How Helpful Are Meta-Analyses in Determining the Best Therapy of Blood Diseases?

Robert Peter Gale^a Hillard M. Lazarus^b

^aImperial College of Medicine, London, UK; ^bUniversity Hospitals Case Medical Center, Case Western Reserve University, Cleveland, Ohio, USA

Key Words

Expert opinion • Meta-analyses • Observational databases • Randomized clinical trials

Abstract

Meta-analyses of data from randomized clinical trials (RCTs) are often used by hematologists to compare the efficacy of therapies of blood diseases. This is especially so when results of RCTs are not decisive. This situation in RCTs arises when the magnitude of differences in treatment outcomes between therapies tested is small, when trials are unpowered to detect differences (these are confounded) and/or when RCTs reach, or seem to reach, contradictory conclusions. Contributing to these limitations of RCTs are the relative rarity of many blood diseases, poor recruitment into RCTs and the greater interest of many hematologists in therapy strategy than in a direct comparison of alternate therapies. These limitations of RCTs are solvable, but only in part, by meta-analyses. Adding data from high-quality observational database studies (ODBs) to meta-analyses is sometimes useful in resolving controversies, but this approach also has limitations: biases may be difficult or impossible to identify and/or to adjust for. However, ODBs have large numbers of diverse subjects receiving diverse therapies and adding these data

to meta-analyses sometimes gives answers more useful to clinicians than meta-analyses of RCTs alone. Side-by-side comparisons suggest analyses from high-quality ODBs often give similar conclusions as meta-analyses of high-quality RCTs. Quantification of expert opinion of high quality is also sometimes useful: experts rarely disagree under precisely defined circumstances and their consensus conclusions are often concordant with results of meta-analyses of high-quality RCTs with and without ODBs. We conclude that meta-analyses are often helpful to determine the best therapy of blood diseases. Accuracy can be improved by including data from high-quality ODBs, when appropriate, and by resolving discordances, if any, with conclusions from high-quality ODBs and from quantification of expert opinion.

Copyright © 2010 S. Karger AG, Basel

Introduction

There are several approaches to determining whether a new therapy is more effective than a current therapy of a blood disease including analyses of data from experimental and observational studies and quantitation of expert opinion. Data from large randomized controlled trials (RCTs) are usually (and rightly) considered the gold

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2010 S. Karger AG, Basel
0001-5792/11/1252-0091\$38.00/0

Accessible online at:
www.karger.com/aha

Robert Peter Gale, MD, PhD, DSc (hon), FACP
11693 San Vicente Boulevard
Suite 335
Los Angeles, CA 90049 (USA)
Tel. +1 310 442 9010, E-Mail robertpetergale@aol.com

standard for comparing therapy outcomes. However, large RCTs with adequate follow-up are rarely reported in blood diseases like leukemia and bone marrow failure syndromes. Why? There are many contributing factors, some of which are unavoidable: (1) These diseases are relatively rare. For example, lifetime risk of developing a leukemia is only 1.4%. (2) Because of their dire prognosis, investigators and potential subjects are often unwilling to participate in RCTs arguing a 'control' cohort is unnecessary because their outcome either known or dreadful. (3) We often think we know the correct answer to results of a therapy comparison a priori and consider RCTs 'unethical'. (4) Often we are more interested in a therapy strategy (ordering of therapies). This requires an intent-to-treat analysis under circumstances where many subjects will not receive the prescribed therapy (therapies). None of these arguments is convincing. For example, investigators' pre-study notion of outcomes (including ours) are repeatedly shown to be wrong (radical mastectomy for breast cancer comes to mind). Nevertheless, these limitations and prejudices hamper our ability to conduct and analyze results of RCTs. Some of these problems, e.g. small sample sizes, can be addressed by meta-analyses – others cannot. But when there are other approaches, like observational database studies (OBDs) and quantifying expert opinion, they can sometimes be compared with results of meta-analyses to increase the likelihood of getting a correct answer. Here, we consider the benefits and limitations of meta-analyses in comparing therapy options in blood diseases.

Randomized Clinical Trials

Several types of experimental studies are used to compare therapies. Most highly regarded are RCTs in which subjects are randomly assigned to receive one or more alternative therapies. The critical advantage of RCTs over other experimental approaches is that randomization ensures comparability of subjects in the therapy arms being studied. This contrasts with nonrandomized studies in which regression techniques are often used to adjust for known prognostic factors. Large randomized trials provide balance for known and unknown prognostic factors and eliminate selection biases whereby some subjects are more likely or less to get certain treatments than others. Another advantage of RCTs is that the data are collected prospectively providing high-quality, well-defined assessment of outcomes as a basis for comparison. Because of these features, RCTs that identify a ben-

efit for a new therapy at a statistically significant level are credible to clinicians. Because there is little debate regarding the advantages of large, properly performed RCTs in comparing alternate therapies, our focus is on limitations.

Results of RCTs are most convincing when they have large numbers of subjects. This is rarely so in RCTs of blood diseases. For example, a study with a 90% power to detect a 10% difference in outcome of two therapies in a blood disease like aplastic anemia requires about 1,000 subjects. No RCT of even 300 subjects is reported.

Limitations of RCTs

Another limitation of RCTs is the small number of subjects treated at any center. Because blood diseases are typically rare (thalassemia and sickle cell disease are exceptions) most centers treat relatively few subjects compared to trial size. Consequently, many centers are typically involved in RCTs. The result may be a consensus protocol design that may engender little enthusiasm or may be less relevant to what physicians really want to know (or both). Disease rarity also contributes to the slow rate of accrual to RCTs of blood diseases. These trials often take 5–10 years from inception to publication. Consequently, the therapies being compared are sometimes not of current interest. Proponents of each therapy being compared often discount results of the late-published RCT claiming techniques and technologies have changed and the conclusions of the trials no longer relevant to current practice ('we have moved on'). This problem of delayed relevance is compounded by the need for long follow-up in some studies.

