Evaluating the Risks of Clinical Research
Annette Rid; Ezekiel J. Emanuel; David Wendler

http://jama.ama-assn.org/cgi/content/full/304/13/1472

Correction
Contact me if this article is corrected.

Citations
Contact me when this article is cited.

Topic collections
Medical Practice; Medical Ethics; Dermatology; Dermatologic Disorders; Quality of Care; Patient Safety/ Medical Error; Statistics and Research Methods; Hypersensitivity; Immunology; Allergy
Contact me when new articles are published in these topic areas.

Related Articles published in the same issue
What Physician-Investigators Owe Patients Who Participate in Research

Subscribe
http://jama.com/subscribe

Permissions
permissions@ama-assn.org
http://pubs.ama-assn.org/misc/permissions.dtl

Email Alerts
http://jamaarchives.com/alerts

Reprints/E-prints
reprints@ama-assn.org
Evaluating the Risks of Clinical Research

Annette Rid, MD
Ezekiel J. Emanuel, MD, PhD
David Wendler, PhD

The ethical appropriateness of clinical research depends on protecting participants from excessive risks. Yet no systematic framework has been developed to assess research risks, and as a result, investigators, funders, and review boards rely only on their intuitive judgments. Because intuitive judgments of risk are subject to well-documented cognitive biases, this approach raises concern that research participants are not being adequately protected. To address this situation, we delineate a method called the systematic evaluation of research risks (SERR), which evaluates the risks of research interventions by comparing these interventions with the risks of comparator activities that have been deemed acceptable. This method involves a 4-step process: (1) identify the potential harms posed by the proposed research intervention; (2) categorize the magnitude of the potential harms into 1 of 7 harm levels on a harm scale; (3) quantify or estimate the likelihood of each potential harm; and (4) compare the likelihood of each potential harm from the research intervention with the likelihood of harms of the same magnitude occurring as a result of an appropriate comparator activity. By explicitly delineating, quantifying, and comparing the risks of research interventions with the risks posed by appropriate comparator activities, SERR offers a way to minimize the influence of cognitive biases on the evaluation of research risks and thereby better protect research participants from excessive risks.

RISK COMPARISONS

Many regulations evaluate the risks of research interventions by comparing them with the risks of specified comparator activities. A finding that the risks of research interventions do not exceed the risks of the comparator activities is regarded as evidence that the research is acceptable and, in some cases, may be subject to fewer restrictions. US regulations direct IRBs to compare the risks of research interventions with the risks “ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests.”10 Under these regulations, a finding that the research risks do not exceed the risks ordinarily encountered in daily life implies that the study may enroll healthy children and may be approved using an expedited review process. Guidelines from the Council for International Organizations of Medical Sciences allow research that does not offer the potential for clinical benefit when the risks do not exceed the “risks attached to routine medical and psychological examinations.”11 Similarly, some have argued that it may be acceptable to enroll children in research that does not offer

See also p 1491.
the potential for clinical benefit when the risks do not exceed the risks of charitable activities. Others have argued that the risks of firefighting or donating a kidney might provide a threshold for determining when competent adults may be enrolled in research without the potential for clinical benefit, on the grounds that society deems it acceptable for individuals to participate in these activities for the benefit of others.

Comparing the risks of research interventions with the risks of other activities provides a context for evaluating research risks. When the comparator activity is sufficiently similar and acceptable, these comparisons allow review committees to appeal to widely endorsed risk evaluations made outside the research context. This approach has the potential to make the evaluation of research risks less vulnerable to errors in intuitive judgment and, thus, more likely to protect research participants. This approach requires identification of appropriate comparator activities and a systematic method for comparing the risks of research with the risks of the comparator activities. We address the latter task by proposing a systematic method for comparing the risks of research interventions with the risks of comparator activities.

