have HIV	*Total 999,000 who do not have HIV*	*Grand total 1,000,000*

$$\text{Positive predictive value} = \frac{999}{1998} = 0.5 \text{ or } 50\%$$

$$\text{Negative predictive value} = \frac{998{,}001}{998{,}002} = 0.999999 \text{ or } >99.9\%$$

But suppose that instead, we tested 1,000,000 intravenous drug abusers, in whom the prevalence of HIV is 10%. Our table would look like this:

1,000,000 drug addicts (prevalence 10%)

	Disease	*No disease present*	
Positive test	99,900	900	*Total 100,800 who test positive*
Negative test	100	899,100	*Total 899,200 who test negative*
	Total 100,000 who have HIV	*Total 900,000 who do not have HIV*	*Grand total 1,000,000*

$$\text{Positive predictive value} = \frac{99{,}900}{100{,}800} = 0.991 \text{ or } 99.1\%$$

$$\text{Negative predictive value} = \frac{899{,}100}{899{,}200} = 0.999889 \text{ or } >99.9\%$$

You can see how the PPV has changed dramatically as a result of the change in prevalence. If we give HIV tests to people with no risk factors and they test positive, then there is only a 1 in 2 chance that they actually have HIV. If we give the same test to people with strong risk factors for HIV and *they* test positive, then it is very (99.1%) likely that they do in fact have HIV.

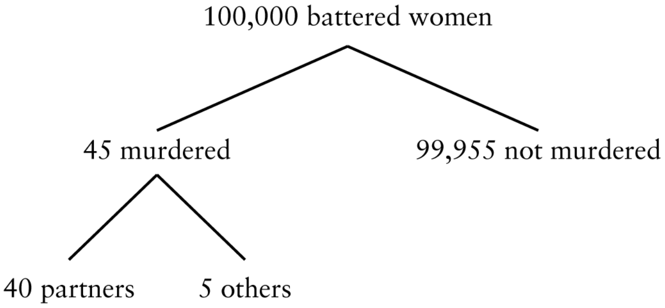
This is all very obvious for groups of people, like our million blood donors and million drug addicts. What happens when we consider just one patient? One way of interpreting this "Bayesian" thinking is that the meaning of a test's results *depends on your prior beliefs about the person*. If you thought that the patient was unlikely to have HIV before you did the test, because he's a blood donor with few risk factors, then you shouldn't be too convinced he has HIV just because he tests positive. If you had good reason to think that the patient had HIV before you ran the test, because he's a drug addict at high risk, then a positive result should leave you pretty certain that he does have HIV. Many people are uncomfortable with this apparently subjective interpretation of test results; nevertheless, it is correct.

Further examples: natural frequencies in the courtroom

Wife battery and murder

No medical relevance here — this last section is just for interest. Courtroom errors can illustrate confused thinking about risk; here are two examples, again from Gigerenzer (2003). In the O. J. Simpson trial, there was evidence that Simpson had been violent to his ex-wife, who had been murdered. The prosecution argued that a history of wife beating reflects a motive to kill: "a slap is a prelude to homicide". The defence countered that battery was *not* a predictor of murder, and should not be admissible as evidence. The defence argued as follows: 2.5–4 million women are battered annually by their partners in the USA, yet only (!) 1432 per year are killed by their partners, so there is less than 1 homicide per 2,500 instances of abuse, or ~40 homicides per 100,000 instances of abuse.

The figures quoted show that *p*(husband murders wife | husband batters wife) ≈ $^{1}/_{2500}$, and the defence argued on this basis. But this is not the relevant probability, as it omits crucial information: *the wife in question had been murdered*. So we are not interested in *p*(husband murders wife | husband batters wife), but in *p*(husband murders wife | husband batters wife *and wife was murdered*). To work this out, we also need to know how many women are murdered by people other than their husbands: it turns out this is about 5 per 100,000 each year in the USA (and we assume this is the same for those who are battered and those who are not). Again, we could play with mathematics, but it'd be hard, so the simplest way to represent this is by a frequency tree, making the numbers concrete:



So if a woman has been battered *and* murdered, the probability that her husband is the murderer is not 1 in 2500 (0.0004) but 40 out of 45 (0.89) — wife battery *is* evidence against the partner of a murdered woman (Good, 1995; 1996). (One could even calculate how much of a difference the information about battery makes, by comparing this probability to the probability that the husband is the murderer if a woman had been murdered but not battered.)

DNA fingerprinting

Here's another area of common misconception. It concerns DNA fingerprinting, in which highly polymorphic "junk coding" regions of human DNA are amplified using the polymerase chain reaction, enzymatically fragmented, and the fragment pattern compared to that from DNA found at a crime scene (DNA fingerprinting is not full-blown sequencing of the sort used by the Human Genome Project). What is the relationship between a DNA match and the guilt of a defendant?

