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.

JAMA. 2010;304(13):1472-1479 www.jama.com

CLINICAL RESEARCH IS JUSTIFIED only when participants are protected from excessive risks. Yet it is often unclear whether the risks of research interventions are acceptable or excessive. Because no systematic framework exists for assessing research risks, investigators, funders, and institutional review boards (IRBs) currently rely on their intuitive judgment to make these determinations.

Although intuition plays an important role in evaluating risks, use of intuition alone raises 6 concerns.

First, intuition alone fails to take into account relevant empirical data. Hence, the resulting judgments are less likely to reflect the actual risks faced by research participants. Second, intuitive judgments of risk are subject to well-documented cognitive biases.[1-4] For example, intuitive judgments are influenced by whether the activity being evaluated is familiar to the rater.[1-4] Familiar interventions are thus more likely to be assessed as low risk, even when they pose greater risks than unfamiliar interventions. Third, intuitive judgments about which research risks are acceptable vary widely,[5,6] raising concern that the extent to which participants are being protected might vary from site to site and from committee to committee.

Fourth, intuitive judgments fail to delineate a threshold for which research risks are acceptable. Fifth, intuitive judgments lack transparency—they do not allow others to understand why IRBs categorize some interventions as low risk and others as high risk. Sixth, reliance on intuitive judgments is inefficient, leading to a system in which the risks of research interventions are evaluated repeatedly by thousands of IRBs,[7] often during meetings that do not allow sufficient time for careful evaluation.[8]

See also p 1491.

To address these concerns, we propose a systematic framework for evaluating the risks of research interventions.[9] Designated review committees (eg, regional or national ones) could use this framework to establish default determinations for the risks of research interventions. Investigators, funders, and IRBs could then focus on whether the default determinations accurately reflect local circumstances.

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 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 examination."[11] Similarly, some have argued that it may be acceptable to enroll children in research that does not offer

Author Affiliations: Department of Bioethics, NIH Clinical Center, Bethesda, Maryland (Drs Rid, Emanuel, and Wendler); and Institute of Biomedical Ethics, University of Zurich, Zurich, Switzerland (Dr Rid).
Corresponding Author: David Wendler, PhD, Department of Bioethics, NIH Clinical Center, Bldg 10, Room 1C118, Bethesda, MD 20892 (dwendler@nih.gov).

 ©2010 American Medical Association. All rights reserved.

Downloaded From: by a UNIVERSITE DE LAUSANNE User on 01/23/2018

the potential for clinical benefit when the risks do not exceed the risks of charitable activities.[12,13] Others have argued that the risks of firefighting[14] or donating a kidney[15] 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 these 2 components: likelihoods to likelihoods and harms to harms.[16]

In principle, comparing 2 likelihoods—whether 1 in 2500 exceeds 1 in 25 000—is straightforward. In practice, likelihood comparisons pose 2 challenges. First, making likelihood comparisons often requires judgment of the quality of the supporting data. Are the data sufficient to make confident judgments? If not, what judgment should be rendered? Second, determining whether nonidentical likelihoods should be treated as normatively equivalent requires judgment. Is a 25 per 100 000 chance of sustaining a bone fracture normatively equivalent to a 20 per 100 000 chance?

Comparing Magnitudes

Comparing 2 harms 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.[17-27] Adverse events in cancer trials are classified in 5 levels,[27] and the Health and Activity Limitation Index distinguishes 6 levels of limitations due to ill health.[20] These approaches are supported by research indicating that 5 to 7 categories are likely to maximize reliability and validity[28,29] without being too complex to use.[30] 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.[31,32]

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).

©2010 American Medical Association. All rights reserved.

