Trials, evidence, and the management of patients with psoriasis

Jonathan L. Rees, FRCP, FRCPE, FMedSci *Edinburgh, United Kingdom*

THE NEED FOR THERAPEUTIC OBSERVATIONS IN MAN

Before any discussion of the value of clinical trials, it is worthwhile remembering why quantitative or numerical studies in man are necessary. That the need for such studies is not self evident is reflected by the fact that it is only relatively recently that numerical based arguments have achieved preeminence over insights based solely on pathophysiologic mechanisms (for reviews see Vandenbroucke[1,2]).

Science is reductionist: science's explanatory approach is to attempt to explain phenomena at one level of description by analysis of those forces acting at a different ("lower") level of description. This epistemic paradigm is phenomenally successful, and despite claims to the contrary, there seems little alternative. It is equally true, but apparently not obvious, that the ability to make precise quantitative predictions from a more basic level to a higher one is often poor. It is therefore the inadequacy of the reductionist enterprise that justifies the need for collecting the results of clinical interventions. A topical example may help.

We may "understand" the details of how ultraviolet radiation (UVR) interferes with cell membranes, how UVR induces changes in gene expression, and how UVR induces particular forms of DNA damage, but our insights do not allow us to make sufficiently precise predictions in the clinical arena: should we use UVB twice or three times a week; what is the risk of melanoma for patients with red hair treated with PUVA; what dosing regimen of UVR should we use for particular diseases? There is an enormous gap, and a gap that cannot be closed, between explanatory pathways of the sort popular in fields such as cell biology and the sort of quantitative predictive models we require in the clinic.

From the Grant Chair of Dermatology, University of Edinburgh.
Reprint requests: Professor Jonathan Rees, Systems Group, Dermatology, University of Edinburgh, First Floor, Lauriston Buildings, Lauriston Place, EH3 9YW, United Kingdom. E-mail: jonathan.rees@ed.ac.uk.
J Am Acad Dermatol 2003;48:135-43.

0190-9622/2003/$30.00 + 0
doi:10.1067/mjd.2003.19

An important corollary of this line of reasoning is easily neglected. Once we accept the limits on our ability to predict with sufficient rigor from a more basic level to a patient, by the same logic the idea of translational research—from bench to clinic, to use the cliché—can only provide a partial view of reality. In many instances a more fruitful approach is in the reverse direction, from clinic to bench.

Ironically, because it is often thought as showing quite the opposite, the rise of human molecular genetics over the past 10 years has provided the best example of clinic to bench. The whole of positional cloning has relied on the definition of a disease at a higher level of explanation guiding the search for the respective "disease gene." The starting point is clinical description and syndrome identification. Only once this is in place can you identify a putative genetic cause. By contrast, if all research were translational in origin we would be left trying to randomly mutate the 50,000 or so genes in the human genome and attempting to delineate particular clinical phenotypes; a task even beyond the National Institutes of Health and the most zealous enthusiasts for knock-out mouse strategies.

The underlying message and rationale is quite simple: just as we can't explain behavior on the basis of an individual synapse, we can't explain much disease or the effects of our interventions on a particular disease without extremely accurate predictive causal pathways, which for most of molecular and cell biology we simply do not have. Clinical medicine is both complex and complicated. Instead, we need to rely on studies in humans under conditions that we can manipulate or at least reason about. Although this latter approach is often described as clinical epidemiology or "trials," it forms the core of what has been the long-standing method of clinical discovery: the comparison of those with a particular state and those without a particular state and an inquiry into either the causal antecedents for the current state or examination of the changes in the status after intervention.

The rest of this article seeks to critically appraise this approach using psoriasis as an example.

THE IDEA OF A TRIAL

Although the modern conception of a clinical trial is often dated to the late 19th and early 20th centuries,[1,2] the logic underpinning clinical trials is not as straightforward as many suppose, nor is it free from internal contradictions.

In everyday clinical practice we wish to know what will happen to a single patient with or without a particular intervention. In rare instances it may be possible to fashion investigative studies around one such individual: side by side comparisons of topical treatments may be one solution, and N of 1 clinical trials remain interesting but neglected. However, the idea of using the patient as his own control is not free from potential error or simplifying assumptions. How do we assess the patient's disease state at the various times? Just because a therapy works once does it work all the time? Can we exclude carry-over effects? And even for the N of 1 design, which approximates in many ways the usual clinical situation, what end points do we choose to measure? How well does our conception of the patient's disease, our measure of disease activity, map to, or explain one person's disease?