**Comparing Likelihoods**

Risk can be analyzed as a function of 2 components: the likelihood that a harm will occur; and the severity or magnitude of the harm should it occur. One way to make the evaluation of research risks more systematic is to independently compare the likelihoods and harms to harms. Comparing likelihoods is relatively straightforward when they are of the same type. For example, it is fairly easy to compare uncomplicated bone fractures that occur during different activities. However, research interventions pose harms frequently not present in other activities. To compare these harms with the harms of comparator activities—whether phlebitis is less severe, equivalent to, or worse than fracturing a bone—the harms first need to be categorized by magnitude. This approach necessitates a scale that divides the continuum of all possible research harms into discrete levels.

There is no objectively correct number of magnitudes into which a given continuum should be divided. Dividing the continuum of temperature from zero to boiling into 100 units is not more or less objectively accurate than dividing it into 212 units. Rather, proposed scales should be evaluated based on how well they serve the goals for which they are created. Does the proposed scale include enough categories to make the needed distinctions, without being too complex to implement? Standard measures of harms to health typically use 5 to 8 levels. Adverse events in cancer trials are classified in 5 levels; and the Health and Activity Limitation Index distinguishes 6 levels of limitations due to ill health. These approaches are supported by research indicating that 5 to 7 categories are likely to maximize reliability and validity without being too complex to use. Although measures of ill health provide a useful starting point, they are limited to disease and disability. In contrast, protection of research participants should take into account all the potential harms the participants face, including psychological, social, and economic harms.

We developed a preliminary scale for research with 5 harm levels and illustrative examples for each. This scale was then systematically refined in 5 steps. First, the initial proposal was presented at 3 academic meetings and evaluated in 2 structured focus groups, yielding a scale with 6 harm levels and revised illustrative examples. Second, the scale was edited based on the input of 5 experts in clinical research and an expert in risk assessment. Third, the scale was discussed with 43 international experts in clinical research, philosophy, research ethics, risk assessment, and patient advocacy, which resulted in a harm scale with 7 harm categories and further revision to the illustrative examples.

Once the 7-category scale was formed, it underwent the fourth step—3 rounds of revisions based on the input of 3 clinicians, 8 bioethicists/research ethicists, and 2 IRB chairpersons. Fifth, the scale was presented and critiqued at 7 meetings, including academic meetings of clinicians and individuals involved in clinical research, and educational meetings involving students, leading to the final harm scale with illustrative examples (Table).

Some harms, such as excruciating pain, are serious no matter how long the harms last. Other harms, such as difficulty hearing, typically are serious only if they are extended in time. Among the many factors that influence the magnitude of particular harms, 7 emerged over the course of the refinement process as especially relevant: (1) the experience, such as pain, associated with the harm; (2) the burden of efforts, including treatment, to mitigate the harm; (3) the effects on an individual’s ability to perform the activities of daily life; (4) the effects on an individual’s ability to pursue life goals; (5) the duration of the harm; (6) the extent to which an individual can adapt to the new circumstances; and (7) the burden imposed by the process of adaptation (Table).

**SYSTEMATIC EVALUATION OF RESEARCH RISKS**

The 4-step process of the systematic evaluation of research risks (SERR) provides a way to systematically compare the risks of research interventions with the risks of comparator activities by independently comparing the 2 components of risk: likelihood and magnitude of harm (BOX).
Testing SERR Using the Risks of Daily Life Standard

Although SERR was the result of an extensive development and refinement process, evaluating its usefulness requires assessment of how well it addresses the limitations of current practice. Does SERR incorporate empirical data, minimize the influence of cognitive biases, reduce variation, delineate a threshold for acceptable risks, and offer a transparent method that can be used by designated review committees?

SERR does not mandate a specific comparator activity. Hence, it can be used to apply different regulatory standards once it has been determined which specific activities will be used to implement the standard in question. For example, the Council for Interna-

