Uncertainties in baseline risk estimates and confidence in treatment effects

Frederick A Spencer,[1] Alfonso Iorio,[1 2] John You,[1 2] M Hassad Murad,[3] Holger J Schünemann,[1 2] Per O Vandvik,[4 5] Mark A Crowther,[1 6] Kevin Pottie,[7] Eddy S Lang,[8] Joerg J Meerpohl,[9] Yngve Falck-Ytter,[10] Pablo Alonso-Coello,[11] Gordon H Guyatt[2]

[1]Department of Medicine, McMaster University, Hamilton ON L8N 4A6, Canada
[2]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton
[3]Division of Preventive Medicine, Mayo Clinic, Rochester, Minnesota, USA
[4]Department of Medicine, Inlandet Hospital Trust, GjØvik, Norway
[5]Norwegian Knowledge Centre for Health Services, Oslo, Norway
[6]Department of Molecular Medicine, McMaster University, Hamilton
[7]Departments of Family Medicine and Epidemiology and Community Medicine, University of Ottawa, Ottawa, Canada
[8]Division of Emergency Medicine, University of Calgary, Calgary, Canada
[9]German Cochrane Center, Institute of Medical Biometry and Medical Informatics, University Medical Center, Freiburg, Germany
[10]Department of Medicine, Case Western Reserve University, Cleveland, USA
[11]Iberoamerican Cochrane Centre, CIBERESP-IIB Sant Pau, Barcelona, Spain

Correspondence to: F A Spencer fspence@mcmaster.ca

Accepted: 29 October 2012

Cite this as: *BMJ* 2012;345:e7401
doi: 10.1136/bmj.e7401

The GRADE system provides a framework for evaluating how risk of bias, publication bias, imprecision, inconsistency, and indirectness may reduce confidence in estimates of relative effects of interventions on outcomes. However, GRADE and all other systems for rating confidence in effect estimates do not fully address uncertainty in baseline risk and its impact on confidence in absolute estimates of treatment effect. In this article the authors examine factors that may reduce confidence in estimates of baseline risk and thus estimates of absolute treatment benefit

The GRADE system provides a framework for assessing confidence in estimates of the effect (“quality of evidence”) of alternative management strategies on outcomes that are important to patients.[1-6] The GRADE system includes consideration of risk of bias, publication bias, imprecision, inconsistency, and indirectness and their impact on confidence in estimates of benefits and harms. The evaluation of each of these issues has, thus far, focused almost exclusively on their potential impact on estimates of relative effect. Because, in most instances, estimates of relative effect of a therapy are similar across different baseline risks, one can apply these relative estimates to the best estimates of overall baseline risk or, if available, estimates from subgroups that differ in baseline risk.

Using the GRADE approach, guideline panellists multiply the best estimate of relative effect by the best available estimate of baseline risk to obtain an estimate of absolute effect (see box). Limitations of the evidence with respect to risk of bias, publication bias, imprecision, inconsistency, or indirectness may reduce confidence in estimates of the relative risk reduction and affect the strength of guideline recommendations.

As with estimates of relative effect, the quality of evidence supporting estimates of baseline risk can vary. At present, GRADE—and all other systems that address confidence in estimates of treatment effect—fails to fully explore issues of confidence in estimates of baseline risk. Nor do these systems incorporate the 95% confidence interval of a baseline risk estimate when deriving their absolute risk estimates. Thus, evaluating uncertainty in baseline risk, and its impact on confidence in absolute estimates of treatment effect, remains an important outstanding issue.

We suggest that the domains currently used in GRADE (risk of bias, publication bias, imprecision, inconsistency, and indirectness) can also help to understand issues of confidence in baseline risk estimates. In this article we use examples from the *Antithrombotic Therapy and Prevention of Thrombosis, 9th edition* (AT9) to examine how these issues may influence estimates of baseline risk and the subsequent impact on derived estimates of absolute effect.