Downloaded From: by a UNIVERSITE DE LAUSANNE User on 01/23/2018

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. Magnitude of Harms Scale With Illustrative Examples[a]

Examples of Harms by Magnitude	Examples and Details of Harms		
	Effect/Disability	Treatment	Duration
Negligible			
Mild nausea	Discomfort; can interfere with ability to pursue some minor life goals (eg, eat)	May require medication	Minutes to several hours
Skin bruise or abrasion	Mild pain	Can require cleaning and coverage	Bruise or abrasion pain, minutes to several hours; healing, ≤10 days
Small			
Headache	Moderate pain, inability to pursue some minor (eg, 1 day hiking) and some major (eg, attend school) life goals	May require medication, rest, or both	Hours
Common cold	Discomfort, inability to pursue some minor (eg, visit museum) and some major (eg, work) life goals	May require medication, rest, or both	Several days
Moderate			
Uncomplicated bone fracture	Moderate pain, inability to pursue some minor life goals (eg, play sports)	Requires some medication and wearing a cast	Fracture pain, hours; recovery, weeks to months
Moderate insomnia for 1 month	Annoying experience, inability to pursue some minor (eg, meet friends) and some major (eg, work) life goals	Can require lifestyle changes and medication	Weeks intermittently
Significant			
Ligament tear of knee with permanent instability	Moderate pain that interferes with pursuing some minor life goals (eg, exercise); permanent instability precludes vigorous exercise and requires adaptation (eg, seek new types of exercise)	Requires surgery and rehabilitation	Tear, hours to days; rehabilitation time following surgery, weeks to months
Intensive care for several weeks (assuming no sequelae)	Often intense pain and physical exhaustion, inability to perform activities of daily life and to pursue essentially all minor and major life goals		Weeks
Major			
Psychotic episode	Terrifying distortions of reality, changes in personality that undermine relationships, precludes performance of daily life activities and many minor and major life goals	Requires medication, can require adaptation of some major life goals (eg, work)	Weeks to a month
Rheumatoid arthritis	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	Requires aggressive medication, physiotherapy, requires major adaptation	Years
Loss of finger	Destabilizes hand, interferes with many activities of daily life, interferes with some minor and major life goals, requires adaptation, distressing transition period	None	Permanent
Severe			
Major depression	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	Requires medication; requires adaptation of some major life goals (eg, relationships)	Decades
Paraplegia	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	Requires daily support and close clinical observation; requires major adaptation	Permanent
Catastrophic			
Severe dementia	Precludes performance of daily life activities and essentially all minor and major life goals, adaptation impossible, distressing transition period	Requires full-time care	Permanent
Death			

[a] 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 life 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.

©2010 American Medical Association. All rights reserved.

Downloaded From: by a UNIVERSITE DE LAUSANNE User on 01/23/2018

tional Organizations of Medical Sciences evaluates research risks by comparing them with the risks of routine medical examinations.[11] To implement this standard, committees first must decide which routine examinations will be used as comparators. Should they use the risks of routine examinations when performed by experts, clinicians of average experience, or medical residents?

To illustrate how SERR works, it will be necessary to select 1 standard, as well as specific activities to apply that standard. Regulations in many countries, including the United States,[10] India,[33] South Africa,[34] Canada,[35] and Uganda,[36] mandate that the risks of research interventions be compared with the risks ordinarily encountered in daily life. This standard has been widely interpreted, by the Institute of Medicine and others, to refer to the risks ordinarily encountered by average, healthy individuals in their daily lives.[37] Given the prevalence of this risks of daily life standard, using the activities of daily life as the comparator offers a practically relevant test of the SERR method.

Currently, the risks of daily life standard is applied inconsistently and unsystematically. For example, a survey of IRB chairpersons in the United States found that 23% judged allergy skin testing to pose minimal risk, 43% judged it as posing a minor increase over minimal risk, and 27% judged it as posing more than a minor increase over minimal risk.[5] This degree of variation is unsurprising given current reliance on intuitive judgments alone. Can SERR help to implement the risks of daily life standard in a way that avoids these problems?

The activities of daily life pose a wide range of risks. Yet, the risks of daily life standard does not specify which activities within this range should be used to evaluate the risks of research. This ambiguity has led to substantial debate over which activities offer appropriate comparators for clinical research. The debate need not be settled for the purposes of evaluating whether SERR offers a systematic and effective method for comparing the risks of research with the risks of comparator activities. The determination of whether SERR offers an effective method will not be influenced by which comparator activities are selected. In the end, SERR can be used to implement whichever interpretation is endorsed.