Clinical trialists are familiar with obstacles to organizing multicenter RCTs of blood diseases. Although centers may agree, for example, to compare a conventional to a new therapy they may not agree on details of the new therapy. These disagreements typically lead to one of two solutions. One is to use a consensus regimen. This results in uniformity in trial conduct but limits generalizability of the conclusion. Moreover, many agree with the architecture critic Meyer Rus who said: 'Consensus is the fastest route to mediocrity.' (To which Frank Lloyd Wright replied: 'Doctors bury their mistakes, I can only advise my clients to grow vines.') This notion limits willingness of investigators and potential subjects to participate in an RCT of a consensus approach. The second solution is to allow each center to do a new therapy however they want. This makes conclusions more generalizable but introduces variability uncontrolled by randomization.

Although there is considerable advantage of having comparable subjects in therapy arms in large RCTs, there are several limitations to extrapolating conclusions from these trials to clinical practice. For example, most RCTs precisely specify study entry criteria and therapy parameters. Although precise study entry criteria ensure comparability of subjects in the therapy arms, this limits applicability of the study conclusions. Subjects on RCTs are often different than typical persons with the blood disease being considered (e.g. younger or fitter). It is common for entrants into RCTs to represent <10% of the universe of potential therapy candidates [1]. It may not be appropriate to assume the benefit of treatment translates to subjects with features different than the study entrants. For example, conclusions of a trial of subjects 18–50 years should not be assumed applicable to younger or older persons.

Sometimes RCTs report a benefit of one therapy over another to a cohort of subjects. However, one or more pre-defined subcohorts within the benefitting cohort may not share this benefit or may even have a worse outcome. Put differently, the benefit of the new therapy may not be uniformly distributed amongst subjects in a cohort of an RCT: some subjects may benefit while others may be harmed [2]. Thus, although results of RCTs inform the clinician how to treat a cohort of subjects, these data may not translate to persons for whom the clinician must decide what therapy to recommend [3]. Moreover, it is sometimes possible to predict who may benefit or be harmed before randomization based on subject characteristics. This indicates an interaction with treatment, in which subjects with different characteristics respond differently to therapy. These interactions may be detected when large ODBs are used; RCTs are rarely powered to detect interactions. These limitations of generalizability apply even more to meta-analyses of RCTs [4]. As an example, although persons with acute myeloid leukemia in first remission with adverse cytogenetics [e.g. del(5/5q), del(7/7q)] may benefit from an allotransplant, this may not be so for a person with favorable cytogenetics [t(15;17), inv(16)]. Most RCTs include subjects from both cytogenetic cohorts; it would be difficult to design an RCT of only one cohort because of sample-size considerations. However, we can predict a priori (as opposed to a posteriori) that whilst the whole allotransplant cohort might have better freedom from relapse than the chemotherapy cohort, this may not translate to a survival advantage for the favorable cytogenetic cohort.

Because of the substantial theoretical advantages of RCTs over alternate approaches, RCTs are regarded as the

gold standard for determining whether a new therapy is better than a current therapy. The question is: how well do RCTs perform in real-world settings? Considerable data indicate that RCTs testing seemingly similar clinical questions sometimes produce discordant results [5]. How would a clinician know which conclusion is correct (if either)? The finding of disparate outcomes of highly cited, initially favorable RCTs is not uncommon. Below we discuss whether meta-analyses are useful in resolving seemingly disparate results of RCTs [6].

Because randomization provides an unbiased assessment of the treatment effect for that specific study, the most likely reason for RCTs with disparate outcomes is that the trials are not really comparable. This lack of comparability is often expected to result in different outcomes of seemingly similar RCTs. Careful consideration of the possible impact of this variability is needed to determine whether it is reasonable to combine these trials in a meta-analysis.

Another potential reason for disparate outcomes of RCTs relates to blinding or randomization concealment. Failure to use adequately concealed random allocation can distort the apparent effects of a treatment in either direction, causing the effects to seem either larger or smaller than they really are. The size of these distortions can be as large as or larger than the size of the effects that are to be detected [7]. Use of biological assignment rather than randomization, as is sometimes done in RCTs of blood diseases, also raises potential for enrollment bias to impact study results [8].

As we discussed above, most hematologists want a generic rather than specific answer to the question whether a new therapy of a blood disease is better than conventional therapy. For example, they want to know whether older persons with myelodysplastic syndrome should receive intensive chemotherapy, a hypomethylating drug or supportive care only. They are less interested in the precise composition of the intensive chemotherapy regimen (daunorubicin, idarubicin or mitoxantrone at what dose and schedule), the precise composition of the hypomethylating regimen (5-azacytidine or decitabine at what dose and schedule) or exactly what constitutes supportive care (platelet transfusions at what blood platelet level). They are unlikely to get a generic answer from a compilation of RCTs since each trial often tests different subjects under different conditions. This limitation can be overcome if the meta-analysis shows no significant interaction between outcome and treatment. It is also important to consider that we rarely have more than a few large RCTs to evaluate in blood diseases. Were there data from only one

RCT, substantial likelihood results would be inconclusive or incorrect for a small-magnitude therapy effect. An example is the results of autotransplants in breast cancer where results of a small RCT (later shown to be spurious) showed a benefit whereas results of larger RCTs and a meta-analysis showed no benefit [9–12].