SUMMARY BOX

Uncertainty in baseline risk estimates and its impact on confidence in absolute estimates of treatment effect are not adequately evaluated in systems of judging confidence in estimates of treatment effect—including GRADE

Risk of bias, publication bias, imprecision, inconsistency, and indirectness can affect confidence in estimates of baseline risk and subsequently confidence in derived estimates of absolute effect of diagnostic and treatment modalities

GRADE’s structure can be easily and effectively adapted to better understand issues regarding confidence in baseline risk. Concerns can be categorised into one or more of the same domains used by GRADE to evaluate evidence supporting a relative risk estimate

Risk of bias

In addressing treatment effects, evidence from observational studies generally warrants lower confidence than evidence from randomised controlled trials. However, community based or population based observational studies can provide better estimates of the baseline risk associated with a given clinical condition than randomised controlled trials, which often enrol highly selected populations. This will be true, however, only if the relevant observational studies are at low risk of bias in ascertaining event rates.

In the AT9 guidelines addressing atrial fibrillation,[7] the panellists derived baseline risk estimates of non-fatal stroke for patients with atrial fibrillation from pooled event rates in the control arms of six randomised controlled trials conducted in the early 1990s.[8] The panellists acknowledged limitations in these estimates, including the fact that

bmj.com

- Use of Mendelian randomisation to assess potential benefit of clinical intervention (*BMJ* 2012;345:e7325)
- When is a further clinical trial justified? (*BMJ* 2012;345:e5913)
- Consort 2010 statement: extension to cluster randomised trials (*BMJ* 2012;345:e5661)

Estimates of absolute effect

When patients and clinicians are trading off desirable and undesirable consequences of an intervention they require estimates of absolute effect. For instance, patients with atrial fibrillation need to trade off risk of strokes versus risk of major bleeding, and they need to know how many strokes anticoagulation will prevent, and how many strokes it will cause. This is best done by applying estimates of relative effect to estimates of baseline risk, such as by means of the CHADS$_2$ scoring system:

Scenario 1

Patients with a CHADS$_2$ score of 1 have a yearly risk of stroke of about 22 per 1000

The relative risk of stroke in patients receiving warfarin is 0.34

Therefore the risk of stroke in treated patients is 22×0.34 per 1000 = 7 per 1000

Thus, the absolute reduction in risk is 22−7 = 15 per 1000

Scenario 2

Patients whose CHADS$_2$ score is 2 have a yearly risk of stroke of about 45 per 1000

The relative risk of stroke in patients receiving warfarin is also 0.34 in this group

Therefore the risk of stroke in treated patients is 45×0.34 per 1000 = 15 per 1000

Thus, the absolute reduction in risk is 45−15 = 30 per 1000

the trials enrolled less than 10% of patients screened. In addition, the authors noted that more recent data from a large administrative database including a broader spectrum of patients suggested lower rates of non-fatal thromboembolism in untreated patients (4.2 *v* 2.1 per 100 patient years).[9] These lower rates may be more reflective of event rates in the current era and would make an important difference in the estimated absolute risk reduction (that is, a more modest effect) associated with anticoagulation in this class of patients.

The panel chose, however, to rely on the trial data because of concern that the lower estimate of stroke derived from the large administrative database reflected under-ascertainment of stroke (that is, a high risk of bias).

Publication bias

Relative risk estimates for the impact of a therapeutic strategy in relation to a comparator on a target outcome are ideally drawn from a systematic review of relevant studies. These estimates are biased if the included studies are unrepresentative because of preferential publication of studies favouring a stronger or weaker effect.[10 11] In GRADE, systematic review and guideline authors may rate down their confidence in effect estimates if they believe publication bias is likely.[12]

Publication bias may similarly affect estimates of baseline risk. Ideally, systematic reviews of large observational studies including a representative sample of the target population will inform estimates of baseline risk. However, observational studies reporting higher undesirable event rates may be less likely to be published than studies reporting lower event rates. This may be particularly true for surgical series, in which surgeons experiencing a higher rate of adverse events than their colleagues may be reluctant to display their less enviable record to the surgical world.

Imprecision

Examination of 95% confidence intervals for estimates of absolute effects provides the optimal approach to determine precision of the estimate.[13] For practice guidelines, rating down the confidence in absolute estimates of effect is warranted if clinical action would differ if the upper versus the lower boundary of the confidence interval represented the truth.