For illustrative purposes, the present evaluation of SERR will use what has been proposed as a reasonable interpretation of the risks of daily life standard: the risks of research should be compared with the risks of activities of daily life that are appropriate for ordinary individuals, even in contexts that do not offer the potential for personal benefit. The activities of daily life for which the most systematic data are available are sports, occupational activities, and driving.[38-44] A number of these activities are widely regarded as acceptable, even when they do not offer the potential for individual benefit. For example, it seems acceptable for individuals to be exposed to the risks of a car trip in order to participate in a charitable activity. Similarly, it seems acceptable to invite individuals, including those who do not enjoy playing sports, to participate in a charity basketball game. Therefore, for present purposes, the risks of driving and sports will be used to evaluate whether SERR offers an effective method for comparing the risks of research with the risks of comparator activities.

To compare the risks of research with the risks of comparator activities, it is important to know the strength of the evidence used to identify the potential harms posed by the research intervention. This information allows review committees to err on the side of caution when evaluating research interventions for which few relevant data are available. For this purpose, FIGURE 1 and FIGURE 2 show the comparison between the risks of daily life and the risks of epicutaneous allergy skin testing and percutaneous liver biopsy, respectively.

The potential harms are based on the available literature and expert opinion. All available data were evaluated by an expert in the field. Four physicians, 1 nurse, and 1 philosopher independently classified the potential harms by magnitude. Disagreement was rare, typically spanned no more than 1 level of magnitude, and was resolved by discussion. It was assumed that treatment of the potential harms would result in the outcomes expected in an average, normal, otherwise healthy patient. Treatment in an intensive care unit was stipulated to bridge periods of acute illness without long-term sequelae other than those expected from the underlying condition.

The final magnitude level assigned to each potential harm is indicated in the figure legends. The lowest plotted likelihood was 0.001 per 100 000. This number was chosen by convention, based on the assumption that likelihoods less than 1 per 100 million are unlikely to make a difference in normative judgment. Figure 1 and Figure 2 indicate strengths of evidence with shades of black.

The preliminary strength of evidence was determined by the number

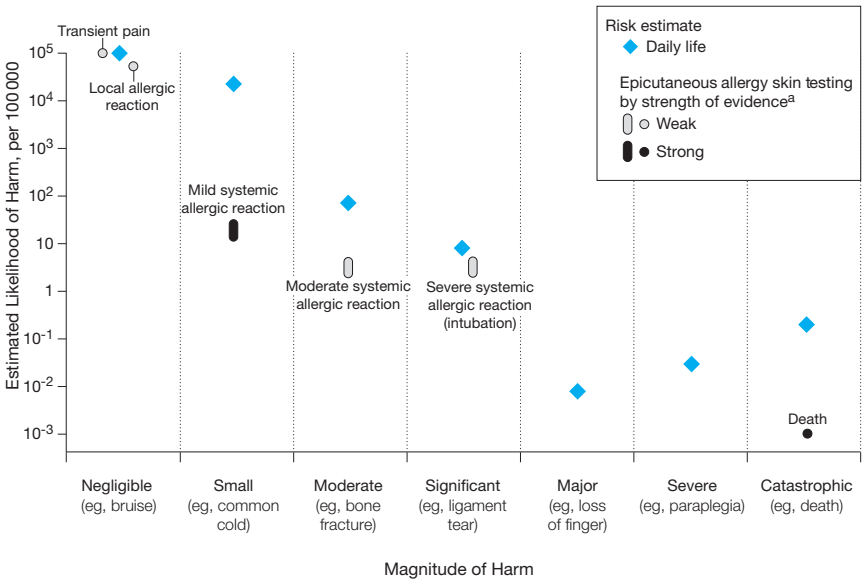
Box. The 4-Step Process of Systematic Evaluation of Research Risks

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.

Downloaded From: by a UNIVERSITE DE LAUSANNE User on 01/23/2018

Figure 1. Comparison of the Risks of Daily Life With the Risks of Epicutaneous Allergy Skin Testing



Estimated risks of epicutaneous allergy skin testing (per 100 000): transient pain (negligible), approximately 100 000; local allergic reaction (negligible), approximately 50 000; mild systemic allergic reaction (small), 11 to 30; moderate or severe systemic allergic reaction (moderate or significant), 2 to 5; and death (catastrophic), approximately 0 (1 case report).[45-52] Daily life risks in the United States (per 100 000): bruise (negligible), approximately 100 000 (all age groups); common cold (1 day [small]), approximately 22 000 (children); bone fracture or dislocation (surfing contest [moderate]), approximately 70 (adults); complete ligament tear of knee (sports practice [significant]), approximately 8 (adolescents); loss of 1 finger (workday in service sector [major]), approximately 0.008 (adults); paraplegia (day of skiing [severe]), approximately 0.03 (all age groups); and death (riskier car trip [catastrophic]), approximately 0.2 (adolescents/adults).[38-44]
[a] Span of elongated data markers indicates range of estimated risk.