Contradictory Results and Confirmation Bias of RCT

How do people deal with disparate information like contradictory results of RCTs? This issue is intimately related to the concept of confirmation bias, a tendency to search for and/or interpret new data in a way that confirms one's preconceptions and avoids data and interpretations which contradict prior beliefs. Confirmation bias is extensively studied by psychologists. In a classic experiment, students from Dartmouth and Princeton universities were shown clippings of the contentious 1951 Dartmouth-Princeton football game and asked to note every instance of cheating. Students from each university were convinced the other side cheated more [13]. Another study involved students who favored or opposed capital punishment. The students were shown two studies: one suggesting executions reduce subsequent murders and the other doubted that conclusion [14]. Whatever their stance, the students judged the study supporting their position well conducted and persuasive and the other study profoundly flawed. Although a balanced analysis of the studies suggests no strong conclusions should be drawn, the opposite happened: students accepted evidence conforming to their original view whilst rejecting contrary evidence. This concept is referred to as attitude polarization. These data suggest that clinicians reviewing data from contradictory RCTs are more likely to believe those confirming their prejudices and reject contradictory data. Experience suggests this is, unfortunately, so. Many readers will be aware of Murphy's Law of Research: enough research will tend to support your theory. Francis Bacon enumerated this concept in 1620: 'The human understanding when it has once adopted an opinion... draws all things else to support and agree with it. And though there be a greater number and weight of instances to be found on the other side, yet these it either neglects and despises... in order that by this great and pernicious pre-determination the authority of its former conclusions may remain inviolate' [15].

Why do clinicians fall prey to confirmation bias? Recent data suggest a physiological basis. For example, a study in 30 men was done just before the 2004 US presidential election. Fifteen described themselves as strong Republicans and 15 as strong Democrats [16]. Subjects

were asked to assess contradictory statements by George W. Bush and John Kerry whilst having a functional magnetic resonance imaging scan. The scans showed that the part of the brain associated with reasoning (dorsolateral prefrontal cortex) was not involved when assessing the candidates' statements whereas the most active regions of the brain were those involved in processing emotions (orbitofrontal cortex), conflict resolution (anterior cingulate cortex) and making judgments about moral accountability (posterior cingulate cortex).

A related but distinct issue is the impact of the setting in which individuals make decisions on their choices [17]. For example, how much people eat depends greatly on the size of the dinner plate rather than the amount of food on the plate [18]. Likewise, how much people drink depends more on the characteristics of the glass than on the volume in it. And what magazines people buy depends on what is displayed at the supermarket checkout counter. In a well-known psychology experiment subjects' soup consumption was shown to be more strongly correlated to the behavior of other diners (disguised experimenters) than to other variables like hunger or portion size. Thus, what peoples' peers think or do has a substantial impact on a person's actions and decisions. These setting biases also likely apply to decisions with more serious consequences, like results of which RCTs to accept. For example, results of an RCT from a 'distinguished' site or a site where one has colleagues may be more readily accepted than a larger, better executed study from other sites.

Another important bias in interpreting results of RCTs relates to sample size. One might assume clinicians give greater weight to results of RCTs of many versus few subjects. However, this may not be so. Substantial data from psychology studies suggest people are more influenced by data about few individuals than many [19]. This is because emotion is an important aspect of decision-making. For example, people typically have a higher emotional response to learning about one starving African child than they do when learning about a famine killing millions. Joseph Stalin expressed this sentiment concisely: 'One man's death is a tragedy. A thousand deaths is a statistic.' The consequence of this interplay between sample size and emotion is that hematologists may give greater consideration to results of a small RCT, especially when its conclusions are concordant with their prejudice before the RCT than to a large RCT whose conclusions are discordant with their perception before the RCT. Can we overcome this prejudice by combining these studies into a meta-analysis? We shall see.

Meta-Analyses

Meta-analyses combine results of several studies that address a set of related research hypotheses. They are highly regarded in the hierarchy of clinical evidence, are highly cited and constitute the basis for most conclusions in this special issue [20–22]. Meta-analyses can be applied to data from RCTs, ODBs or both. Here we focus on meta-analyses of data from RCTs.

Conclusions from meta-analyses are widely used to determine the relative efficacy of a new therapy, especially when it is difficult to perform large RCTs. Their predominant use in the context of blood diseases is to increase power when effect size is small and individual studies lack sufficient power to detect an effect size of this magnitude. As discussed, these are often the limitations of RCTs in blood diseases. Meta-analyses are used to adjudicate disparate results of RCTs.

There are several intrinsic weaknesses of meta-analyses including the pooling of biases of the studies included, biases introduced by the process of selecting RCTs for inclusion and heterogeneity between the RCTs included. Publication bias is an important limitation of meta-analyses: RCTs with favorable outcomes are more likely to be published than those with unfavorable outcomes. A relevant example is an analysis of publications regarding prognostic variables in cancer: 90–95% of all reports in 2005 were positive [23]. Recent reviews find about 50% of meta-analyses were affected by publication bias. Addition of the unreported studies to the meta-analysis substantially changed the conclusion in 10–20% of analyses [24, 25]. This deficiency can be partly adjusted for by sensitivity analyses in which variation (uncertainty) in the conclusion of a meta-analysis can be apportioned, qualitatively or quantitatively, to different sources of variation in the input of the model. Also, conclusions of meta-analyses may be influenced by the funding source [26] and other variables. Several studies compared whether conclusions of meta-analyses are concordant with those of large clinical trials with disparate conclusions. For example, one study that compared the magnitude and uncertainty of treatment effects found discordances in 10–23% of clinical settings. Few of these studies were in blood disease, reflecting their rarity [6].

