How to Ascertain Drug Safety in the Context of Benefit. Controversies and Concerns

LEE S. SIMON, C. VIBEKE STRAND, MAARTEN BOERS, PETER M. BROOKS, PETER S. TUGWELL, CLAIRE BOMBARDIER, JAMES F. FRIES, DAVID HENRY, LARRY GOLDKIND, GORDON GUYATT, ANDREAS LAUPACIS, LARRY LYND, TOM MacDONALD, MUHAMMAD MAMDANI, ANDREW MOORE, KEN S. SAAG, ALAN J. SILMAN, RANDALL STEVENS, and ALAN TYNDALL

ABSTRACT. There is great concern about clearly defining benefit and risk in the context of both drug development and clinical practice. In view of this pressure, the OMERACT Executive identified the need to bring together clinical trialists, pharmacoepidemiologists, clinicians, clinical epidemiologists, statistical experts, and regulatory representatives to discuss different approaches to define risk and perhaps improved ways to express it. Each attendee spoke on a given topic and the group was charged to consider the issue of risk in the context of formally posed questions. This article provides a summary of the presentations and outlines the discussions that followed. (J Rheumatol 2009; 36:2114–21; doi:10.3899/jrheum.090591)

Key Indexing Terms:
CLINICAL TRIALS
DRUG TOXICITY
RISK ASSESSMENT
PHARMACOEPIDEMIOLOGY

The need for infinite knowledge concerning all possibilities of safety is a stifling development of new therapies. Is the answer to take drug safety monitoring out of "Pharma" development? If yes, what are the alternatives? What are the roles of the clinician and the patient, and who is the risk manager?

Two disciplines have been very active in this area and each could benefit from a more formal working arrangement. Clinical epidemiology has tended to focus on benefit, whereas pharmacoepidemiology has focused on safety. Pharmacoepidemiology has typically used tools developed for pharmacoviligance that capitalize on postmarketing

From the Harvard Medical School, Boston, Massachusetts; Division of Immunology, Stanford University, Palo Alto, California, USA; VU University Medical Center, Amsterdam, The Netherlands; Faculty in Health Services, University of Queensland, Royal Brisbane Hospital, Herston, Australia; Centre for Global Health, University of Ottawa, Ottawa; Health Care Research and Clinical Decision Making, Toronto General Research Institute and Institute for Work and Health, Toronto, Canada; Division of Immunology and Rheumatology, Stanford University School of Medicine, Palo Alto, California; Department of Gastroenterology, Uniformed Services University of the Health Sciences, Bethesda, Maryland, USA; Department of Clinical Epidemiology and Biostatistics, Department of Medicine, McMaster University, Hamilton; Institute for Clinical Evaluative Services, Toronto; Keenan Research Centre and Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto; Faculty of Pharmaceutical Sciences, University of British Columbia, Centre for Clinical Epidemiology and Evaluation, Vancouver Coastal Health Research Institute, Vancouver, Canada; Medicines Monitoring Unit, Medicine and Therapeutics, Ninewells Hospital and Medical School, University of Dundee, Nethergate, Dundee, Scotland; Pain Research Unit and Nuffield Department of Anaesthetics, University of Oxford, Oxford Radcliffe Hospital, Headington, Oxford, England; University of Alabama, Center for Education and Research on Therapeutics in Musculoskeletal Disorders (CERTs), Birmingham, Alabama, USA; Arthritis Research Campaign, Chesterfield, England; Inflammation and Immunology, Celgene Corporation, and University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School, New Brunswick, New Jersey, USA; and Department of Rheumatology, University of Basel, Basel, Switzerland.