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[45-52] 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 the knee (significant), and death (catastrophic).

Based on the literature[45-52] 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[53-64] 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 analgesics; (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) hematothorax, requiring aspiration; (10) pneumothorax, requiring no treat-

©2010 American Medical Association. All rights reserved.

Downloaded From: by a UNIVERSITE DE LAUSANNE User on 01/23/2018

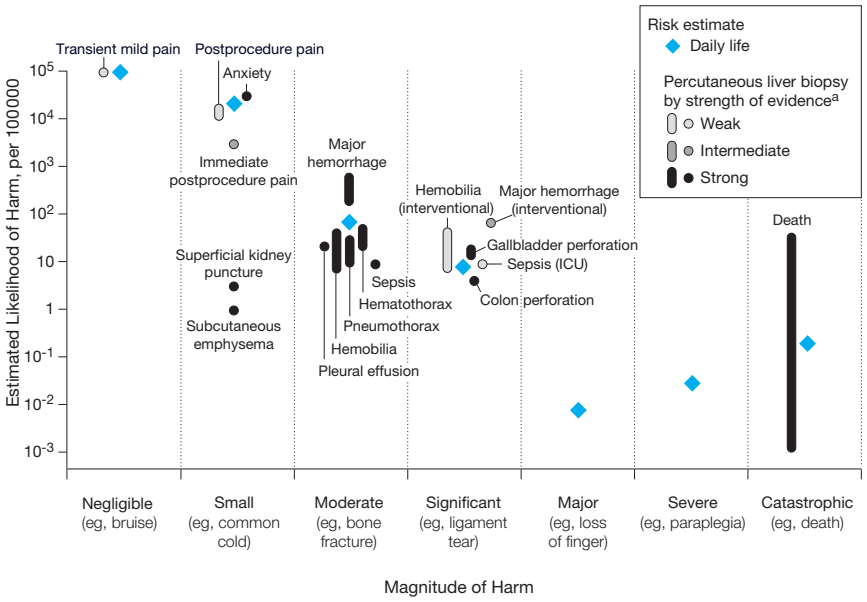
ment or drainage and analgesics for 2 to 5 days; (11) hemobilia, involving colic and/or black stool and/or jaundice for 1 week; (12) sepsis, requiring antibiotics; (13) major hemorrhage, requiring interventional radiography or surgery; (14) hemobilia, requiring interventional radiography or surgery; (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[53-64] 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.[41]

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.

Figure 2. Comparison of the Risks of Daily Life With the Risks of Percutaneous Liver Biopsy



Estimated risks of percutaneous liver biopsy (per 100 000): anxiety (small), 31 000; transient mild pain (negligible), approximately 100 000; immediate postprocedure pain (small), approximately 3000; postprocedure pain (small), approximately 10 000 to 20 000; superficial kidney puncture (small), 3; subcutaneous emphysema (small), 1; pleural effusion (moderate), 21; hematothorax (moderate), 18 to 63; pneumothorax (moderate), 8 to 35; major hemorrhage requiring transfusion (moderate), 160 to 733; hemobilia requiring conservative treatment (moderate), 6 to 50; sepsis requiring antibiotic treatment (moderate), 9; major hemorrhage requiring interventional radiography/surgery (significant), 67; hemobilia requiring interventional treatment (significant), 6 to 50; sepsis requiring intensive care unit (ICU) treatment (significant), 9; gallbladder perforation (significant), 12 to 22; colon perforation (significant), 4; death (catastrophic), 0 to 40.[53-64] For daily life risks see Figure 1 legend.
[a] Span of elongated data markers indicates range of estimated 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[65,66] 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,[8] a situation that is likely to be similar in other countries. Systematic risk evaluations are not possible in that time frame. In

©2010 American Medical Association. All rights reserved.