Meta-analyses of RCTs are most useful in analyzing new therapies of blood diseases when the clinical trial design and therapies (including drugs, doses and schedules) and the subjects being studied are quite similar or identical. However, variations in these introduce heterogeneity which can be approached, but not solved, using a

random-effect model. However, if we think any one or more of these variables interacts with treatment or if there is significant heterogeneity across studies we should not combine these trials in a meta-analysis. Instead, we should conclude that the treatment effect cannot be considered ‘generic’ but depends on the type of subjects treated or on specifics of the intervention.

How often are results of meta-analyses of RCTs correct? Answering this assumes the correct answer is known. In one study, results of meta-analyses performed before data from a large RCT were available were compared to the results of the large RCT, the results of which were assumed to be correct [27]. About one third of the time results of the RCT were discordant with results of the prior meta-analysis. False-positive and -negative results were detected. Similar analyses have been reported by others with error rates varying from 10 to 40% [28]. These data suggest results of meta-analyses, especially in the sphere of new therapies of blood disease where there are few large RCTs, are unlikely to completely resolve controversies regarding therapy strategies.

Experts in meta-analysis admit the limitations of this approach; some consider it an oversimplification of complex data. In considering results of meta-analyses it is important to separately evaluate the quality of the studies included and the consistency (or inconsistency) of the results. Equally important is transparency in how studies were selected for inclusion. In doing so, the reviewer must rely on his or her expert opinion. Not unexpectedly, experts may seemingly reach different conclusions about the value of a specific meta-analysis. How to deal with this variability in opinions on meta-analyses, RCTs or other issues is discussed below.

Another issue is whether results of ODBs (discussed below) should be included in meta-analyses typically restricted to data from RCTs? In a random analysis of meta-analyses from the Cochrane Collaboration, about one third of meta-analyses had data from two RCTs [29]. Some data suggest adding data from ODBs to these meta-analyses of RCTs is useful [30]. This is especially so if the anticipated benefits of adding data from ODBs (which depends on variables like size and quality) exceed the anticipated risks (inherent biases) by a substantial margin. For example, when there are few if any large RCTs and/or when there is substantial variability and/or biases in these trials. This situation, unfortunately, applies to most trials of new therapies of blood diseases.

Finally, a fundamental limitation of most meta-analyses is the lack of subject level data. This limits the ability to assure comparability of seemingly similar studies, to

explore variables correlated with outcomes and to combine results of RCTs and ODBs amongst others. Some meta-analyses overcome this limitation by obtaining access to subject level data.

Observational Databases

Several types of observational studies can also help determine whether a new therapy of a blood disease is better than current therapy including: (1) controlled cohort studies and (2) case-control studies. For example, for certain rare blood diseases or complex infrequent therapies, RCTs cannot be done. Here, we focus on the use of ODBs in the context of determining therapy benefit.

The necessary starting point for using ODBs is a high-quality dataset. High-quality ODBs should contain subject level data. Consecutiveness, accuracy and completeness of reporting must be known. Consensus on requirements for high-quality ODBs were recently published [31]. The potential advantages of high-quality ODBs over experimental studies, including RCTs, in determining therapy benefit are several-fold. The most important advantage is the vastly larger numbers of subjects available for analyses. This increases the power of analyses but poses specific problems (see below). Also, subjects are more diverse in ODBs than in RCTs increasing potential applicability and generalizability of conclusions. Less precisely defined therapy approaches are often used, e.g. several techniques of T-cell depletion may be tested. The advantage of this is to allow a more generic conclusion like whether T-cell depletion is effective, as well as to explore whether these various techniques have similar effects. ODB studies are far less costly than RCTs.

Limitations of ODBs

There are important limitations to ODBs in the context of determining therapy benefit. For example, heterogeneity of subject-, disease- and therapy-related variables makes it important to ensure comparability of subjects in the therapy arms. Although adjustments can be made for known prognostic variables, there can be no such adjustment for unknown variables or variables that might operate in a new therapy setting. This is probably the most important difference between ODBs and large RCTs where comparability of subjects in therapy arms is not ensured by random assignment. Also, less precisely specified therapy approaches may result in combining approaches of differing efficacy.

Another limitation of ODBs is that diagnostic criteria and observation schedules vary between centers. These can result in important ascertainment biases. For example, different centers may use different criteria to diagnose 'leukemia relapse' (for example, in chronic myelogenous leukemia, cytogenetic relapse at one center and molecular relapse at another). Also, bone marrow examinations may be performed at different frequencies at different centers. Another issue is time-to-treatment biases. This is especially important in some blood diseases with a rapidly fatal course, like some leukemias and lymphomas [32]. Other selection biases also operate: clinicians may preferentially use a specific therapy more in one group of subjects than another. Adjustment can be if these selection biases are known. However, some variables leading to selection biases may be unknown or unrecorded. Studies using ODBs are restricted to the data being routinely collected and it is difficult or impossible to obtain additional non-standard data on subjects. This may make ODB studies difficult to implement in certain settings. Finally, like RCTs, ODB studies are often behind the times in a field because of the need to accumulate sufficient numbers of subjects with adequate follow-up.

The fundamental criticism of observational studies is that unrecognized confounding factors may distort results. However, the common belief that confounding with subsequent distortion is common and unpredictable is incorrect in most instances (see below).

How Do Results of RCTs, Meta-Analyses and ODBs Compare?