<table>
<thead>
<tr>
<th>Examples of Harms by Magnitude</th>
<th>Examples and Details of Harms</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Negligible</strong></td>
<td><strong>Negligible</strong></td>
</tr>
<tr>
<td>Mild nausea</td>
<td>Discomfort; can interfere with ability to pursue some minor life goals (eg, eat)</td>
</tr>
<tr>
<td>Skin bruise or abrasion</td>
<td>Mild pain</td>
</tr>
<tr>
<td><strong>Small</strong></td>
<td><strong>Small</strong></td>
</tr>
<tr>
<td>Headache</td>
<td>Moderate pain, inability to pursue some minor (eg, 1 day hiking) and some major (eg, attend school) life goals</td>
</tr>
<tr>
<td>Common cold</td>
<td>Discomfort, inability to pursue some minor (eg, visit museum) and some major (eg, work) life goals</td>
</tr>
<tr>
<td><strong>Moderate</strong></td>
<td><strong>Moderate</strong></td>
</tr>
<tr>
<td>Uncomplicated bone fracture</td>
<td>Moderate pain, inability to pursue some minor life goals (eg, play sports)</td>
</tr>
<tr>
<td>Moderate insomnia for 1 month</td>
<td>Annoying experience, inability to pursue some minor (eg, meet friends) and some major (eg, work) life goals</td>
</tr>
<tr>
<td>Intensive care for several weeks (assuming no sequela)</td>
<td>Often intense pain and physical exhaustion, inability to perform activities of daily life and to pursue essentially all minor and major life goals</td>
</tr>
<tr>
<td><strong>Major</strong></td>
<td><strong>Major</strong></td>
</tr>
<tr>
<td>Psychotic episode</td>
<td>Terrifying distortions of reality, changes in personality that undermine relationships, precludes performance of daily life activities and many minor and major life goals</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>Daily episodes of serious pain and permanent stiffness, unable to pursue some minor (eg, vacation) and some major (eg, work) life goals, sometimes unable to perform some activities of daily life</td>
</tr>
<tr>
<td>Loss of finger</td>
<td>Destabilizes hand, interferes with many activities of daily life, interferes with some minor and major life goals, requires adaptation, distressing transition period</td>
</tr>
<tr>
<td><strong>Severe</strong></td>
<td><strong>Severe</strong></td>
</tr>
<tr>
<td>Major depression</td>
<td>Depressive episodes with hopelessness/worthlessness, loss of interest in usual activities, insomnia, and eating; can preclude performance of some daily life activities and some minor and major life goals; often baseline anxiety and low mood</td>
</tr>
<tr>
<td>Paraplegia</td>
<td>Inability to perform some activities of daily life, inability to pursue many minor (eg, hiking) and some major (eg, having children) life goals, often distressing transition period</td>
</tr>
<tr>
<td>Catastrophic</td>
<td>Precludes performance of daily life activities and essentially all minor and major life goals, adaptation impossible, distressing transition period</td>
</tr>
</tbody>
</table>

**Deaths**

<table>
<thead>
<tr>
<th>Magnitude</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death</td>
<td>Precludes performance of daily life activities and essentially all minor and major life goals, adaptation impossible, distressing transition period</td>
</tr>
</tbody>
</table>

**Table. Magnitude of Harms Scale With Illustrative Examples**

---

3 Important factors that influence the magnitude of a harm include associated experience (no sensory impact, nuisance, uncomfortable, distressing, suffering); burden of efforts to mitigate condition (low/moderate/high; weeks/months/permanent); inability to perform activities of daily life (partial/complete); inability to realize life goals (minor/major goals, some goals in one category/some goals in both categories/all goals in one or both categories); duration (minutes/hours/days/weeks to months/years/permanent, intermittent/continuous); potential to adapt to new (residual) condition (minor/moderate/major adaptation, impossible to adapt); and burden of adaptation period (low/moderate/high). The examples were chosen based on input from 43 international experts in clinical research, research ethics, and risk assessment. The examples have an illustrative function to show how the harm scale might be applied. Factors not mentioned in the description of an example are considered not relevant. It is assumed that the given harms occur in otherwise healthy, normal, average individuals (adults), which implies that the selected examples might fall into a different category on the harm scale in individuals who are not healthy, normal, or adults. No examples of economic or social harms are given due to their strong context dependence.
EVALUATING THE RISKS OF CLINICAL RESEARCH

Box. The 4-Step Process of Systematic Evaluation of Research Risks

1. Identify the potential harms posed by the research intervention.
2. Categorize the magnitude of each potential harm using the harm scale.
3. Quantify or estimate the likelihood of each potential harm.
4. Compare the likelihood of each potential harm from the research intervention with the likelihood of potential harms of the same magnitude occurring in an appropriate comparator activity. If the likelihoods of the potential research harms are all comparable with the likelihoods of potential harms of the same magnitude in the comparator activity, then the risks of the research intervention do not exceed the risks of the comparator activity. Depending on the regulations in question, this finding implies that the risks of the research are acceptable and, in some cases, the research may be subject to fewer restrictions.