L.S. Simon, MD, Associate Clinical Professor of Medicine, Harvard Medical School; C.V. Strand, MD, Clinical Professor of Medicine, Adj, Division of Immunology, Stanford University; M. Boers, MD, Professor of Clinical Epidemiology, VU University Medical Center; P.M. Brooks, MD, Faculty in Health Services, University of Queensland, Royal Brisbane Hospital; P.S. Tugwell, MD, Director, Centre for Global Health, University of Ottawa; C. Bombardier, MD, Professor, Canada Research Chair in Knowledge Transfer for Musculoskeletal, Health Policy Management and Evaluation; Director, Health Care Research and Clinical Decision Making, Toronto General Research Institute; and Clinical Research Coordinator, Institute for Work and Health; J.F. Fries, MD, Professor, Division of Immunology and Rheumatology, Stanford University School of Medicine; L. Goldkind, MD, Department of Gastroenterology, Uniformed Services University of the Health Sciences; G. Guyatt, MD, MSc, FRCPC, Professor, Department of Clinical Epidemiology and Biostatistics, Department of Medicine, McMaster University; D. Henry, MD, President and CEO, Institute for Clinical Evaluative Services; A. Laupacis, MD, Keenan Research Centre, Executive Director, Li Ka Shing Knowledge Institute, St. Michael's Hospital; L. Lynd, PhD, Professor, Faculty of Pharmaceutical Sciences, University of British Columbia, Centre for Clinical Epidemiology and Evaluation, Vancouver Coastal Health Research Institute; T. MacDonald, MD, Professor, Head, The Medicines Monitoring Unit, Medicine and Therapeutics, Ninewells Hospital and Medical School, University of Dundee; M. Mamdani, PharmD, MA, MPH, Director, Applied Health Research Centre, Keenan Research Centre, Li Ka Shing Knowledge Institute, St. Michael's Hospital; A. Moore, DSc, Pain Research Unit and Nuffield Department of Anaesthetics, University of Oxford, Oxford Radcliffe Hospital; K.G. Saag, MD, Associate Professor, Medicine Director, University of Alabama, Center for Education and Research on Therapeutics in Musculoskeletal Disorders; A.J. Silman, FRCP, FMedSci, DSc(Hons), Arthritis Research Campaign; R. Stevens, MD, Vice President and Head, Inflammation and Immunology, Celgene Corporation, Clinical Associate Professor of Medicine, University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School; A. Tyndall, MD, Department of Rheumatology, University of Basel, Felix Platter-Spital.

Address correspondence to Dr. L.S. Simon. E-mail: omeract@uottawa.ca

data. We who analyze and summarize the evidence have a responsibility to provide a balanced view of both perspectives to approval agencies, policymakers, patients, and practitioners.

As noted in an introductory article[1], a group of international experts convened to address issues regarding drug safety to assess how OMERACT might contribute in this area.

1. ASSIGNING CAUSALITY (Larry Goldkind)

"Safety signals" are defined as any new piece of information or any spontaneous report about a drug that is clinically relevant; signals can arise from many different sources, including postmarketing data, preclinical data, and case study series. Because the term is extremely broad, it may be necessary to assign a threshold of concern when a safety signal triggers a need for further study. After a signal is identified, it should be assessed to determine whether it represents a potential safety risk; whether the event was caused by the product; and whether any action should be taken.

The term "causality assessment" is most commonly applied to individual case reports, which then become part of an analysis of a case series of events. In this setting each case is assessed for causality using the following criteria: (a) a temporal relationship; (b) a previous report with the same or a related drug; (c) any confounding drug use; (d) any confounding by disease; and (e) clinical plausibility.

Once this information is extracted from records, the case is assigned to one of 5 causality likelihood categories ranging from certain to unlikely or unclassifiable. Causality assessment methodologies apply criteria to each case, and likelihoods are assigned by expert opinion and templated questionnaires/algorithms. The Bradford Hill Criteria used by the US Food and Drug Administration (FDA) are incomplete and could be supplemented by using biologic plausibility or Bayesian methodologies. For imbalances in event rates in controlled studies, the role of causality assessment is more limited since rare events in randomized clinical trials rely heavily on case by case causality assessments. Causality assessment of single cases has little role in evaluating adverse events that have significant background rates, are confounded by the underlying disease being studied, or have long latency to clinical outcome. Ultimately the most critical issue when looking at a case or small case series of postmarketing adverse event reports or an imbalance within a clinical trial database is to clearly articulate the settings where the severity of the adverse event in question warrants definitive quantification of risk. For such cases there is no way around the need for large databases.