- *Reported match → true match.* When the lab reports that the defendant's DNA matches the crime scene DNA, they may be right (true positive) or wrong (false positive). The false positive rate is probably of the order of 1 in 100 — due to technical errors (e.g. enzyme failures producing misleading DNA banding patterns) or human errors (e.g. contamination or mislabelling of samples) (Koehler, 1997).
- *True match → defendant was the source.* If the defendant's DNA does match the crime scene DNA, that may be because the crime scene DNA *is* the defendant's DNA, or because the crime scene DNA is somebody else's DNA that is very similar to, or identical to, the defendant's (e.g. close relatives).
- *Defendant was the source → defendant was present at the crime scene.* If the crime scene DNA is the defendant's DNA, that may be because the defendant was at the crime scene, or because his DNA was transferred there, deliberately or accidentally, by somebody else.
- *Present at the crime scene → guilt.* If the defendant was at the crime scene, he may be guilty, or he may have been present at the crime scene before / during / after the crime but not have committed it.

The *random match probability* is the probability that a random person matches the crime scene DNA. The *source probability* is the probability that the defendant was the source of the crime scene DNA. The *guilt probability* is the probability that the defendant committed the crime. There are a couple of mistakes commonly made in courtrooms:

- The random match probability *p*(match) is confused with *p*(not source | match). This is called the *source probability error*. For example, if the random match probability is 1 in 100,000, the source probability error is to assume that this is the probability that the defendant is not the source, or, equivalently, that the probability that the defendant is the source is 99,999 out of 100,000.
- The random match probability *p*(match) is confused with *p*(not guilty | match). This is called the *prosecutor's fallacy*. For example, if the random match probability is 1 in 100,000, the prosecutor's fallacy is to assume that the probability of the defendant being innocent is also 1 in 100,000, meaning that the probability of his being guilty is 99,999 out of 100,000.

Here's an exercise illustrating this. The following instructions were given to law professionals as part of a study (see Hoffrage *et al.*, 2000); they are phrased in terms of conditional probabilities. Try to work out the answers. The answers are at the end of this handout.

> *You are trying a rape and murder case. The expert witness testifies that there are about 10 million men who could have been the perpetrator. The probability of a randomly selected man having a DNA profile that is identical to the trace recovered from the crime scene is approximately 0.0001 percent. If a man has this DNA profile, it is practically certain that a DNA analysis shows a match. If a man does not have this DNA profile, current DNA technology leads to a reported match with a probability of only 0.001 percent.*
>
> *A match between the DNA of the defendant and the traces on the victim has been reported.*
>
> 1. *What is the probability that the reported match is a true match, that is, that the person actually has this DNA profile?*
> 2. *What is the probability that the person is the source of the trace?*
> 3. *Please render your verdict for this case: guilty or not guilty?*

Conclusion

Just to repeat:

> **Whenever faced with a problem involving probabilities, try to create a natural frequency representation. You will find the problem easier, and you will explain it more clearly to other people.**

3. Suggested reading: risk and empathy

- Gigerenzer (2003), *Reckoning With Risk: Learning To Live With Uncertainty*. Penguin, £8.99 paperback (ISBN 0-140-29786-3). Excellent and very readable indeed (shortlisted for the 2003 Aventis science book prize), it deals with uncertainty in medical diagnosis (and what to do about it) but also the use of forensic evidence in the courtroom, and how to mislead people by exploiting their innumeracy. [Copies in Experimental Psychology, Churchill, and Heffers, at the least. Previously published in 2002 by Simon & Schuster as *Calculated Risk: How To Know When Numbers Deceive You*.]

- Your ability to empathize with staff and patients will strongly influence your skill as a doctor. Can you imagine difficult situations from the other person's point of view, even after the event? It helps when you taking things too personally; there is much more thoughtlessness than malice, and much more fear than hate. This is empathy: the ability to read other people's minds. If you're not very empathetic, do you know? Simple tactics can improve your empathy skills; try Tucker-Ladd (1996-2000), chapter 13, section 2, or similar (free at http://mentalhelp.net/psyhelp/); some of the rest of the book is fairly good, too.

Acknowledgements

Thanks to Roz McCarthy for creating the exercise, Kate Plaisted for discussion of the content, Trevor Robbins for pointing out Gigerenzer's book, Ian McLaren for supplying the content for the learning/retrieval strategy section, and all for comments on an earlier draft.

References cited in this handout

If you want to look up papers, try starting with PubMed (http://www.pubmed.com/) and type in the author, year, and something of the title into the search box. When you find and click on the paper, it may give you a link to a full-text PDF copy. It's often best to try to retrieve them from a computer within the *.cam.ac.uk* domain, since this gives you access to the University's journal subscriptions.

Bayes, T. (1763). An Essay towards solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London* **53**.

Brant-Zawadzki, M. (2002a). CT screening: why I do it. *AJR American Journal of Roentgenology* **179:** 319-326.