Downloaded From: by a UNIVERSITE DE LAUSANNE User on 01/23/2018

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,[67-71] complementary conceptual analysis will be needed to determine which comparator activities are appropriate.[71] 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

1. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. *Science*. 1974;185(4157):1124-1131.
2. Slovic P. Perception of risk. *Science*. 1987;236(4799):280-285.
3. Slovic P. *The Perception of Risk*. London, England: Earthscan Publications; 2000.
4. Weinstein ND. Optimistic biases about personal risks. *Science*. 1989;246(4935):1232-1233.
5. Shah S, Whittle A, Wilfond B, Gensler G, Wendler D. How do institutional review boards apply the federal risk and benefit standards for pediatric research? *JAMA*. 2004;291(4):476-482.
6. Lenk C, Radenbach K, Dahl M, Wiesemann C. Nontherapeutic research with minors. *J Med Ethics*. 2004;30(1):85-87.
7. Office for Human Research Protections database. http://ohrp.cit.nih.gov/search/. Accessed January 12, 2010.
8. Department of Health & Human Services Office of the Inspector General. Institutional review boards: a time for reform, June 1998 (publication OEI-01-97-00193). http://oig.hhs.gov/oei/reports/oei-01-97-00193.pdf. Accessed September 10, 2010.
9. Wood A, Grady C, Emanuel EJ. Regional ethics organizations for protection of human research participants. *Nat Med*. 2004;10(12):1283-1288.
10. Department of Health and Human Services. US Code of Federal Regulations, 45 CFR 46; revised 1991. Protection of Human Subjects.

 ©2010 American Medical Association. All rights reserved.