Discussion. Assigning causality needs an appreciation of the limitations of current methods to assign causality. For example, there are differences of opinion in deciding what a signal is, how to define benefit/risk, and which methodologies applied are robust. The majority of cases fall between probable and possible. Signals can come from any new pieces of information about a drug that are clinically relevant; these include any spontaneous report. There is no agreement on the magnitude of frequency, i.e., when to say the number of events in this exposed group is higher than expected. The Bradford Hill criteria used by the FDA are perhaps incomplete and could be supplemented by biologic plausibility and Bayesian logic. Some participants considered it important, when an association is found, to assess the magnitude of imbalance, the severity of the event, whether it was biologically plausible, and, if so, whether it was consistent. Rare events will require mathematical modeling. There was debate around the statement that "Association is per group; causality, on an individual basis." Epidemiologists strongly disagree with the former: causality also applies at group level.

2. UTILITY OF LARGE STREAMLINED TRIALS TO DEFINE RISK (Andreas Laupacis, Tom MacDonald)

Many questions could be answered by large, streamlined, randomized controlled trials (RCT). Large streamlined safety studies done in the setting of usual care, using designs such as the Prospective Randomised Open Blinded Endpoint (PROBE), can generate data on drug effectiveness and safety with excellent external validity. Such data are very useful to inform policy decisions.

Large streamlined trials for more common rheumatological disorders such as osteoarthritis require moderate-sized base populations, but those for rarer disorders need much larger base populations. However, whole-country databases for conducting such trials make these goals feasible. In Europe such trials can be mounted in the United Kingdom, Denmark, Sweden, Finland, and The Netherlands, and probably in some parts of Italy and Spain. Randomization of large cohorts to two (or more) treatment arms, each of which uses an efficacious treatment, allows for comparisons of safety and effectiveness within an experimental design. Such designs are far superior to observational studies, where channeling and other biases make the comparisons unreliable. While it is true that such designs become more observational with time (this is true of all studies with prolonged followup), the baseline randomization allows for valid comparison of treatment groups.

While very rare idiosyncratic side effects, such as liver necrosis, may not be detected reliably, trials with a numerator and denominator allow estimations of rates of such disorders. For example, if no events are found in a population of 30,000 participants, then the rate of such adverse drug reactions (ADR) can be confidently said to be rarer than 1:10,000. However, very rare serious ADR will still require pharmacovigilance and postmarketing for their detection, but not for quantification. One advantage of large streamlined studies is that they allow for prolonged subject followup.

Discussion. Some argue that randomization is important only for intended effects and that it is far less important for unintended effects. Sadly, the history of medicine is littered with instances where unintended drug effects were related to patient characteristics or other biases or confounders that were either unmeasured or unmeasurable, so this distinction has not proved helpful.

The alternatives to large streamlined trials are observational studies, which have major challenges, including confounding by indication and channeling bias. Given that a culture change would be needed before large, streamlined simple trials are routinely implemented, these need to be supported or mandated by drug approval authorities. For this to happen, structures and networks that enable these trials to be mounted efficiently and at reasonable cost need to be organized. A key factor that supports the ability to carry out large streamlined trials is a universal healthcare number. It is this feature that the European countries have in common, enabling them to do these trials. The ethical, political, data-protection, and other debates required to implement the universal number need strong support from the healthcare community.

Patients often comment that it would be better to be part of a system that trialled new drugs properly within the setting of healthcare because, by default, patients are participating in n-of-one trials when they take a new agent that hasn't been studied yet in the real world. A culture of wishing to participate in trials needs to be fostered so that patients expect or request to be included in studies. Informational websites like www.getrandomised.org are examples of a venue to reach out to the public and promote participation in research projects.

3. UTILITY OF METAANALYSIS OF RCT IN EVALUATING ADVERSE EFFECTS OF DRUGS (Andrew Moore, David Henry)

Systematic reviews with metaanalysis of RCT have been used increasingly to answer important questions about the efficacy of medical interventions. When asked whether metaanalyses are useful for safety analyses, 78% of respondents answered "yes" and 22% "no."

Summary analyses can provide precise estimates of treatment effects, while regression analyses can explore variation in effect sizes across different populations. In addition, it is sometimes possible to measure effects of treatment on outcomes that were not primary endpoints of trials, including capture of adverse events. In the absence of large definitive trials, metaanalysis of small trials may provide accurate estimates of risk of uncommon adverse events. This is based on the assumption that summation of low frequency events in multiple small trials is statistically equivalent to the estimate obtained from a single large trial with a similar denominator for exposed individuals. Care needs to be taken that the sample size is large enough even with pooled studies; a rule of thumb is that 200 to 300 events are needed to develop accurate, credible estimates of the number needed to treat to harm. It is important to also look at time to event.