©2010 American Medical Association. All rights reserved.
of observations on which each potential harm is based. Fewer than 100 observations was considered weak evidence, 100 to 1000 observations was considered moderate evidence, and more than 1000 observations was considered strong evidence. Expert opinion is treated by definition as weak evidence. Four factors were then evaluated: strength of the methodology, generalizability of the study population, relevance of the study environment, and timeliness of the clinical or diagnostic practice. If these factors undermined the strength of the data, the preliminary strength determination was reduced, thus yielding the final strength determination.

EXAMPLE: ALLERGY SKIN TESTING
How would SERR evaluate the risks of epicutaneous allergy skin testing using the present interpretation of the risks of daily life standard? The literature suggests that allergy skin testing poses 6 potential harms (step 1, Box) in average adults: (1) transient pain from the skin pricks; (2) local allergic reaction with itching for 5 to 15 minutes; (3) mild systemic allergic reaction with self-limiting hay fever symptoms or hives requiring antihistamines; (4) moderate systemic allergic reaction with asthmatic symptoms or low blood pressure, typically requiring epinephrine treatment; (5) severe systemic allergic reaction, requiring intubation; and (6) death.

Based on the input of 3 physicians, 1 nurse, and 1 philosopher, the transient mild pain and local allergic reaction were categorized as negligible harms on the harm scale (step 2, Box). The mild allergic reaction qualifies as a small harm. The moderate systemic allergic reaction constitutes a moderate harm, and a significant harm if intubation is required. Death is catastrophic. Using riding in a car and playing sports as the primary daily life comparators yields the following potential harms from daily life with these magnitudes: a bruise (negligible), a common cold (small), an uncomplicated bone fracture (moderate), a complete ligament tear of knee (significant), and death (catastrophic).

Based on the literature and expert opinion, the likelihood estimates (step 3, Box) for the 6 harms of allergy skin testing are provided in Figure 1.

An effective way to compare likelihoods of potential harms that are of comparable magnitude (step 4, Box) is to place them on the same log-linear graph. The resulting comparison reveals that the potential harms from epicutaneous allergy skin testing are not greater in number, and are all less likely to occur than comparable harms from the activities of daily life (Figure 1). This suggests that allergy skin testing qualifies as minimal risk under the present interpretation of the risks of daily life standard.

EXAMPLE: LIVER BIOPSY
To further evaluate SERR, consider how it would categorize the risks of percutaneous liver biopsy. The literature suggests that liver biopsy poses 18 potential harms (step 1, Box) in the average adult: (1) transient mild pain during administration of local anesthesia; (2) anxiety in anticipation; (3) immediate postprocedure pain of moderate intensity for 1 to 2 hours, requiring analgetics; (4) postprocedure pain of mild intensity for several days, self-limiting; (5) superficial kidney puncture with no symptoms or blood in urine; (6) subcutaneous emphysema, self-resolving; (7) major hemorrhage with hypotension or decrease in hemoglobin concentration greater than 2 g/dL, requiring transfusion; (8) pleural effusion, requiring aspiration; (9) hematotherax, requiring aspiration; (10) pneumothorax, requiring no treat-
ment or drainage and analgesics for 2 to 5 days; (11) hemobilia, involving colic and/or black stool and/or jaundice for 1 week; (12) sepsis, requiring antibiotics; (13) major hemorrhage, requiring interventional radiography or surgery; (14) hemobilia, requiring interventional radiography or surgery; (15) gallbladder perforation with severe pain, requiring surgery; (16) colon perforation with severe pain, requiring surgery; (17) sepsis, requiring intensive care; and (18) death.