Most commonly, we wish to anticipate the effects of interventions for a particular patient by extrapolating from studies of other individuals who have the same disease and who are therefore supposed to approximate the clinical state of the patient we wish to treat. It is widely accepted that in most instances the most robust way to carry out such experiments will involve blinding of the patient and the researcher; randomization so as to reduce the possibility of bias in assignment to whichever agent is used; and, of course, a control intervention.

The use of contemporaneous and randomized controls is a major feature in attempting to improve the robustness of the design ever since the work of R. A. Fisher in the United Kingdom almost a century ago.[3-5] These features of a randomized clinical trial (RCT) cannot, however, always be achieved in practice, nor does the advantage of the class of randomized trials mean that the results of a particular study are more reliable than those from the *class* of observational studies. A well-performed randomized trial on a selectively skewed group of patients cannot be expected to provide a basis for clinical therapy. The RCT provides no probabilistic index of what inference we can make beyond the trial's randomization. RCTs are rarely, if ever, reported with appropriate sampling to allow us to extrapolate the results to the population of patients who did not take part in the study (for an exception see the PACE strategy described by Charlton, Taylor, and Proctor[6]). It follows that there is no formal (probabilistic) reason to imagine that the results of even a well performed RCT provide a more reliable guide to clinical action than other forms of clinical opinion, simply because these alternatives were not directly compared in the study. Or to put it more precisely, the sampling and statistical reasoning we associate with an RCT allow us to make inferences about the population of patients who take part in trials, not the population of patients we see in the clinic. Once we move outside the study population we leave (random) sampling theory behind.

TRIAL METHODOLOGY FOR DERMATOLOGY: THE DOG THAT DIDN'T BARK IN THE NIGHT

The dominant ideology underpinning the design of clinical trials sits uneasily with what would seem the most appropriate clinical questions a dermatologist will want to answer. This merely reflects the dominating influence of trials in areas such as cardiovascular disease or cancer, and the resulting obsession with large trials with simple end points.[6-9] In turn this reflects the fact that in these areas of medicine, treatments often don't work very well, and that these diseases are characterized less well at a clinical level than, say, psoriasis. For instance, many of the large studies reported for cardiovascular disease or cancer relate to discrete and unique end points such as death or survival. In most dermatologic practice such measures are only occasionally important. Death cannot be repeated; exacerbations of psoriasis can.

First of all, imagine a simple question: does treatment A work better than treatment B for psoriasis? We might even choose a relatively discrete event such as clearance after 3 months of therapy, or time to clearance, as end points. But of course what we really want to know is not just the effect of this agent over a time period such as 3 months, but also what effect the agent has on the pattern of psoriasis after cessation of therapy, or what happens when we use the drug again. Yet, studies of disease activity following treatment are, by any sensible definition, ignored in most clinical trials. One can only marvel at the contrast between the weight of data collected during the study period, and that which follows. I am reassured that those interested in infectious disease showed more insight, otherwise prednisolone would remain everybody's favorite for pulmonary tuberculosis; patients feel better the next day.

Secondly, in a disease that is chronic and variable in nature we would like to know about any interaction between efficacy and an individual patient. If a topical agent is known to work on a particular occasion in 6 patients out of 10, one of the critical

factors we should expect from a trial is whether, when at some future date the patient relapses, the probability of success for any patient remains 6/10 or is conditional on the response obtained in the first trial.

Imagine however that the chances of success are even lower, 1/10. You might be tempted, and reasonably so, to argue that the drug is not terribly effective as it only works in 10% of patients. Should we use it at all? But the question, framed in this way is irrelevant: at issue is whether it consistently works in those 10% of patients. Unlike trials in many other areas of medicine, in dermatology we have the opportunity to feed back results from a treatment course to guide further treatment. Trials therefore need to take into account not just variability between individuals but variability in response to a particular agent over time within an individual. Such conditional probabilities may be far more important than those currently reported, and without them current trials remain limited in value.

Thirdly, and almost as important as whether an individual will respond or not, is: when does the evidence allow us to say a treatment is failing and that therefore we should swap to another treatment modality? For instance, consider an individual treated with a topical agent for fairly mild psoriasis as an outpatient. In terms of clearance most patients are unlikely to respond. When should we abandon a particular treatment? How long do we wait before we can make sensible predictions about whether a particular treatment is working? And what is the conditional probability of response with another agent given the failure of the first? Is there a class effect (with topical agents), and when are we better moving on to another treatment modality such as phototherapy? Or consider phototherapy itself. If the TL01 phototherapy seems not to be working in a particular patient how long do we wait before trying PUVA? Is early response predictive of later clearance and to what degree? Alternatively is there a class effect with phototherapy also?