Of great importance is the under-reporting of low frequency events that were not identified, *a priori*, as trial outcomes of interest. For instance, it has been common practice to report "adverse reactions" in trials only after physician assessment of causality. All adverse or beneficial events should be reported, irrespective of their assumed association to drug exposure, and events are more likely in large than in small trials.

Controlled observational studies of adverse drug effects have also been included in metaanalyses. The individual studies tend to be large and designed to assess uncommon outcomes. The main threat to the validity of the summary estimates of effect is residual confounding, which is sometimes related to the indications for therapy. In this situation, metaanalysis may produce a very precise estimate of a confounded relationship. There is potential conflict here between statistical and epidemiological inference. In practice, all available data, both randomized and nonrandomized, tend to be retrieved during evaluation of emergent adverse events.

Discussion. The importance of distinguishing between systematic reviews and metaanalyses was emphasized. Thresholds should be defined for the minimal level of quality that should be present before including each study in pooling. The metaanalysis can only be as good as the quality of the RCT included. The new Cochrane Handbook[2] permits indirect comparisons. The credibility of this approach is still not established. More direct and indirect comparisons are needed to decide if these should be actively encouraged.

Observational studies: Studies aiming to assess global drug safety need to balance health gains in terms of treatment efficacy against health loss and increased drug-related morbidity. RCT are often too small or too short, include the wrong patients, or use the wrong comparator. Observational studies are useful, but careful attention must be paid to size, selection, appropriate comparator, equivalent, risk periods, choosing comprehensive outcomes, followup, and appropriate analysis with adjustment for confounders.

4. POSTMARKETING SURVEILLANCE (Jim Fries)

Should a postmarketing surveillance program be implemented with all deliberate speed, be required by the FDA, and be funded by a pool of sponsors of all drugs in a category, and should all drugs in a class be analyzed identically and concurrently? The result of voting on this question: 78% of respondents voted yes and 22% voted no.

The majority of participants agreed that a formal, carefully designed, large-scale, postmarketing surveillance program is an absolutely essential component of drug safety evaluation and harm reduction. Many critical issues can be addressed realistically only by these approaches, including

delayed or cumulative effects, direct comparison across all drugs in a class, and identification of signals for truly rare adverse effects. Postmarketing surveillance studies must include: more than one defined population, a full protocol with surveillance every 6 months, concurrent control subjects, patient-reported and electronic medical record-reported cross-validated data, predefined hypotheses, full death ascertainment by national death indices (NDI), and close oversight by both a data management advisory council and the FDA. Programs such as the Stanford Toxicity Index have documented the ability to validly address issues such as relative toxicity of specific disease modifying antirheumatic drugs and nonsteroidal antiinflammatory drugs; they have also highlighted issues such as "patient-reported" versus "physician ascribed" data, attribution versus frequency counting, value tradeoffs, duration of followup, definition of specific toxicities, selection biases, and the need for consensus.

A postmarketing surveillance program should be implemented, implementation should be required by the FDA, funding should be provided by a pool of sponsors of all drugs in a category, and all drugs in a class should be analyzed identically and concurrently. The program should make use of current and evolving electronic medical record systems in combination with patient-reported data where the medical record does not contain the data required; and sponsors should be at arm's length from data collection procedures, analyses, and interpretation of data, while fully involved in hypothesis specification, protocol development, and public discussions.

Discussion. Government approval agencies have been much slower to approve the use of patient outcomes for safety than for benefit; yet patient-reported outcomes are at least one essential component of benefit:harm tradeoffs presented to patients by practitioners. Opinions differ on whether treating clinicians should continue to be asked to assess attribution: the issue of the trialist-clinician assigning causality needs to be systematically evaluated. On the one hand, we no longer ask trialists to ascribe benefit; on the other, a study was quoted where the number of events needing analysis was reduced 10-fold when events deemed clearly unrelated were excluded[3]. Other important issues needing study included value tradeoffs, duration of followup, definition of specific toxicities, selection biases, and need for consensus about these issues.