Using the harm scale (step 2, Box), and input from 3 physicians, 1 nurse, and 1 philosopher, the magnitude of these potential harms was categorized as follows: (1) negligible, (2-6) small, (7-12) moderate, (13-17) significant, and (18) catastrophic. Potential harms of comparable magnitude from the activities of daily life are previously listed.

Based on the literature and expert opinion, the likelihood estimates (step 3, Box) for the harms of percutaneous liver biopsy are provided in Figure 2. The likelihood estimates for potential harms in daily life are previously listed, plus a 0.03 per 100 000 risk of paraplegia.

Comparing the likelihoods of the potential harms with the likelihoods of potential harms in daily life (step 4, Box) reveals that liver biopsy poses a number of serious harms, such as gallbladder and colon perforation and death, that are more likely than comparable harms in daily life (Figure 2). Therefore, under the present interpretation of the risks of daily life standard, percutaneous liver biopsy poses greater than minimal risk.

TEST RESULTS: ADVANTAGES OF USING SERR
Application to epicutaneous allergy skin testing and percutaneous liver biopsy suggests that SERR has the potential to significantly improve the evaluation of research risks by addressing the 6 concerns posed by current reliance on intuition alone. First, SERR evaluates the risks of research interventions based on the empirical data. This should increase the accuracy of risk judgments.

Second, SERR reduces the influence of cognitive biases by requiring reviewers to explicitly identify and compare risks. For example, by comparing the risks of research interventions with the risks of familiar comparator activities, SERR counters the tendency to regard unfamiliar activities as necessarily more risky.

Third, by providing a common method, SERR promotes consistency in evaluation across interventions, studies, and committees. SERR also provides the means to identify sources of disagreement and consider strategies for addressing them. Disagreement about the magnitude of a harm points to the need for conceptual analysis on the nature of the harm or better understanding of its consequences. Disagreement about likelihoods suggests the need for better data or determination of how to proceed, given uncertainty or the absence of relevant data.

Fourth, by comparing the risks of research interventions with the risks of comparator activities, SERR helps to delineate a threshold for acceptable risks based on the assumption that absent a reason to think otherwise, evaluations of risks should be consistent across similar activities in different realms of life. To ensure a proper threshold, analysis will be needed to identify appropriate comparator activities for clinical research.

Fifth, SERR provides a transparent method for evaluating research risks. For example, review committees could make the data and graphs they use to evaluate research risks publicly available on a Web site.

Sixth, data suggest that IRBs in the United States have as little as 8 minutes to review new protocols, a situation that is likely to be similar in other countries. Systematic risk evaluations are not possible in that time frame.
addition, requiring countless IRBs to repeat the same evaluations for common research interventions represents an enormous waste of resources. Establishing review committees with the requisite expertise and representation to implement SERR would locate the vital responsibility of evaluating research risks in meetings dedicated to this task. IRBs could then focus on whether local circumstances provide reason to alter the default risk judgments made by the designated committee(s).

SERR potentially offers these advantages while still retaining the critical role of normative judgment. Review committees must use their judgment to categorize the potential harms of research procedures by magnitude, to identify comparator activities that are appropriate and relevantly similar to research, and to evaluate whether the default risk judgments apply in the local circumstances.

**POTENTIAL LIMITATIONS**

SERR raises several potential limitations. First, SERR does not provide criteria for determining whether the comparator activities are acceptable and relevantly similar. SERR is intended as a method to systematically compare the risks of research with the risks of comparator activities. Absent a broadly recognized account of acceptable risk, complementary conceptual analysis will be needed to determine which comparator activities are appropriate. Because SERR does not specify the comparator activities, it can be used to implement the different standards prescribed by governmental regulations.