These questions are both straightforward and obvious, but at the same time awkward; awkward simply because although we have become inured to the idea that we should moderate clinical practice in the light of clinical trial evidence, we have yet to appreciate that, as is so often the case in science, asking the right question is harder than collecting the data.

WHAT DO WE MEASURE?

Up to this stage I have glossed over an important series of methodologic problems concerning the relation between clinical trials and clinical practice. I started out by saying that ideally we would like to have evidence of the merits of a particular intervention on the patient in front of us. This we seldom have. We therefore base part of our decision-making on data from studies of individuals with a similar disease, frequently in the form of a study comparing treatment A with treatment B. Often, I suggested, such studies do not provide the information we require. For the present let us ignore these deficiencies, and look in more detail at what trials do provide. I will argue that the closer you look, the less sure you become.

WHAT OUTCOME VARIABLE DO YOU STUDY?

If you wish to compare treatments you have to chose a variable to make the comparison with: you need a measure of clinical state, often seen (mistakenly) as being synonymous with disease activity. The relation of activity or outcome measures to the criteria used in everyday clinical practice is of course not clear. Although this issue was formerly ignored, in recent years there has been a torrent of activity reporting new measures of disease activity, often with accompanying claims that these new scales are more clinically meaningful than those used previously or, and most fatuously of all, that they share a holistic empathy with the patient's condition and psychological state. Sadly, as ever when the word "holistic" is mentioned, hard thinking and rigor have been left well behind: beware existentialism with barcodes.

At a mundane level, the classification of benefit or outcome in a trial may influence greatly what you can say about it. You can support this argument with some calculus, but it is intuitive to most clinicians that such rules must apply. For instance if one was to choose a 100% clearance as a primary end point, the differences between many agents may be reduced considerably. If one was to use a less stringent criterion such as 50% improvement in disease area then the relative merits of the two treatments would appear to be different. Furthermore, if you use scores that are conglomerates of individual aspects of the disease, such as scaling, inflammation, or area, which are known to respond differently to different agents, then the issue may be even more complicated. Therefore the idea that one can summarize the benefits of an outcome in one sort of summary statement is mistaken. Dressing up large numbers of clinically heterogeneous studies using different outcomes, the usual systematic review fodder is not so much an academic exercise, as a real waste of time. For those of us who profess some sympathy with statistical theory and what probabil-

ity space is meant to convey (where probability is a "measure of sets in an abstract space of events"), the pain is correspondingly more acute.

AND THE PASI?

What of the ubiquitous Psoriasis Area and Severity Index (PASI) score, the stalwart of the psoriasis trial? This score of psoriasis severity remains widely used in trials despite its known inadequacies in terms of reproducibility.[10,11] Most trialists just ignore the papers outlining the deficiencies of the PASI and plug in their statistics package just the same. They rely on the fact that clinicians may know what happens to patients, but clinicians frequently don't know what happens to patients' data in the presence of a computer and the absence of a statistical conscience. They are far too trusting.

At this level of criticism, the issues are fairly easily dealt with. Reproducibility could be easily improved with training. Standardized pictures of what the scores mean would help, and individual plaque mapping would all make the scores more useful: nothing difficult here. And there are a number of alternative approaches; side by side comparisons (for topical treatments) based on either plaque scores, or strict blinding with observers forced to "call" which side is better. All these approaches have been used, and have much to commend them. But the real problem is more obstinate, and centers on two issues: how do rating or scoring scales relate to individual patients; and does the concept of a validated rating scale have any meaning, or, as I will argue, does it represent a logical error?

WHAT ARE RATING SCALES FOR?

Those most enamored with clinical trials, evidence-based medicine (EBM), and systematic reviews and the like see clinical trials as a facsimile of everyday practice. This was how it was done in the trial—this is how you should do it in the clinic, or vice versa: all else is wrong (even unethical I hear the whispering . . .). For instance, enthusiasts for pragmatic trials implicitly believe there is a clear calculus linking, say, the PASI score with an individual's clinical state.