5. WHAT ARE THE UTILITIES OF DRUG REGISTRIES TO DEFINE RISK?

(Claire Bombardier, Alan Silman)

Should well conducted and analyzed register-based observational studies be seen as the gold standard for assessing drug safety in the real world by health professionals and regulators? To this question, 52% of respondents said they are superior in value to other sources, 32% that they are equal in value, and 16% responded that they are inferior in value. Geographically constituted registries ascertaining drug exposure and subsequent morbid events are increasingly used by regulators, pharmaceutical companies, and academia to assess and quantify potential hazards; in the US they are particularly assigned a role in risk management programs. The well described advantages include: knowledge of denominator to calculate rates, accumulation of experience of risk in the real world as opposed to clinical trial patients, and the possibility to undertake longterm followup to detect events requiring prolonged exposures or with a long latency. In addition, registers that capture the entire treatment history may allow study of complex exposures and their resulting effects. There are, however, disadvantages to registries: lack of appropriate in-built comparison group(s), data quality, patients not obligated to follow any protocol, patients subjected to polypharmacy because they have a wide range of comorbidities that would usually exclude them from RCT, loss to followup, and changes in medical care while the registry is being set up.

These disadvantages can be overcome using various methods. Unlike clinical trials, allocation to treatment is not random and is subject to confounding factors that might influence risk in unpredictable directions. Gathering detailed data on potential confounders could allow statistical adjustment through mechanisms such as propensity-modeling to overcome this. Data quality often varies or is unknown because registers frequently utilize data derived from routine clinical practice. There is no *a priori* reason why data quality in a register should not reach appropriate standards, although the size of the enterprise may require resources beyond the scope of the funding body. Real-world patients are not subject to protocol demands; and treatment stops, starts, additions and changes are frequent, are difficult to capture, and add considerable complexity to analysis. In any longitudinal followup study, loss to followup is a concern, especially with increasing duration. Record linkage systems may be a useful substitute for direct data collection, although the quality and depth of the data may not be sufficient. Where possible, efforts should be made to link primary clinical data with administrative data to maximize the advantages of each source of data.

Discussion. There is an increasing acceptance of the concept that prospective minimal data collection is the "standard of care" and will have the major benefit of increasing the quality of care. This raises the question of what is research and what is public safety. The role of patient agreement with prospective analyses is being debated.

There was agreement with the importance of prospectively linking registries with large administrative databases and defining the goals *a priori*. Collaboration between private and publicly funded databases needs to be worked out. Challenges include deciding whether to orient to drug, class of drug, or disease (each with its own confounders), data quality, losses to followup, and how to ensure data integrity.

6. PHARMACOEPIDEMIOLOGIC STUDIES
(Muhammad Mamdani, Ken Saag)

Randomized clinical trials are typically designed and powered to assess drug efficacy in a scientifically rigorous manner and have, therefore, traditionally been the primary accepted source of efficacy information. These trials, however, do not reliably report meaningful information on drug safety because of their insufficient size and duration, patient homogeneity, and lack of ability to assess patient harm due to ethical issues. Many different observational pharmacoepidemiologic study designs can be used to assess drug safety and each has their own set of strengths and limitations. The group was asked whether observational studies of large administrative databases should be used for assessment of therapeutic safety, with 87% responding yes and 13% no.

Cohort studies are generally better than case-control studies as they provide direct estimates of both absolute and relative risks and are able to characterize the temporal nature of adverse events associated with drug therapies. However, since observational pharmacoepidemiologic research does not involve randomization, issues related to selection bias and confounding by indication plague many cohort studies. Advanced statistical techniques can be used to "adjust" for clinical differences between patient groups. However, it is often uncertain whether the level of adjustment is sufficient to provide valid and reliable conclusions.

The group was asked whether observational studies of large administrative databases should utilize a cohort-based design to assess absolute risk estimates of drug safety, with 90% answering yes and 10% answering no. It is thus recommended that where feasible, cohort-based study designs should be preferred over case-control study designs.

It is always important to characterize absolute risks as well as relative risks. What can be done: (a) utilize large national datasets with fewer selection biases; and (b) provide generalizable studies of comparative effectiveness and safety, which complement other approaches to adverse effect detection. What is not possible: to classify disease states, differentiate incident versus prevalent rate, and assess disease severity.