Downloaded From: by a UNIVERSITE DE LAUSANNE User on 01/23/2018

11. Council for International Organizations of Medical Sciences (CIOMS). International ethical guidelines for biomedical research involving human subjects 2002. http://www.cioms.ch/publications/layout_guide2002.pdf. Accessed April 18, 2009.
12. Wendler D. Protecting subjects who cannot give consent. *Hastings Cent Rep*. 2005;35(5):37-43.
13. Wendler D. *The Ethics of Pediatric Research.* New York, NY: Oxford University Press; 2010.
14. London AJ. Two dogmas of research ethics and the integrative approach to human-subjects research. *J Med Philos*. 2007;32(2):99-116.
15. Miller FG, Joffe S. Limits to research risks. *J Med Ethics*. 2009;35(7):445-449.
16. Wendler D, Varma S. Minimal risk in pediatric research. *J Pediatr*. 2006;149(6):855-861.
17. Murray CJ. Quantifying the burden of disease. *Bull World Health Organ*. 1994;72(3):429-445.
18. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). *Med Care*. 1992;30(6):473-483.
19. Brooks R. EuroQol: the current state of play. *Health Policy*. 1996;37(1):53-72.
20. Erickson P. Evaluation of a population-based measure of quality of life. *Qual Life Res*. 1998;7(2):101-114.
21. Rosser R, Kind P. A scale of valuations of states of illness. *Int J Epidemiol*. 1978;7(4):347-358.
22. Furlong WJ, Feeny DH, Torrance GW, Barr RD. The Health Utilities Index (HUI) system for assessing health-related quality of life in clinical studies. *Ann Med*. 2001;33(5):375-384.
23. Feeny D, Furlong W, Boyle M, Torrance GW. Multi-attribute health status classification systems. *Pharmacoeconomics*. 1995;7(6):490-502.
24. Kaplan RM, Bush JW, Berry CC. Health status index. *Med Care*. 1979;17(5):501-525.
25. EuroQol—a new facility for the measurement of health-related quality of life. *Health Policy*. 1990;16(3):199-208.
26. Murray CJ. Rethinking DALYs. In: Murray CJ, Lopez AD, eds. *The Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability From Diseases, Injuries, and Risk Factors in 1990 and Projected to 2020.* Cambridge, MA: Harvard University Press; 1996:1-98.
27. National Cancer Institute (NCI). NCI common terminology criteria for adverse events v3.0 (CTCAE), August 9, 2006. http://ctep.info.nih.gov/protocolDevelopment/electronic_applications/docs/ctcaev3.pdf. Accessed January 20, 2010.
28. Cox EP. The optimal number of response alternatives for a scale. *J Mark Res*. 1980;17(4):407-422.
29. Preston CC, Colman AM. Optimal number of response categories in rating scales. *Acta Psychol (Amst)*. 2000;104(1):1-15.
30. Miller GA. The magical number seven plus or minus two. *Psychol Rev*. 1956;63(2):81-97.
31. Feinberg J. *The Moral Limits of the Criminal Law: Harm to Others.* Vol 1. New York, NY: Oxford University Press; 1984.
32. Levine RJ. *Ethics and Regulation of Clinical Research.* Baltimore, MD: Urban & Schwarzenberg; 1981.
33. Indian Council of Medical Research. *Ethical Guidelines for Biomedical Research on Human Participants.* New Delhi, India: Director-General Indian Council of Medical Research; 2006.
34. South African Medical Research Council. *Guidelines on Ethics for Medical Research: General Principles, Including Research on Children, Vulnerable Groups, International Collaboration and Epidemiology.* Cape Town, South Africa: South African Medical Research Council; 2002.
35. Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, Social Sciences and Humanities Research Council of Canada. Tri-Council policy statement: ethical conduct for research involving humans; Ottawa, Canada; 2005. http://www.pre.ethics.gc.ca/policy-politique/tcps-eptc/docs/TCPS%20October%202005_E.pdf. Accessed September 20, 2010.
36. Uganda National Council for Science and Technology. *National Guidelines for Research Involving Humans as Research Participants.* Kampala, Uganda: UNCST; 2007.
37. Committee on Clinical Research Involving Children. *Ethical Conduct of Clinical Research Involving Children.* In: Field JJ, Behrman RE, eds. Washington, DC: The National Academies Press; 2004.
38. Nathanson A, Bird S, Dao L, Tam-Sing K. Competitive surfing injuries. *Am J Sports Med*. 2007;35(1):113-117.
39. Ingram JG, Fields SK, Yard EE, Comstock RD. Epidemiology of knee injuries among boys and girls in US high school athletics. *Am J Sports Med*. 2008;36(6):1116-1122.
40. Brown J. Amputations: a continuing workplace hazard (originally published January 30, 2003). US Dept of Labor; Bureau of Labor Statistics Web site. http://www.bls.gov/opub/cwc/sh20030114ar01p1.htm. Accessed September 20, 2010.
41. Floyd T. Alpine skiing, snowboarding, and spinal trauma. *Arch Orthop Trauma Surg*. 2001;121(8):433-436.