There is, however, an alternative view of what a clinical trial is. This latter approach argues that the trial is an artificial construct—like all the rest of scientific experimentation—that allows us to interpret the evidence about whether an agent works. It doesn't claim to mimic the everyday but seeks to produce evidence that guides practice. Trials, it asserts, are no use as facsimiles of nature, but rather allow us to take nature apart. These differences may seem subtle, fodder for academic philosophers only perhaps; they are not. Indeed, they lead to completely different interpretations of clinical trial data.

The most obvious comment is noteworthy by its absence: patients don't come to a clinic demanding a lower PASI score; they want to get better. And physicians don't treat rating scales, they treat patients.[12]

The rating scale in a clinical study is used as a measure of whether a treatment has an effect in the context of a clinical trial, not in the context of a clinic. The language of clinical description such as, "this patient has awful psoriasis: he needs cyclosporine," is not that of science. You can produce all sorts of operational definitions of what a PASI score of 18 means, but there is no equivalent axiomatic definition of bad psoriasis. A woman 2 months before her wedding day may have the same PASI score as she had 10 years before, but she doesn't have the same disease. If the relation between a rating scale and clinical description was fixed you would be obliged to treat her the same on both occasions, but most wouldn't.

The assessment of an individual patient visit is couched in the language of that encounter. The word "bad" does not belong in the same epistemic domain as saying the patient has a systolic pressure of 180 mm Hg or that the PASI is 42. And just as in any spoken language such as English or German, you cannot meaningfully define (as part of an axiomatic system) even simple nouns like table or chair (because they are concepts). So it is in the clinic. Judgements such as "bad" or "moderate" are not part of any closed axiomatic system as they are in mathematics. Writing a macro in Excel to calculate a PASI takes 5 minutes; try writing the code for "bad psoriasis" instead. It can't be done simply because such parts of the consultation lie outside the operational definitions that science is defined by. This, please note, is not the same as saying that these concepts are useless or irrational or that they cannot convey information precisely: Shakespeare managed pretty well on both accounts; it is just that his writings don't fall within the domain of science or mathematics. There is therefore, in philosophical parlance, a noncommensurability between the trial outcome data and the measures used in the clinic. That the respective measures may often show covariance is not in doubt; but neither is it in doubt that if we equate particular scores with terms like "bad" or "moderate" then we soon end up with logically inconsistent statements.

BUT IS THE PATIENT BETTER?

Notwithstanding the criticisms made above, the issue of how to measure clinical state needs to be

pursued further: how do you know whether an individual is in a better clinical state at the end of a study than at the beginning?

One can of course, focus on a particular aspect of the disease as with the PASI. However, the situation is different from that of cancer, for example, where the disease imperative casts a greater shadow over an individual's existence. For psoriasis the issue is not so straightforward, and the solution of just looking at the PASI will not suffice.

We would obviously not want the patient to be worse after treatment than before the intervention. But "worse" is not a scientific term—couched in this way the same issues about epistemic domains arise as I have described in the previous paragraph. We can say the PASI improved but how do we say the patient is better?

There are those who believe that quality-of-life scores or disability indices somehow circumvent this issue. I disagree, and the reasoning has nothing to do with how you weight questionnaire answers or what statistical instruments you employ. Quality of life scores fail simply because we haven't the faintest idea what, in the context of an individual person's life, quality actually means. It is simply not a question that science knows how to approach, nor will it ever, I suspect. It remains firmly entrenched in metaphysics like many other important questions. We have little to add to what the Greek philosophers said over two millennia ago.

Imagine a trial of PUVA versus TL01 phototherapy. PUVA shows a benefit in terms of disease score reduction over that of TL01 phototherapy.[13] This we might agree on without dissent. What, however, remains a problem is how we know whether the patients have benefited overall. Are they better? So, although we know that PUVA therapy in certain circumstances is associated with an increased risk of squamous cell malignancy and possibly melanoma,[14,15] there is no formal calculus that allows us to weigh those disadvantages against the benefits of the treatment. We can all make individual judgments about this, and most of us in the clinic can make some sort of crude suggestion as to what we think the relative merits are for an individual patient (in discussion with that patient) but we have no way to turn this into a summary value for a population, let alone one with confidence limits. Some welfare or health economists would argue that it is possible to reduce these different things using some sort of universal currency but, despite these necessary claims, necessary because there is no justification for health economists otherwise, there is no solution in sight. And once again, this isn't a technical problem, it isn't that we don't have the right computers or questionnaires, but simply that, at least as currently formulated, it is one of those domains of knowledge that science can make no contribution to.