Discussion. Issues suggested for research include: Deciding whether only the "new user" should be included (not prevalent users); Misclassification and "immortal time bias"; Prescriptions versus dispensing; Ascertaining whether all or only selected outcomes need to be validated; Deciding whether a clearing-house of validated outcomes should be established; Deciding whether validation in one source is good for another; Establishing what type of documentation constitutes an adequate validation; Creating a set of minimal guidelines for confounding (How far back should confounding factors be measured and how to minimize; Assure appropriate use of propensity scores, instrumental variables, and risk scores).

7. SIMPLE VERSUS COMPLEX METRIC
(Maarten Boers)

Data on risk and benefit of a drug treatment from trials and observational studies need to be placed in the proper perspective to decide on the value of a treatment. Standardized measures are available to assess benefit. However, benefit and harm have not yet been usefully combined into one scale. The question of whether we should continue efforts to develop a single metric was asked, with 70% responding yes and 30% responding no. Difficulties with comparing benefit and harm include placing a value judgment on scientific facts, trading off short and longterm effects, and assessing multiple benefits and risks concurrently.

The first simple metric, developed by the OMERACT executive, proposes as a first step 3 ranks for both beneficial and harm outcomes: for benefit, these are "none," "substantial," and "(near) remission"; for harm, these are "none," "severe," and "(near) death." Patients are ranked for both benefit and harm and subsequently counted in a 3×3 table. A second simple metric in use is the "Principle of Three," which can be used to summarize all the available evidence from trials and observational studies on a qualitative scale. Three separate 3×3 tables describe the disease, the benefits, and the harms of treatment. In each, the dimensions "seriousness," "duration," and "incidence" are scored on a 4-point scale: 0 = absent or no effect; 1 = low; 2 = medium; 3 = high. Both scores are plotted on a 2-dimensional diagram.

Complex: A third metric, the multicriteria decision analysis, is a complicated technique that provides ways to disaggregate a complex problem allowing multiple criteria, uncertainty around the estimates, and tradeoffs. All the above methods need weights to be attached to categories of benefit, harm, and severity of the disease being treated. GRADE[4] (Grading of Recommendations Assessment, Development, and Evaluation) has attempted to address this issue by placing emphasis on the quality of evidence; GRADE provides nomenclature for a decision based on implicit or explicit weighting of the evidence for the tradeoff (see details below).

Methodology to assess benefit and risk on a single scale and weights to categorize these should be developed further.

Discussion. A simple metric such as the 3×3 table places value judgments on scientific facts, perhaps oversimplifying multiple comparisons and tradeoffs. "Weighting" is less of an issue if categories are kept simple, but weighting does imply a value judgment. This depends upon the patient, so it is argued that the facts should be presented in as simple a fashion as possible and then the patient can decide on the tradeoff depending on the relative values they place on the key benefits versus harms. For policy-makers, information should be aggregated on what most patients would value given the best information.

Research challenges include: The utility of supplying a

single metric measure in individual trial reports; How to incorporate uncertainty, since event rates (evaluated in terms of both risks and benefits) are merely the point estimates; Establishing how to accommodate the fact that measurement of benefit is specific and involved, risk not so much so, and they need to be on the same scale; How to make the decision at the micro versus macro level; How to resolve the issue of rare serious adverse events that have a significant effect on patients' quality of life but involve only a few patients, versus modest benefit (clinical improvement) in a large number of patients, such that the potential benefits in a large number of patients may appear to outweigh the potential harm in a few patients.

8. COMPLEX FRAMEWORKS (Larry Lynd)

There is a Next Steps working group, which has been established by the Pharmaceutical Research and Manufacturers of America (PhRMA) and the FDA involving representatives from industry, regulatory bodies (United States, Canada, The European Medicines Agency), and academics to look at different quantitative methods for benefit-risk analysis rather than at one particular metric. They are specifically looking at multicriteria decision analysis, BRAT, conjoint analysis, and incremental net-benefit/modeling. A number of incremental net-benefit models have been developed that incorporate epidemiologic data with preference weights, including an evaluation of rofecoxib relative to naproxen in RA, and alosetron for the management of irritable bowel syndrome. In the latter, differences have been demonstrated between different patient subgroups. In particular the incremental net benefit is greater in patients with moderate and severe symptoms, as opposed to in those patients with only mild symptoms. In these models, the decision-making approach has been favored over the research approach, under the assumption that these models will serve as decision aids to decision-makers (either regulatory or clinical) in conjunction with expert judgment as opposed to acting as a replacement.