Imprecision in estimates of baseline risk will affect the derived absolute effect of a given therapy. The AT9 guidelines suggest venous thromboprophylaxis with low dose, low molecular weight heparin (LMWH) for women undergoing assisted reproduction who develop severe ovarian hyperstimulation syndrome.[14] The authors estimate that use of low dose LMWH will prevent 26 venous thromboembolic events (95% confidence interval 13 to 42) per 1000 patients treated. Their estimate comes from applying indirect evidence of the relative risk reduction associated with low dose LMWH from existing surgical literature (relative risk 0.36 (95% confidence interval 0.20 to 0.67)) to a baseline venous thromboembolic event rate of 4.1%. The quality of evidence for the resulting recommendation was rated down for indirectness (relative risk estimate derived from a general surgical population).

This baseline risk of 4.1% was, however, derived from a sample of just 49 patients with severe ovarian hyperstimulation syndrome from a cohort of 2748 cycles of assisted reproduction therapy.[15] The 95% confidence interval around the 4.1% point estimate is 1.1% to 13.7%. Therefore, depending on selection of baseline risk (and multiplying by the 95% confidence interval of the relative risk reduction), use of low dose LMWH in such patients may result in as few as four events prevented to as many as 110 events prevented per 1000 treated. The lower estimate of four events per 1000 treated would make any recommendation for thromboprophylaxis in this population far less attractive than the latter. Such imprecision is likely to arise in rare conditions.

Inconsistency

In GRADE, confidence in estimates of effect from a body of evidence may be rated down if the magnitude of treatment effect varies substantially across relevant studies.[16] Inconsistency may also undermine estimates of baseline risk. Guideline developers often derive baseline risk estimates by pooling event rates from observational studies using similar populations. Event rates among individual studies may vary greatly from the pooled estimate, thus decreasing confidence in this estimate.

In the chapter of the AT9 guidelines addressing prophylaxis for venous thromboembolism in surgical patients, the authors suggest an average risk of 2.1% for venous thromboembolism in patients undergoing craniotomy and suggest use of lower extremity external compression devices as prophylaxis.[17] This risk estimate was derived from a pooled estimate of event rates observed in eight studies providing event rates in neurosurgical patients using external compression devices.[18] Based on this estimate, and multiplying by a relative risk estimate of 0.56, the authors calculated that use of LMWH instead of external compression devices would prevent nine non-fatal symptomatic venous thromboembolic events per 1000 patients treated. Using a similar approach, they calculated that LMWH will cause 11 more non-fatal intracranial bleeds. Based on these estimates of absolute benefit and harm, they provided a weak recommendation for mechanical prophylaxis over LMWH for venous thromboembolism.

The venous thromboembolic event rates in the included studies varied from 0% to 10%. This inconsistency decreases our confidence in the baseline risk estimates and consequently in the recommendation. If true venous thromboembolic event rates are closer to 10% despite use of external compression devices, LMWH would prevent 44 non-fatal venous thromboembolic events per 1000 treated. Based on this estimate of absolute effect, it is less clear which prophylactic strategy should be recommended.

Indirectness

Direct evidence in the GRADE framework includes studies that have enrolled the populations of interest, delivered the intervention in the manner of interest, and measured the outcomes important to patients over the time frame of interest.[19] A guideline panel will have concerns about indirectness when the population, intervention, or outcome differs from those in which they are interested—what one might otherwise call limitations of applicability.

The evidence supporting a baseline risk estimate can also be indirect. This occurs when baseline risk estimates are derived from a population that differs significantly from the population to whom the resulting guidelines are directed. Given the lack of high quality evidence documenting outcome event rates for specific disease states in community settings, estimates of baseline risks for outcome events are often derived from event rates in the control arms of randomised controlled trials. In general, patients enrolled in such trials are younger, have less comorbidity, and have better outcomes than patients encountered in clinical practice. Therefore, application of relative risk estimates for a given intervention to a baseline risk rate derived from a randomised controlled trial may underestimate both the absolute benefits and harms associated with that intervention in the community setting.