TO WHAT DEGREE DO WE REALLY HAVE RELIABLE INFORMATION ABOUT OUR THERAPIES?

As a rhetorical device, let's ignore the criticisms I have already registered. Let us instead imagine that there is indeed some magical algebraic formula that we can use to describe the trade-offs between, say, having to attend the hospital 2 or 3 times a week for phototherapy, or 5 times a week for outpatient tar, or the long-term risks of renal insufficiency in those treated with cyclosporine. Suppose we had such a formula, do we have the real numbers to plug into it in order to calculate what to do?

One view is what I call the naive trialist view: namely the idea that we have—or soon will have—sufficient information in order to be confident about the side effects and the applicability of summary measures of both treatment effects and side effects to individual patients. By contrast, my own view of reality is quite different.

There are at any one time large numbers of clinically relevant questions that we are unable to answer. However, this is not a transitory phenomenon. We will not know all the answers in another 5 or even 10 years. If we look at the accumulation of our knowledge about phototherapy one might, with some reason and lack of any common sense, be frustrated that at least one large and long-term RCT of phototherapy has not been carried out over the last 30 years. We are never going to have this information because such trials will never be done. Instead, we will always have arguments by analogy, and the problems about weighting particular pathophysiologic mechanisms in terms of trying to define the underlying risks of phototherapy. What might be quite a reasonable summary of our knowledge today might look completely different in 5 years. Imagine for instance that the relation between the risk of PUVA and melanoma shows a peak incubation period of 25 years, our ranking of the various treatment modalities for patients with psoriasis will, at least in the eyes of most clinicians, be turned completely upside down.

Look at the use of cyclosporine. We know that in individuals with organ transplants, use of cyclosporine and other immunosuppressive agents results in a dramatically increased risk of skin and other cancers.[16] How do we now reason about patients who are treated with psoriasis with intermittent cyclosporine? We have no systematic data and no practical way to get it. Will repeated intermittent use result

in significant increases in risk? How will this risk interact with other risk factors for cervical cancer or the putative effects of UVR on lymphoma? We don't know; neither does Novartis.

But let's go a step further. Do our figures for drug side effects vary with time? Or are they genuine biological constants? Few diseases remain constant in incidence. We know, for instance, that skin cancer rates have changed considerably over the last 40 years. How does this influence our certainty about the relation of PUVA or cyclosporine to skin cancer? In reality it would seem likely that the basal rates, dependent as they are on exposure to natural UVR, will modulate the subsequent risk from PUVA or cyclosporine. Our estimates for harm from PUVA are therefore not time neutral, but contingent on UVR exposure over a longer time frame. What may be acceptable in one decade may not be in another.

In reality we just don't have precise answers to many of these questions. In a perfect, and unreal, world we can imagine trials of all the permutations of therapies people might receive during their lifetime. What, however, even in this perfect world, we wouldn't be able to do is look at the permutations of therapies not yet invented. This is not a trivial point. For instance, at present there is a fashion for rotating second line therapy in psoriasis.[17] This may or may not be a good idea but, depending on the biological model used, you could imagine that this is exactly the way you might predict would cause an increase in tumor rates, rather than a diminution of risk to the patient. But for cohorts of individuals who have received these therapies, other therapies that may come on stream in the next few years may interact with those that we have already used, and the answers will remain at best imprecise and at worst unknown.

The take home message is clear: our knowledge will remain incomplete. We may make broad brushstroke comparisons between agents, but we are deluding ourselves if we believe that we are able to pick up risk differences of other than orders of magnitude for psoriasis. Rather than odds ratios reported to the first column on the right hand side of the decimal point, the nearest or second nearest on the left may be all we can achieve. If this is not sufficiently bleak, things can get worse.

INDIVIDUALS VERSUS POPULATIONS

It is a truism to remark that we treat individuals but base our choice of treatments on that obtained from treatments of populations. There is little trivial to this issue. In the assessment of outcomes, summary measures don't refer to individuals.[12] If there is a discrete outcome, individuals either suffer it or they don't. The probability of an adverse effect of 0.1 does not have the same clinical meaning as the fact that it is based on ten individuals, nine of whom did not have it and one who did. More importantly, and obviously, it doesn't equate to a 10% reduction in some clinical status for a particular person.