Clinicians can be more certain a therapy strategy is correct when their decision is based on concordant results of several large RCTs and of appropriately conducted meta-analyses of these trials. However, as discussed, these data are often unavailable. Here, analyses from ODBs may help. Several recent studies compared results of RCTs and ODBs. One study showed remarkably similar point estimates for the effect of therapy between RCTs and ODBs [33]. A second study compared results of RCTs and ODBs for 19 diverse treatment-settings [34]. Estimates of treatment effects were concordant in 17; in only 2 settings was the magnitude of the effect in the ODBs outside the 95% confidence interval for the RCTs. A third study compared results of ODBs and RCTs in 8 settings [35]. Concordant results were found in 7. In a fourth study, results of RCTs and ODBs were compared in 45 settings [28]. Odds ratios were similar for both techniques (about 85% of comparisons) despite substantial between-

study heterogeneity for both (greater for ODBs than RCTs). Discordant results were found for <10% of prospective studies. As expected, few of these studies involved blood diseases. A fifth study found concordant conclusions in six of eight settings compared [36]. Combining these data, estimates of treatment effect from ODBs and RCTs were concordant in >90% of settings. In the rare discordances, ODBs estimated greater benefit than seemingly comparable RCTs. Often the magnitude of discordance in estimating treatment effect correlates with study quality: low for high-quality studies but high for low-quality studies [7, 37].

What should hematologists conclude when results of RCTs, meta-analyses and/or ODBs are discordant? First, they need to consider three issues: (1) Are the subjects and the therapies which are being examined comparable? (2) Is follow-up comparable? (3) Were similar end points used? ODBs almost always include more diverse subjects than RCTs or meta-analyses and almost always include more diverse doses and schedules of drugs than RCTs and meta-analyses. Because ODBs frequently have substantially longer follow-ups than RCTs, they are more likely to use a survival end point than RCTs or meta-analyses which may focus on disease- or progression-free survival. A reasonable conclusion is that ODBs, RCTs and meta-analyses will likely reach concordant conclusions when the magnitude of bias inherent in ODBs is small compared to the variability and biases inherent in RCTs and meta-analyses (e.g. choice of study subjects, disease, disease states, drugs, doses and schedules) [33].

These data suggest results of RCTs, although useful, are often incomplete and/or contradictory. These studies also typically have relatively brief follow-up. These limitations are only partially solved by meta-analyses which cannot overcome flaws inherent in the RCTs included. There are several circumstances when data from ODBs may be more useful to the clinician and often more definitive than RCTs or meta-analyses because the benefit achieved by analyzing large numbers of subjects with diverse features exceeds potential biases. Examples include: (1) when RCTs are not feasible because of small numbers of subjects; (2) when several small RCTs do not reach a definitive conclusion or reach contradictory conclusions that cannot be resolved by meta-analyses; (3) when long-term follow-up is of interest, and (4) when an RCT is not feasible because an unresolved issue is no longer the focus of clinical research interest. In areas of current clinical interest, clinicians are certainly most comfortable when a conclusion is supported by concordant outcomes of RCTs and ODBs.

Expert Opinion

Francis Galton was apparently surprised when he noted that the averaged guesses of a crowd at a county fair estimated the weight of an ox more accurately than the estimate of any one expert butcher [recounted in ref. 38]. There are many similar examples of the advantage of aggregated group opinions over individual judgments. This advantage of quantifying a diversity of opinions hinges on several factors: (1) each person should have comparable data; (2) each opinion should be independent; (3) each person should use personal knowledge, and (4) there needs to be a structured mechanism for turning private judgments into an aggregated estimate. In some regard, quantifying expert opinion is a less-structured form of meta-analysis.

Some scientists regard expert opinion as the highest level of evidence in therapy decision-making. This is because (hopefully) experts' calculi include diverse lower-order data including biological plausibility, case reports and series, personal experiences and data from RCTs, ODBs and meta-analyses thereof. However, not everyone shares this view. Some consider expert opinion unreliable. In the US Public Health Service Preventative Services Task Force hierarchy of evidence and elsewhere, expert opinion is judged the lowest-level type of evidence [20, 22]. Likewise in the Grading of Recommendations, Assessment, Development and Evaluations (GRADE) system (<http://clinicalevidence.bmj.com/ceweb/about/about-grade.jsp>) no credit is given to expert opinion. How can we explain these discordant views?

One problem with expert opinion that may explain its low ranking is that different experts seem to reach different conclusions about the best therapy for a specific person or disease state. Or do they? As we will discuss, experts rarely disagree on substantial issues. Most disagreements can be traced to a few variables: (1) different datasets; (2) different perceptions about the target subject, and (3) different perceptions about the question being posed (e.g. is it the best outcome, the most cost-effective or the most interesting research question?). When these ambiguities are removed, experts rarely disagree. This, of course, does not mean their consensus opinion is correct: there was once agreement the sun circled the earth! (Galileo apparently disagreed, but later changed his mind, for good reason.)

Consensus Methods and Building

Consensus methods are another approach to compare the effectiveness of two therapies of blood diseases. Although these methods vary considerably, several important elements are shared: (1) panel composition; (2) data compilation and analysis; (3) quantifying opinions; (4) analysis of possible divergent opinions, and (5) expression of outcomes.