It was asked how this metric incorporates quality adjusted life-years, in addition to the raw number of events. It was suggested that the Simple versus Complex methods need to be looked at separately for regulatory purposes.

9. GRADING OF RECOMMENDATIONS ASSESSMENT, DEVELOPMENT, AND EVALUATION (GRADE) WORKING GROUP (Gordon Guyatt)

The GRADE working group has developed a system that classifies the quality of evidence into 4 levels: high, moderate, low, and very low depending on the study designs, potential weaknesses (risk of bias, imprecision, inconsistency, indirectness, and publication biases), and special strengths (large effect, dose-response). The GRADE system offers 2 levels of recommendations: strong and weak. When an intervention's benefits clearly outweigh its risks and burden, or clearly do not, strong recommendations are warranted. On the other hand, when the tradeoff between benefits and risks is less certain, either because of low-quality evidence or because high-quality evidence suggests benefits and risks are closely balanced, weak recommendations become appropriate. There are 4 factors that bear on the strength of a recommendation. The clearer the tradeoff between the desirable or the undesirable consequences of implementing an intervention, the more likely a strong recommendation. High-quality evidence is more likely to result in a strong recommendation than is low-quality evidence. Wide variability in patient values and preferences across the population of interest will make a strong recommendation less likely. Finally, if the intervention is associated with large use of resources (i.e., high cost), a strong recommendation is less likely. Its combination of relative simplicity with conceptual and methodological rigor (Table 1) has led a large number of groups (including the World Health Organization, the American College of Physicians, the National Institute for Clinical Excellence, the Cochrane Collaboration, UpToDate Inc., and BMJ Clinical Evidence) to adopt the GRADE approach.

The group was asked whether the regulatory process should adopt GRADE, or some other system that provides a structure that ensures optimal transparency and decision-making on the basis of that evidence: 79% of the respondents answered yes, and 21% answered no.

Discussion. There was some debate around the merits of randomized trials automatically beginning high, although they may be rated down, and longitudinal observational studies starting low, although they may be rated up. The way in which values are incorporated for different indications needs to be clear, e.g., clopidogrel versus ASA for peripheral vascular disease versus cardiovascular disease.

10. OTHER MODELS OF RISK: NONTREATMENT (Randall Stevens)

The safety and tolerability of a drug or an investigational

Table 1. GRADE summary of findings. Alendronate tradeoffs with randomized controlled trials for benefit and longitudinal observational studies for harms.

Outcome	Difference	Quality	Number Needed to Treat
Vertebral effects	–7/1000	Moderate	142
Hip effects	–10/1000	High	100
Esophageal ulcer	1/2000	Very low	2000
Withdrawals	0	Moderate	∞
Osteonecrosis	1/20,000	Very low	20,000
Atrial fibrillation	5/1000	Low	200
Inconvenience	1000/1000	High	1
Cost		High	—

product is assessed on the type, incidence, and severity of the adverse events (AE). An assessment of the AE must take into account the intervention, the risk, cost, and duration of that intervention, and whether the AE occurs immediately or develops over a number of months or years. Alternatives to not treating patients with the drug in question are often not taken into account. In the context of safety evaluations of a therapy, the risk for the patient is contextual to the diseases the patient has, the risk of no treatment versus alternative therapies, and the value system the patient places on the type and risk of available treatment or nontreatment options. In addition, whether the therapy is first-line, second-line, or salvage therapy must be determined in order to place the AE into context. The need then is to define risk in terms of acute, subacute, and chronic injuries, and if they are manageable, treatable or not, and whether the risks can be mitigated. The participants were asked whether the rheumatological community should develop a simple one-page database form for tracking patients receiving a cell-based therapy for an autoimmune/rheumatological disorder. In response, 93% said yes and only 7% answered no.