For a treatment regimen, summary values may show an overall benefit, but of course some individuals may get worse. For instance UVR may make some patients with psoriasis worse. Merely describing a summary measure of what happened to 100 individuals is not the same as describing to an individual that there might be a probability of 0.8 of some improvement but a 0.1 probability that the therapy will make their disease worse. We can't just summarize this evidence with one figure. Indeed, to do so would be misleading.

There are two further points to be made about summary measures (such as odds ratios) and what they mean in the context of individual decision making.

In the majority of instances such measures are of little use in guiding individual decisions unless they are reported in such a way that they can be converted into natural frequencies. For instance, using a ratio and saying that 4 times as many persons experience a side effect with treatment A as with treatment B is less use than saying that for treatment A, 4 suffer it and 96 don't; whereas for treatment B only one suffers it and 99 don't. Individual decision making requires absolute figures: is A worse than B by a factor of 4, or by 3/100 (0.99-0.96/100)? Significant odds ratios are far too easy to obtain, and it is the absolute rates that may influence a patient more, especially when therapy is not mandatory but a matter of choice. Similarly, quoting an odds ratio for clearance of psoriasis is less helpful than quoting absolute rates of clearance with different therapies. It is the latter that can translate into something meaningful for a particular patient.[18,19] The question is what will happen if I follow this course of action, not how do the different courses of treatment compare as part of a statistical hypothesis test.

There is, however, a more pernicious and deep seated problem with the attempt to base individual patient treatment on the results from clinical trials using comparators such as odds ratios: it is simply that such measures and their accompanying confidence intervals do not provide the appropriate measures of variance to guide clinical decisions. The reported confidence intervals reflect not a characteristic of the patients, nor even patient population, but a measure of the precision of the differences between population means. That this is the case can be seen by the fact that merely sampling a larger population (and increasing the denominator n) allows

more precise estimates of treatment effect to be seen when comparing two treatments. However, although this precision lends greater power to the testing of difference between two means, it is not the relevant measure of variance that we require in the clinic. In the clinic we wish to know what variation there is in treatment outcome for a single agent. We wish to answer the question: if you use this treatment what range of outcomes can you expect, what probability distribution of outcome is there? We are not asking for the distribution of differences between two treatments; this is of little consequence to the patient except in an indirect way, as it simply reflects the needs of statistical testing, rather than evidence on which to base therapy.

STATISTICS AND OBJECTIVITY

The interpretation of statistical issues, although not the primary purpose of this article, also needs mentioning. There is a widespread perception, often encouraged by statisticians, that statistical analysis is objective and independent of the views held by the scientist.[3,20,21] Data is conveniently confused with evidence as though the act of interpretation of data as evidence is automatic. Of course Bayesians do not accept this and point out that within the classical statistical paradigm there is a large degree of subjectivity involved in interpreting even simple statistical tests. It is why the idea of EBM is a misnomer—what is usually thought of as EBM is really data-based medicine.

That there are profound differences of opinion about what sort of inferences can be made on the basis of basic concepts as *P* values, confidence intervals, or multiple or repeated testing, is often ignored, or alternatively thought worthy of advanced statistical discussion only.[3,5,20-22] In reality, the differences between the Fisherian view of *P* values, the Neyman-Pearson paradigm of hypothesis testing and, say, approaches based on likelihood or Bayesian thinking are of major and neglected clinical relevance to how we advise individual patients. As we move closer to the individual, such issues place a larger and larger limit on the precision of our numerical predictions. The delusion that there is only one model of statistical inference seems curiously widespread amongst trialists or EBM enthusiasts: all too often clinicians (and journal editors) are unaware that, depending on your choice of statistical theory, you may end up with different *P* values and therefore different conclusions (for examples see[20,21]).

PERSONAL PROBABILITY AND UTILITY

Individuals, of course, may respond quite differently to evidence apparently presented in an identical fashion. There can be little surprise at this—unless one lives in a Stalinist state (or works for a HMO, perhaps). Prior knowledge, prior experience, plus perceptions of risk will differ between people. And data cannot become evidence until these views are taken into consideration.[3,23-25]

Although it is very tempting to glibly describe humans as irrational in terms of risk perception, psychologists are beginning to learn that much human behavior may be far more ecologically sound and rational than the proponents of mathematical logic once thought.[23,24,26-28] So for an individual being considered for PUVA, the given and objective rates of cancer may result in quite different decisions by different individuals. In many, if not the majority of instances, we will not be able to produce empirical information sufficient to overturn an individual's own choice: that is, there are many reasonable decisions that can be defended. An example from my own practice made this clear.