Determining who appropriate panel members are is fundamental to the success of the consensus process [39]. Panels composed entirely of believers or non-believers in a new therapy are unlikely to reach a balanced, reproducible conclusion because of confirmation bias (see above). Moreover, mixed panels of believers and non-believers may fail to reach consensus because of attitude bias (also see above). In considering panel composition, it is important to replicate the target environment in which therapy decision-making will occur. For example, it may be important to have diverse geographic representation if the conclusions are to be applied nationally or internationally. Considerable data show substantial geographical variation in the use of radiation therapy versus surgery versus the watch-and-wait approach in managing prostate cancer in comparable subjects [40]. Often, these geographical variations are not explained by appropriate or inappropriate technology use [41]. Also, Americans and Europeans often view new therapies from different perspectives driven by issues other than efficacy, like cost and resource use. It is important that proponents and opponents of a new therapy be represented. Panel size is also important: panels of fewer than 9 members rarely have reproducible results; larger panels are unwieldy and expensive.

Panelists should be informed of all data relevant to the topic being considered. Published data can be retrieved, evaluated (for example, for level of evidence) and summarized. Inclusion of data from RCTs, ODBs and meta-analyses thereof is important. Several systems for ranking the quality of evidence are widely used [20, 22]. There are factors which limit these rankings, like publication biases (discussed above). Above, we discussed disagreements over the quality-of-evidence rankings of data inputs like ODBs and expert opinion. A more important challenge/problem is ensuring all panelists know unpublished data including abstracts, meeting reports and personal experiences. This can be accomplished in interactive sessions (see below).

Many seemingly discordant opinions amongst experts arise from failure to precisely define the question(s). For

example, when asked which of two therapies is more appropriate, experts may envision rather different subjects with different disease-related prognostic variables and prior therapies. This can be overcome by precisely defining disease-related prognostic variables (like age and gender) by hierarchical permutation of relevant variables followed by ranking of precisely defined subjects. Experts can also be polled before ranking to determine which variables they believe are needed to estimate the best therapy. Typically, experts identify fewer than 10 variables, often fewer than 5, which influence their decision-making. As expected, lists of variables from different experts often overlap substantially. Examples include age, gender, disease stage, performance score, prior therapy response and response duration. However, recursive-partitioning analyses of expert opinion after ranking often shows experts use only a subset of the variables they believed were important before ranking.

Another requisite for success is a clearly defined question. Usually this is what the best therapy is for the subject being considered. Experts need to be explicitly told not to consider collateral issues like costs, resource use and societal values (like answering a question that benefits future subjects but not the subject being considered).

Delphi Analyses

There are several consensus techniques for quantifying expert opinion [reviewed in ref. 42]; a detailed discussion is beyond our scope. At one extreme, experts independently review data and the sum of their opinions is expressed quantitatively or qualitatively, ideally with a description of the variance. Other techniques involve bringing experts together in a structured or unstructured format. Again, the sum of their opinions can be expressed quantitatively or qualitatively. Unfortunately, there are few data on internal and external validation of these processes. For example, how many experts should be included, should there be one or several sessions, will individual opinions be confidential and are results expressed quantitatively or qualitatively? Other techniques are more elaborate. For example, Delphi consensus panels, developed at the RAND corporation to analyze issues like nuclear weapons and fighter aircraft design, use defined numbers of panelists in multiple iterative, interactive sessions combined with anonymous voting and sophisticated data analyses [43–48]. The Delphi panel technique yields specific criteria of appropriateness that can be used as the basis for treatment guidelines.

Reliability of these expert consensus techniques requires internal and external validation. There are few data addressing this issue. For example, most techniques have not been rigorously tested for internal validity: are results of a consensus panel reproducible if done by the same panel 6 months or 1 year later? More importantly, are they reproducible when the same data are concurrently reviewed by a different expert panel? An exception is Delphi panels where there are considerable data indicating recommendations of similarly constituted panels are reproducible when concurrent panels review similar data [49]. One test of external validity is whether recommendations of Delphi analyses are concordant with subsequently performed RCTs. Recent analyses indicate reasonably high concordance [50]. Other data are likewise supportive [51]. Similar tests of external validity of other consensus methods are less often reported.

Added to the cautions we discuss above is another important and interesting limitation, partially related to but distinct from confirmation bias: experts are easier to fool than non-experts. This is because they jump to unwarranted conclusions seeing immediately the direction an issue is headed and because they tend to focus on subtleties rather than larger issues [summarized in ref. 52]. For example, 54 wine experts were given red wines to evaluate several of which were really white wines with a tasteless red dye added. None of the experts detected the subterfuge: they were off considering the subtleties of red wines like terroir, bouquet, tannins and the like. In contrast, non-experts were more likely to correctly identify the red wines as white because they were not so easily subverted. In a similar vein, several fish experts claimed to be able to distinguish between the unappealing Patagonian toothfish and the ever-popular and, consequently, now nearly extinct Chilean sea bass. They are the same fish. That said, careful and critical quantification of expert opinion can be useful, especially if there is concordance with conclusions from RCTs, ODBs and results of meta-analyses.

Conclusions

Although large RCTs are rightly considered the gold standard study design, many RCTs of new therapies of blood diseases have modest sample size and may require other compromises such as pooling of therapy regimens or subjects to increase sample sizes. This can lead to inconsistency in results across multiple trials. These limitations can be addressed, but only in part, using meta-anal-

yses. Reliance on ODBs and structured quantification of expert opinion provide alternatives which have strengths that complement conclusions of RCTs and meta-analyses, including more and more diverse subjects, greater generalizability, lower cost and the ability to examine consistency of treatment effects. Although there is greater potential for bias in ODB-based analyses, several studies indicate frequent concordant conclusions for RCTs, ODBs, meta-analyses and structured quantification of expert opinion. Care must be taken. We suggest that whilst results of RCTs and meta-analyses thereof are of considerable value in evaluating efficacy of new therapies of blood diseases, data from ODBs and structured quantification of expert opinion are also helpful. Results of analyses of ODBs can be included into meta-analyses to improve their outcome. We need all the help we can get in making complex therapy decisions in blood diseases.