11. OTHER MODELS OF RISK: SAFETY OF CELLULAR THERAPEUTICS (Alan Tyndall)

Hematopoietic stem cell transplantation (HSCT) for severe autoimmune diseases is being performed in the context of prospective controlled trials in Europe, North America, and other regions. These transplants occur in established specialized transplant units and all data, including acute and longterm toxicity, are collected routinely as part of clinical trial good clinical practice and accreditation guidelines. Multipotent mesenchymal stromal cells (MSC) are under consideration for the treatment of autoimmune disease based on their *in vitro* antiproliferative properties, efficacy in animal models, apparent low acute toxicity, and the early positive anecdotal outcomes in human acute graft versus host disease. Based on these experiences, it is recommended that the rheumatological community develop a simple one-page database form for tracking patients receiving a cellular therapy for an autoimmune/rheumatological disorder. This database is compatible with other such databases and managed within the EULAR Standing Committee for Clinical Affairs and American College of Rheumatology Quality of Life committees.

12. CONDITIONAL APPROVALS (Andreas Lapaucis)

The question was asked whether conditional approval should be applied to all new drug approvals to gain additional safety data. The final votes were split, with 48% of respondents answering "always" or "frequently," 46% answering "sometimes," and only 6% answering "rarely" or "never."

The discussion revolved around the need to carefully define what is meant by the term "registry." The definition could include: (a) a group of patients about whom primary data are collected before and after starting the drug; (b) primary data collection when starting the drug, with followup using administrative data; or (c) a registry identifying and following patients entirely through administrative data. In addition to this, the purpose of a registry needs to be carefully defined. This can include determining real-world compliance/concordance, evaluating concomitant drug utilization, assessing real-world effectiveness, assessing real-world harm, detecting rare side effects that might not be detected in randomized trials, and assessing the characteristics of patients who receive the drug in the real world.

It should be noted that evidence from a registry about effectiveness will often be unconvincing, because of concern that confounders have not been adequately dealt with when comparing the registry patients with those not taking the drug. On the other hand, registries can be a method of regulating uptake of a drug, by requiring patients to meet strict criteria before being allowed into the registry. However, this is an expensive method of regulating drug use in the long term. Although prospective registries may detect important but relatively rare side effects, careful consideration should be given to whether these harms could be detected much more cheaply with administrative data.

Discussion. There was much discussion about the applicability and utility of conditional approvals. In the United States, a variant of conditional approvals exists under the condition that surrogate markers are used for approval or approval is predicated on attenuated data sets due to the importance of the therapeutic benefit newly offered. Prior to 2007, the US Secretary of Health and Human Services, a cabinet level position, had to determine with appropriate input whether a drug could or should be removed from marketing. The FDA Amendment Act has now implemented new laws that allow the FDA to remove products from the market without as much political influence as was required prior to 2007. Thus, it should be easier for the FDA to make such decisions, allowing conditional approvals to be more commonly applied.

Elsewhere in the world consideration has also been given to conditional reimbursement. Thus drugs could be reimbursed and made available until their influence upon clinically important outcomes has been assessed in routine practice.

CONCLUSIONS/FUTURE RESEARCH

Participants were asked to vote on whether safety assessment should be taken out of the hands of the industry. The split was fairly even, with 55% voting "yes" and 45% voting "no." When asked who should take over safety assessment, 13% said the government, 17% said academia, and 71% said a combination of government and academia.

Overall, it was decided that future research be conducted in 4 main areas:

1. Research on value tradeoffs and weighting is required.

2. Postmarketing surveillance studies need to be expanded.
3. A single metric to assess benefit and risk that takes into account both generic and specific studies needs to be developed. This metric must be interpretable for both physicians and patients. The GRADE approach should be one of the options developed.
4. Registries in Rheumatology that link to national databases need to be developed following well developed methodological guidelines.

REFERENCES

1. Simon LS, Strand CV, Boers M, et al. Observations from the OMERACT Drug Safety Summit, May 2008. J Rheumatol 2009;36:2110-4.
2. Higgins PT, Green S, editors. Cochrane handbook for systematic reviews of interventions. New York: Wiley; 2008.
3. Fries JF, Spitz PW, Williams CA, Bloch DA, Singh G, Hubert HB. A toxicity index for comparison of side effects among different drugs. Arthritis Rheum 1990;33:121-30.
4. Atkins D, Best D, Briss P, et al, and the GRADE Working Group. Grading quality of evidence and strength of recommendations. BMJ 2004;328:1490-9.