A red-haired woman, heavily freckled, with a medical background, didn't wish to receive phototherapy, but by contrast, wanted to be admitted to hospital for 4 weeks for tar treatment. My original preference (prejudice) was for phototherapy, but her arguments were cogent. She responded to tar in the past and her concerns about skin cancer were already prominent because of her red hair and local public health campaigns on melanoma. The crucial point is that pragmatic-type trials are unlikely to be able to provide much of the data we need.[12] She wasn't concerned with averages but with N of 1, herself. She didn't want other people's utilities for the costs and benefits, she had her own; she didn't want the odds ratio of clearing with tar versus PUVA, as she had clear conditional probabilities that such trials don't consider. And her perception of tumor risk is, by definition, personal, and would no doubt be different if she hadn't been exposed to "public health" campaigns.

When the melanoma rate is quoted as 1:1000, one person gets the melanoma—the others don't. For patients it would be a mistake to subscribe to a strictly frequentist statistical paradigm: they don't get to throw the dice 1000 times, just once. In the example above, once the patient assumes her beliefs, the attempt to produce evidence to render her views, "mistaken" or even untenable, is nigh on impossible.

There is a whole experimental literature on choice, risk perception and psychology. Humans don't respond to risk in a linear way.[23,24,26-28] For instance, the perception of increase in risk from melanoma from 0.39 to 0.4 is viewed differently from that of between 0 and 0.01. Nor should we imagine that people's decisions are necessarily inferior to some of our statistical models, even when

judged by our own statistical criteria. If you are poor, is playing a lottery irrational or rational? The basis for clinical practice therefore has to be not the summary statistic or a notional individual who doesn't exist, but a demonstration of the various choices for which an individual can sign up. The informed decision will differ between individuals; often it may not be possible to argue convincingly against a particular choice (whatever one's own prejudice is) and the possibility of discrete events (such as death or severe side effects) may exert larger effects than might be imagined based on population based studies.

CONCLUSION

In this article I have attempted to address why numerical studies in man are required. Reductionist models from the laboratory will not furnish us with the sort of quantitative information required in the context of the clinic. Human studies are therefore unavoidable, although an increase in our physiological understanding of many of our treatments (such as phototherapy) will also allow more precise and interesting clinical studies.

I have outlined many of the limitations of current approaches in using data gained from human studies as evidence on which to inform therapy. This isn't meant to be dismissive of trials, but merely to point out the need for more hard thinking in a clinical context. Although there is a resurgence of interest in trials and "outcomes" in psoriasis, I see little evidence of serious thought. My own view is that many trials, even case studies, carried out in the distant past, despite their apparent statistical simplicities, provided more useful information than many present day RCTs. Simplicity is, of course, in this context a virtue.

The major reasons for the current state of affairs are twofold. The first is, I believe, the uncritical assumption that trials in cardiovascular disease or cancer should be the model we should follow.[8,9] And uncritical confusion between statistical niceties and scientific ones, all too common in these other fields, has compounded the issue. A statistical hypothesis is not the same as a scientific hypothesis[29,30]; for any trial of phototherapy the idea of a null hypothesis with no effect is absurd. Priors need factoring in to such assessments. And rather than the silliness of meta-analyses for psoriasis one should go back and read Fisher: with reference to statistical tests, significance is defined by the experimenter repeatedly being able to see the same effect with confidence.[3-5,31] We don't need more statistical postmortems: if the effects are real, we will seldom even need statistics. Indeed, when Ken Rothman, the founding editor of *Epidemiology* banished *P* values from its pages, he set us an example to follow[32] (nor should those lovers of confidence intervals ('confidence tricks') be smug—similar confusions arise too).

The second issue to remember is that the empirical test of a trial's usefulness doesn't lie with another trial. Trials are a means to an end. We don't primarily want to know how therapies compare in trials; we want to know how they compare in patients in everyday clinical practice. The test of a trial's conclusions is therefore not further trials, but clinical practice. This is an issue that concerns all of medicine, and at present we have little idea of how to solve, or even frame, the appropriate methodological questions. Indeed many will be alarmed by the problem being viewed in this way at all.

I thank Professors Brian Diffey and Peter Friedman, and Dr Bruce Charlton, for reading an earlier draft of this manuscript.