Acknowledgments

We tackled several of the issues in this paper in a previous report [53] with three additional co-authors (M. Eapen, B. Logan and M.J. Zhang) with a far greater knowledge of statistics than ours. The focus of that report was on the value of ODBs in addressing therapy questions in bone marrow and blood cell transplants. Here, we focus on a broader topic, therapy of blood diseases and on the value and limitations of meta-analyses of RCTs to complement a special issue on the use of meta-analyses in blood diseases in which it appears. As many of the concepts overlap, there is, unavoidably, similarity between these reports. We are grateful to our prior co-authors for sharpening our thinking. David W. Golde provided knowledge on the dicey issue of the Patagonian toothfish and Chilean sea bass. Sabine Jacob kindly prepared the typescript. R.P.G. is an employee of Celgene Corporation.

References

- 1 Herland K, Akselsen JP, Skjongsberg OH, Bjermer L: How representative are clinical study patients with asthma or COPD for a larger 'real life' population of patients with obstructive lung disease? *Respir Med* 2005; 99:11–19.
- 2 Rothwell PM: Can overall results of clinical trials be applied to all patients? *Lancet* 1995; 345:1616–1619.
- 3 Mant D: Can randomized trials inform clinical decisions about individual patients? *Lancet* 1999;353:743–746.
- 4 Lambert PC, Sutton AJ, Abrams KR, Jones DR: A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *J Clin Epidemiol* 2002;55:86–94.

- 5 Horwitz RI: Complexity and contradiction in clinical trial research. *Am J Med* 1987;82: 498–510.
- 6 Ioannidis JP: Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 2005;294:218–228.
- 7 Kunz R, Oxman AD: The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* 1998;317:1185–1190.
- 8 Logan B, Leifer E, Bredeson C, Horowitz M, Ewell M, Carter S, Geller N: Use of biological assignment in hematopoietic stem cell transplantation clinical trials. *Clin Trials* 2008;5: 607–616.
- 9 Beswoda WR, Seymour L, Dansey RD: High-dose chemotherapy with hematopoietic rescue as primary treatment for metastatic breast cancer: a randomized trial. *J Clin Oncol* 1995;13:2483–2489.
- 10 Stadtmauer EA, O'Neill A, Goldstein LJ, Crilley PA, Mangan KF, Ingle JN, Brodsky I, Martino S, Lazarus HM, Erban JK, Sickles C, Glick JH: Conventional-dose chemotherapy compared with high-dose chemotherapy plus autologous hematopoietic stem-cell transplantation for metastatic breast cancer. Philadelphia Bone Marrow Transplant Group. *N Engl J Med* 2000;342:1069–1076.
- 11 Tallman MS, Gray R, Robert NJ, LeMaistre CF, Osborne CK, Vaughan WP, Gradishar WJ, Pisansky TM, Fetting J, Paietta E, Lazarus HM: Conventional adjuvant chemotherapy with or without high-dose chemotherapy and autologous stem-cell transplantation in high-risk breast cancer. *N Engl J Med* 2003;349:17–26.
- 12 Farquhar CM, Marjoribanks J, Lethaby A, Bassier R: High dose chemotherapy for poor prognosis breast cancer: systematic review and meta-analysis. *Cancer Treat Rev* 2007; 33:325–337.
- 13 Manjoo F: *True Enough: Learning to Live in a Post-Fact Society*. Hoboken, Wiley, 2008.
- 14 Lord C, Ross L, Lepper M: Biased Assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence. *J Pers Soc Psych* 1979;37:2098–2109.
- 15 Bacon F: *Novum organum*. 1620.
- 16 Westen D, Blagov PS, Harenski K, Kilts C, Hamann S: Neural bases of motivated reasoning: an fMRI study of emotional constraints on partisan political judgment in the 2004 US Presidential election. *J Cogn Neurosci* 2006;18:1947–1958.
- 17 Thaler RH, Sunstein CR: *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven, Yale University Press, 2008.
- 18 Wansink B: *Mindless Eating: Why We Eat More than We Think*. New York, Bantam Books, 2006.
- 19 Ariely D: *Predictably Irrational: the Hidden Forces that Shape our Decisions*. New York, Harper Collins, 2008.
- 20 Preventive Services Task Force: Guide to Clinical Preventive Services: Report of the US Preventive Services Task Force, ed 2. Baltimore, Williams & Wilkins, 1996.
- 21 Patsopoulos NA, Analatos AA, Ioannidis JP: Relative citation impact of various study designs in the health sciences. *JAMA* 2005;293: 2362–2366.
- 22 Harbour R, Miller J: A new system for grading recommendations in evidence-based guidelines. *BMJ* 2001;323:334–336.
- 23 Kyzas PA, Denaxa-Kyza D, Ioannidis JP: Almost all articles on cancer prognostic markers report statistically significant results. *Eur J Cancer* 2007;43:2559–2579.
- 24 Palma S, Delgado-Rodriguez M: Assessment of publication bias in meta-analyses of cardiovascular diseases. *J Epidemiol Community Health* 2005;59:864–869.
- 25 Sutton AJ, Cooper NJ, Abrams KR, Lambert PC, Jones DR: Bayesian approach to evaluating net clinical benefit allowed for parameter uncertainty. *J Clin Epidemiol* 2005;58:26–40.
- 26 Jørgensen AW, Hilden J, Gøtzsche PC: Cochrane reviews compared with industry supported meta-analyses and other meta-analyses of the same drugs: systematic review. *BMJ* 2006;333:782.
- 27 LeLorier J, Grégoire G, Benhaddad A, Lapiere J, Derderian F: Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *N Engl J Med* 1997;337:536–542.
- 28 Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, Contopoulos-Ioannidis DG, Lau J: Comparison of evidence of treatment effects in randomized and non-randomized studies. *JAMA* 2001; 286:821–830.
- 29 Shrier I: Cochrane Reviews: new blocks on the kids. *Br J Sports Med* 2003;37:473–474.
- 30 Shrier I, Boivin JF, Steele RJ, Platt RW, Furlan A, Kakuma R, Brophy J, Rossignol M: Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? A critical examination of underlying principles. *Am J Epidemiol* 2007;166:1203–1209.
- 31 Vandembroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Schlesselman JJ, Egger M: Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *Ann Intern Med* 2007;147:w163–w194.
- 32 Sekeres MA, Elson P, Kalaycio ME, Advani AS, Copelan EA, Faderl S, Katarjian HM, Estey E: Time from diagnosis to treatment initiation predicts survival in younger, but not older, acute myeloid leukemia patients. *Blood* 2009;113:28–36.
- 33 Concato J, Shah N, Horwitz RI: Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;342:1887–1892.
- 34 Benson K, Hartz AJ: A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000;342:1878–1886.
- 35 Britton A, McPherson K, McKee M, Sander-son C, Black N, Bain C: Choosing between randomised and non-randomised studies: a systematic review. *Health Technol Assess* 1998;2:i–iv.
- 36 Guyatt GH, DiCenso A, Farewell V, Willan A, Griffith L: Randomized trials versus observational studies in adolescent pregnancy prevention. *J Clin Epidemiol* 2000;53:167–174.
- 37 MacLehose RR, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AM: A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technol Assess* 2000;4:1–154.
- 38 Surowiecki J: *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. New York, Random House, 2004.
- 39 Herrin J, Etchason JA, Kahan JP, Brook RH, Ballard DJ: Effect of panel composition on physician ratings of appropriateness of abdominal aortic aneurysm surgery: elucidating differences between multispecialty panel results and specialty society recommendations. *Health Policy* 1997;42:67–81.
- 40 Krupski TL, Kwan L, Afifi AA, Litwin MS: Geographic and socioeconomic variation in the treatment of prostate cancer. *J Clin Oncol* 2005;23:7881–7888.
- 41 Chassin MR, Koseoff J, Park RE, Winslow CM, Kahn KL, Merrick NJ, Keesey J, Fink A, Solomon DH, Rook RH: Does inappropriate use explain geographic variations in the use of health care services? A study of three procedures. *JAMA* 1987;258:2533–2537.
- 42 Goodman C, Baratz SR (eds): *Improving Consensus Development for Health Technology Assessment: An International Perspective*. Council on Health Care Technology. Washington, Inst Med. Natl Academy Press, 1990.
- 43 Brook RH, Chassin MR, Fink A, Solomon DH, Koseoff J, Park RE: A method for the detailed assessment of the appropriateness of medical technologies. *Int J Technol Assess Health Care* 1986;2:53–63.
- 44 Leape LL, Hilborne LH, Leape LL, Fiske ME, Park RE, Kamberg CJ, Brook RH: The appropriateness of coronary artery bypass surgery in New York State. *JAMA* 1993; 269:753–760.
- 45 Bernstein SJ, Hilborne LH, Leape LL, Fiske ME, Park RE, Kamberg CJ, Brook RH: The appropriateness of coronary artery bypass surgery in New York State. *JAMA* 1993;269: 766–769.