42. Wendler D, Belsky L, Thompson KM, Emanuel EJ. Quantifying the federal minimal risk standard. *JAMA*. 2005;294(7):826-832.
43. Garibaldi RA. Epidemiology of community-acquired respiratory tract infections in adults. *Am J Med*. 1985;78(6B):32-37.
44. Arruda E, Pitkäranta A, Witek TJ Jr, Doyle CA, Hayden FG. Frequency and natural history of rhinovirus infections in adults during autumn. *J Clin Microbiol*. 1997;35(11):2864-2868.
45. Bernstein DI, Wanner M, Borish L, Liss GM; Immunotherapy Committee, American Academy of Allergy, Asthma and Immunology. Twelve-year survey of fatal reactions to allergen injections and skin testing. *J Allergy Clin Immunol*. 2004;113(6):1129-1136.
46. Codreanu F, Moneret-Vautrin DA, Morisset M, et al. The risk of systemic reactions to skin prick-tests using food allergens. *Eur Ann Allergy Clin Immunol*. 2006;38(2):52-54.
47. Devenney I, Fälth-Magnusson K. Skin prick tests may give generalized allergic reactions in infants. *Ann Allergy Asthma Immunol*. 2000;85(6 pt 1):457-460.
48. Lin MS, Tanner E, Lynn J, Friday GA Jr. Nonfatal systemic allergic reactions induced by skin testing and immunotherapy. *Ann Allergy*. 1993;71(6):557-562.
49. Lockey RF, Benedict LM, Turkeltaub PC, Bukantz SC. Fatalities from immunotherapy (IT) and skin testing (ST). *J Allergy Clin Immunol*. 1987;79(4):660-677.
50. Reid MJ, Lockey RF, Turkeltaub PC, Platts-Mills TA. Survey of fatalities from skin testing and immunotherapy 1985-1989. *J Allergy Clin Immunol*. 1993;92(1 pt 1):6-15.
51. Turkeltaub PC, Gergen PJ. The risk of adverse reactions from percutaneous prick-puncture allergen skin testing, venipuncture, and body measurements. *J Allergy Clin Immunol*. 1989;84(6 pt 1):886-890.
52. Valyasevi MA, Maddox DE, Li JT. Systemic reactions to allergy skin tests. *Ann Allergy Asthma Immunol*. 1999;83(2):132-136.
53. Bravo AA, Sheth SG, Chopra S. Liver biopsy. *N Engl J Med*. 2001;344(7):495-500.
54. Cadranel JF, Rufat P, Degos F; for the Group of Epidemiology of the French Association for the Study of the Liver (AFEF). Practices of liver biopsy in France. *Hepatology*. 2000;32(3):477-481.
55. Gilmore IT, Burroughs A, Murray-Lyon IM, Williams R, Jenkins D, Hopkins A. Indications, methods, and outcomes of percutaneous liver biopsy in England and Wales. *Gut*. 1995;36(3):437-441.
56. Grant A, Neuberger J; British Society of Gastroenterology. Guidelines on the use of liver biopsy in clinical practice. *Gut*. 1999;45(suppl 4):IV1-IV11.
57. Greenwald R, Chiprut RO, Schiff ER. Percutaneous aspiration liver biopsy using a large-caliber disposable needle. *Am J Dig Dis*. 1977;22(12):1109-1114.
58. Malnick S, Melzer E. Routine ultrasound-guided liver biopsy. *J Clin Gastroenterol*. 2005;39(10):900-903.
59. McGill DB, Rakela J, Zinsmeister AR, Ott BJ. A 21-year experience with major hemorrhage after percutaneous liver biopsy. *Gastroenterology*. 1990;99(5):1396-1400.
60. Minuk GY, Sutherland LR, Wiseman DA, MacDonald FR, Ding DL. Prospective study of the incidence of ultrasound-detected intrahepatic and subcapsular hematomas in patients randomized to 6 or 24 hours of bed rest after percutaneous liver biopsy. *Gastroenterology*. 1987;92(2):290-293.
61. Okuda K, Musha H, Nakajima Y, et al. Frequency of intrahepatic arteriovenous fistula as a sequela to percutaneous needle puncture of the liver. *Gastroenterology*. 1978;74(6):1204-1207.
62. Perrault J, McGill DB, Ott BJ, Taylor WF. Liver biopsy: complications in 1000 inpatients and outpatients. *Gastroenterology*. 1978;74(1):103-106.
63. Piccinino F, Sagnelli E, Pasquale G, Giusti G. Complications following percutaneous liver biopsy. *J Hepatol*. 1986;2(2):165-173.
64. Van Thiel DH, Gavaler JS, Wright H, Tzakis A. Liver biopsy: its safety and complications as seen at a liver transplant center. *Transplantation*. 1993;55(5):1087-1090.
65. Finkel AM. Toward less misleading comparisons of uncertain risks. *Environ Health Perspect*. 1995;103(4):376-385.
66. Andrews CJ, Hassenzahl DM, Johnson BB. Accommodating uncertainty in comparative risk. *Risk Anal*. 2004;24(5):1323-1335.
67. Starr C. Social benefit versus technological risk. *Science*. 1969;165(899):1232-1238.
68. Shrader-Frechette KS. *Risk and Rationality: Philosophical Foundations for Populist Reforms.* Berkeley: University of California Press; 1991.
69. Sunstein C. *Risk and Reason.* Cambridge, MA: Cambridge University Press; 2002.
70. Fischoff B, Lichtenstein S, Slovic P, Derby SL, Keeney RL. *Acceptable Risk.* New York, NY: Cambridge University Press; 1981.
71. Hansson SO. Ethical criteria of risk acceptance. *Erkenntnis*. 2003;59:291-309. doi:10.1023/A:1026005915919.

©2010 American Medical Association. All rights reserved.

Downloaded From: by a UNIVERSITE DE LAUSANNE User on 01/23/2018