A second potential limitation is that the risks posed by some activities of daily life are inappropriate comparators for evaluating the risks of research interventions. Our interpretation of the risks of daily life standard appeals to the risks of activities in daily life that seem acceptable, even in contexts in which the participants do not realize personal benefit. Although many individuals enjoy sports and driving, these activities can be acceptable even for individuals who do not enjoy them in charitable contexts. Thus, while the examples of epicutaneous allergy skin testing and percutaneous liver biopsy are included in this study to evaluate the usefulness of SERR, the activities used to implement the risks of daily life standard seem reasonable for evaluating research risks.

A third limitation might be that SERR relies on risk data, but such data are never fully complete. Careful consideration of the available data, including consideration of its shortcomings, seems preferable to ignoring relevant data and making judgments based on intuition alone. In addition, clinical research uses many interventions, such as magnetic resonance imaging, glucose tolerance tests, and lumbar punctures, for which considerable empirical data are available.

A fourth potential limitation is that SERR is too complex. Whether a method is too complex depends on the importance of the task and the quality of the alternatives. The importance of protecting research participants and the absence of systematic alternatives suggest that SERR is worth pursuing. Moreover, SERR is intended to be used by designated review committees. Future testing of SERR will be needed to assess its feasibility when used by committees trained in its use. The present analysis reveals that SERR provides a systematic method to evaluate the risks of research interventions based on the empirical data and in comparison to the risks of comparator activities. SERR thus has the potential to minimize the influence of cognitive biases and better protect research participants from excessive risks.

**Author Contributions:** Dr Wendler had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design: Rid, Emanuel, Wendler. Acquisition of data: Rid, Wendler. Analysis and interpretation of data: Rid, Wendler. Drafting of the manuscript: Rid, Wendler. Critical revision of the manuscript for important intellectual content: Rid, Emanuel, Wendler. Study supervision: Wendler. Financial Disclosures: None reported.

**Funding/Support:** This study was funded by the Department of Bioethics at the National Institutes of Health (NIH) Clinical Center. Dr Rid reports receipt of support by a grant from the Swiss National Science Foundation.

**Role of the Sponsors:** The NIH had no role in the analysis of the data, writing of the manuscript, or decision to submit for publication.

**Disclaimer:** The opinions expressed in this article are the authors’ own. They do not represent any position or policy of the NIH, the US Public Health Service, or the US Department of Health and Human Services.

**Additional Contributions:** We thank the participants of the Systematic Evaluation of Research Risk (SERR) workshop and members of the Department of Bioethics for their input on previous versions of SERR. None received compensation; participants in the workshop received travel reimbursement and a per diem. We thank Jon Jay, JD, Patti Zettler, JD, and Ari Hoffman, MD, NIH, for their research assistance. Mr Jay received compensation for his contribution to this article; Dr Hoffman and Ms Zettler did not. We thank Calman Prussin, MD, and Theo Heller, MD, NIH, for discussion on the risks of allergy skin testing and liver biopsy, respectively; Susan Hilsenbeck, PhD, Duncan Cancer Center, and Smith Breast Center at Baylor College of Medicine, for designing the graphs; Bob Goodin, PhD, NIH, Reidar Lie, MD, PhD, and Franklin Miller, PhD, Department of Bioethics, NIH Clinical Center, for their comments on previous drafts of the manuscript; Eric Meslin, PhD, Indiana University Center for Bioethics, Robert Nelson, MD, PhD, Office of the Commissioner FDA, Steven Joffe, MD, MPH, Dana Farber Cancer Institute, and Pearl O’Rourke MD, Partners Healthcare Systems, and Harvard Medical School, for input on the harm scale. Neal Dickert, MD, PhD, Division of Cardiology, Emory University School of Medicine, Emily Largent, BSN, Department of Bioethics, NIH Clinical Center, and Amy Agrawal, MD, Sunrise Medical Group, Rockville MD assisted with the harm scale. None of these individuals received compensation in association with their contributions to this article. We thank the Foundation for NIH for help with organizing the SERR workshop and obtaining funding, and Pfizer and Western Institutional Review Board for financial support of the workshop.

**REFERENCES**

29. Indian Council of Medical Research. Ethical Guidelines for Biomedical Research on Human Participants. New Delhi, India: Director-General Indian Council of Medical Research; 2006.