Brant-Zawadzki, M. N. (2002b). Screening CT: rationale. *Radiographics* **22:** 1532-1536; discussion 1536-1539.

Brenner, D. J. & Elliston, C. D. (2004). Estimated radiation risks potentially associated with full-body CT screening. *Radiology* **232:** 735-738.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: problems and opportunities. In *Judgement Under Uncertainty: Heuristics and Biases* (Kahneman, D., Slovic, P. & Tversky, A., eds.), pp. 249-267. Cambridge University Press, Cambridge.

Fahey, T., Griffiths, S. & Peters, T. J. (1995). Evidence based purchasing: understanding results of clinical trials and systematic reviews. *British Medical Journal* **311:** 1056-1059; discussion 1059-1060.

Gigerenzer, G. (2003). *Reckoning with Risk: Learning to Live with Uncertainty*, Penguin, London.

Gigerenzer, G., Hoffrage, U. & Ebert, A. (1998). AIDS counselling for low-risk clients. *AIDS Care* **10:** 197-211.

Good, I. J. (1995). When batterer turns murderer. *Nature* **375:** 541.

Good, I. J. (1996). When batterer becomes murderer. *Nature* **381:** 481.

Hoffrage, U. & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine* **73:** 538-540.

Hoffrage, U., Lindsey, S., Hertwig, R. & Gigerenzer, G. (2000). Medicine. Communicating statistical information. *Science* **290:** 2261-2262.

Koehler, J. J. (1997). One in millions, billions, and trillions: Lessons from *People vs. Collins* (1968) for *People vs. Simpson* (1995). *Journal of Legal Education* **47:** 214-223.

Nystrom, L., Andersson, I., Bjurstam, N., Frisell, J., Nordenskjold, B. & Rutqvist, L. E. (2002). Long-term effects of mammography screening: updated overview of the Swedish randomised trials. *Lancet* **359:** 909-919.

Pitches, D., Burls, A. & Fry-Smith, A. (2003). How to make a silk purse from a sow's ear--a comprehensive review of strategies to optimise data for corrupt managers and incompetent clinicians. *British Medical Journal* **327:** 1436-1439.

Salzmann, P., Kerlikowske, K. & Phillips, K. (1997). Cost-effectiveness of extending screening mammography guidelines to include women 40 to 49 years of age. *Annals of Internal Medicine* **127:** 955-965.

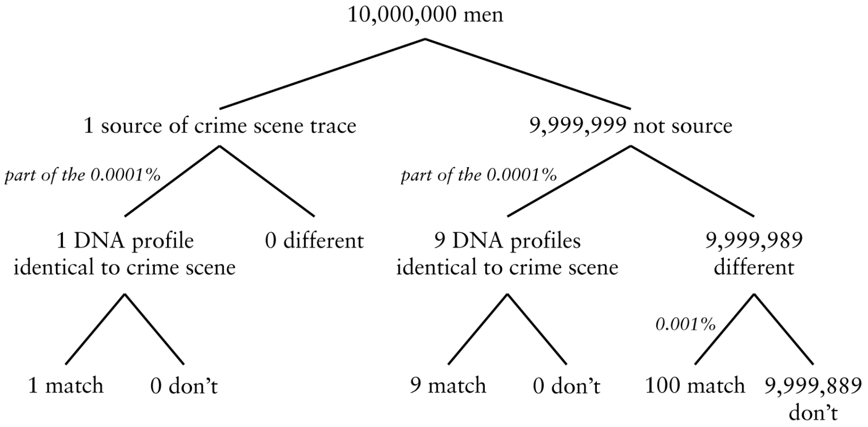
Tucker-Ladd, C. E. (1996-2000). *Psychological Self-Help*, Mental Health Net (online, http://mentalhelp.net/psyhelp/).

Answers to courtroom scenario

Short answers.

1. *p*(true match | reported match) = 0.09
2. *p*(source | reported match) = 0.009
3. Up to you…

Full answer. It is, of course, easiest to re-express everything as natural frequencies in a frequency tree. Here's one way of drawing it:



Now the questions are easy:

1. The probability of a given reported match being a true match is the number of matches where the DNA is identical to the crime scene (1+9) divided by the total number of matches (1+9+100), which is 10 in 110, or 0.09, or 9%.

2. The probability that the person is the source of the trace, given a reported match, is the number of matches where the person is the source (1), divided by the total number of matches (1+9+100), which is 1 in 110, or 0.009, or 0.9%

3. Would you convict on that evidence?

As you'd expect, even professional lawyers did badly on this question, until it was rephrased for them into natural frequencies (see Hoffrage *et al.*, 2000). With the question expressed in probabilities, about 1% of law students and 10% of law professionals got the answers right; when expressed in natural frequencies, 40% of students and 70% of professionals succeeded. Guilty verdicts were more common when the evidence was presented as probabilities than when it was presented as frequencies.