- 46 Bernstein SJ, McGlynn EA, Siu AL, Roth CP, Sherwood MJ, Keesey JW, Kosecoff J, Hicks NR, Brook RH: The appropriateness of hysterectomy. A comparison of care in seven health plans. Health Maintenance Organization Quality of Care Consortium. JAMA 1993;269:2398–2402.
- 47 Bengtson A, Herliz J, Karlsson T, Brandrup-Wognsen G, Hjalmarson A: The appropriateness of performing coronary angioplasty and coronary revascularization in a Swedish population. JAMA 1994;271:1260–1265.
- 48 Gray D, Hampton JR, Bernstein SJ, Kosecoff J, Brook RH: Audit of coronary angiography and bypass surgery. Lancet 1990;335:1317–1320.
- 49 Pearson SD, Margolis CZ, Davis S, Schreier LK, Sokol HN, Gottlieb LK: Is consensus reproducible? A study of an algorithmic guidelines development process. Med Care 1995;33:643–660.
- 50 Shekelle PG, Chassin MR, Park RE: Assessing the predictive validity of the RAND/UCLA appropriateness method criteria for performing carotid endarterectomy. Int J Technol Assess Health Care 1998;14:707–727.
- 51 Kravitz RL, Laouri M, Kahan JP, Guzy P, Sherman T, Hilborne L, Brook RH: Validity of criteria used for detecting underuse of coronary revascularization. JAMA 1995;274:632–638.
- 52 Dolnick E: Fish or foul? Opinions. New York Times, Sept 2, 2008.
- 53 Gale RP, Eapen M, Logan B, Zhang MJ, Lazarus H: Are there roles for observational database studies and structured quantification of expert opinion to answer therapy controversies in transplants? Bone Marrow Transplant 2009;43:435–446.