REFERENCES

1. Vandenbroucke JP. Clinical investigation in the 20th century: the ascendancy of numerical reasoning. Lancet 1998;352(Suppl 2)SII12-6.
2. Vandenbroucke JP. Evidence-based medicine and "medecine d'observation." J Clin Epidemiol 1996;49:1335-8.
3. Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Krüger L. The empire of chance: how probability changed science and everyday life. Cambridge (UK): CUP; 1989.
4. Hacking I. Logic of statistical inference. Cambridge (UK): CUP; 1965.
5. Salsburg D. The lady tasting tea. New York: WH Freeman; 2001.
6. Charlton BG, Taylor PR, Proctor SJ. The PACE (population-adjusted clinical epidemiology) strategy: a new approach to multicentered clinical research. Q J Med 1997;90:147-51.
7. Charlton BG. Mega-trials: methodological issues and clinical implications. J R Coll Physicians Lond 1995;29:96-100.
8. Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. J Clin Epidemiol 1995;48:23-40.
9. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? Stat Med 1984;3:409-22.
10. Tiling-Grosse S, Rees J. Assessment of area of involvement in skin disease: a study using schematic figure outlines. Br J Dermatol 1993;128:69-74.
11. van de Kerkhof PC. On the limitations of the psoriasis area and severity index (PASI). Br J Dermatol 1992;126:205.
12. Rees JL. Two cultures? J Am Acad Dermatol 2002;46:313-6.
13. Gordon PM, Diffey BL, Matthews JNS, Farr PM. A randomized comparison of narrow-band TL-01 phototherapy and PUVA photochemotherapy for psoriasis. J Am Acad Dermatol 1999;41:728-32.
14. Stern RS. The risk of melanoma in association with long-term exposure to PUVA. J Am Acad Dermatol 2001;44:755-61.
15. Stern RS, Liebman EJ, Väkevä L, PUVA FUS. Oral psoralen and ultraviolet-A light (PUVA) treatment of psoriasis and persistent risk of nonmelanoma skin cancer. J Natl Cancer Inst 1998;90:1278-84.
16. Kinlen LJ, Sheil AG, Peto J, Doll R. Collaborative United Kingdom-Australasian study of cancer in patients treated with immunosuppressive drugs. Br Med J 1979;2:1461-6.
17. Koo J. Systemic sequential therapy of psoriasis: a new paradigm for improved therapeutic results. J Am Acad Dermatol 1999;41:S25-S8.

18. Gigerenzer G, Hoffrage U. How to improve Bayesian reasoning without instruction: frequency formats. Psychol Rev 2001;102: 684-704.
19. Hoffrage U, Lindsey S, Hertwig R, Gigerenzer G. Communicating statistical information. Science 2000;290:2261-2.
20. Berger JO, Berry DA. Statistical analysis and the illusion of objectivity. Am Scientist 1998;76:159-65.
21. Royall R. Statistical evidence: a likelihood paradigm. Boca Raton: Chapman & Hall/CRC; 1997.
22. Howson C, Urbach P. Scientific reasoning: the Bayesian approach. Chicago and La Salle (IL): Open Court; 1993.
23. Chase VM, Hertwig R, Gigerenzer G. Visions of rationality. Trends Cognit Sci 1998;2:206-14.
24. Gigerenzer G. Adaptive thinking: rationality in the real world. Oxford: OUP; 2000.
25. Gigerenzer G, Selten R. Bounded rationality: the adaptive toolbox. Cambridge (MA): MIT Press; 2001.
26. Kahneman D, Tversky A, editors. Choices, values, and frames. Cambridge: Published for the Russel Sage Foundation by Cambridge University Press; 2000.
27. Kahneman D. New challenges to the rationality assumption. In: Kahneman D, Tversky A, editors. Choices, values, and frames. Cambridge: Published for the Russel Sage Foundation by Cambridge University Press; 2000. p. 758-74.
28. Kahneman D, Tversky A. On the reality of cognitive illusions. Psychol Rev 1996;103:582-91.
29. Rees J. Science versus statistics. Br Med J 2001;322:1184a.
30. Rees J. Evidence-based medicine: the epistemology that isn't. J Am Acad Dermatol 2000;43:727-9.
31. Hacking I. The emergence of probability: a philosophical study of early ideas about probability, induction and statistical inference. Cambridge (UK): CUP; 1975.
32. Rothman KJ. Writing for epidemiology. Epidemiology 1998;9: 333-7.