Indirectness may also lead to overestimates of absolute effects. As discussed above, baseline risk estimates of non-fatal stroke for patients with atrial fibrillation in the AT9 guidelines were derived from the pooled event rates in the control arms of six randomised controlled trials comparing warfarin with aspirin in the early 1990s.[2] For $CHADS_2$ (stroke risk) scores of 0, 1, 2, and 3–6, respectively, baseline event rates of 0.8%, 2.2%, 4.5%, and 9.6% per year were used to generate estimates of absolute benefit with warfarin. Rates of non-fatal thromboembolism in untreated patients were significantly lower in a more current and representative population than seen in the older trials (for $CHADS_2$ scores of 0, 1, 2, 3, and 4–6, respectively, absolute event rates of 0.4%, 1.2%, 2.5%, 3.9%, and 6.3% were reported).[9]

Use of the estimates from the more current observational database would have resulted in a substantial decrease in the calculated absolute benefit of warfarin over one year. For example, using the baseline risk estimates from the older trials, warfarin use is predicted to prevent 30 non-fatal strokes per 1000 (95% confidence interval 23 to 35 strokes prevented) in patients with a $CHADS_2$ score of 2. With the lower baseline risk estimates, however, the absolute benefit of warfarin decreases—resulting in prevention of only 16 (13 to 19) non-fatal thromboembolic events per 1000 treated. Similarly, absolute benefit for patients with a $CHADS_2$ score of 1 would have declined from 15 (11 to 17) fewer events to eight (6 to 9) fewer events without a change in estimated harm due to bleeding. These revised absolute benefits would potentially alter recommendations—possibly changing the direction of the recommendation for warfarin in patients with a $CHADS_2$ score of 1 and reducing the strength of the recommendation from strong to weak for warfarin over aspirin in patients with a $CHADS_2$ score of 2.

Discussion

Adopted by over 60 groups worldwide, the GRADE approach represents an important innovation in interpreting evidence from systematic reviews, health technology assessments, and clinical practice guidelines. At present, the approach focuses on evaluating confidence in estimates in the relative effect of one treatment strategy over another, and then—in most cases—assuming that this confidence also applies to estimates of absolute effects. Estimates of baseline risks, however, directly affect estimates of absolute risks and benefits of a treatment. We suggest that the confidence in estimates of baseline risks is subject to the same issues as evidence for relative effects of a treatment strategy.

To date guidelines have rarely considered issues of baseline risk. As our examples illustrate, GRADE's structure can be usefully adapted to better understand issues regarding confidence in baseline risk.

This discussion has only illustrated the problem. We are not yet ready to offer specific guidance on how to rate down confidence in estimates of baseline risk. As with other methodological problems previously encountered, a great deal of work studying specific examples needs to be done before we can offer concrete solutions. This article represents a first step in this process.

Contributors: FAS and GHG conceived the study and take responsibility for the integrity of the data and the accuracy of the data analysis. FAS is the guarantor. FAS, AI, and GHG designed the study. FAS, AI, JY, MHM, POV, and GHG analysed and interpreted the data. FAS, AI, MHM, and GHG drafted the manuscript. All authors critically revised the manuscript.

Funding: This study was not externally funded.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; all authors are members of the GRADE working group and were contributors to *Antithrombotic Therapy and Prevention of Thrombosis, 9th edition*, but this manuscript is not submitted on behalf of either group.

Ethical approval: Not required.

Provenance: All authors were contributors to the *Antithrombotic Therapy and Prevention of Thrombosis, 9th edition*, which used the GRADE system to assess quality of evidence underlying subsequent recommendations. During the development of these guidelines, authors (in particular FAS, AI, POV, and GHG) struggled with issues of confidence in estimates of baseline risk and how to evaluate and categorise uncertainty in baseline risk estimates. Further discussion of these issues among panel members prompted the development of this manuscript.

1 Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *J Clin Epidemiol* 2011;64:380-2.

2 Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, et al. Going from evidence to recommendations. *BMJ* 2008;336:1049-51.

3 Guyatt GH, Oxman AD, Kunz R, Jaeschke R, Helfand M, Liberati A, et al. Incorporating considerations of resources use into grading recommendations. *BMJ* 2008;336:1170-3.

4 Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schunemann HJ. What is "quality of evidence" and why is it important to clinicians? *BMJ* 2008;336:995-8.

5 Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924-6.

6 Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008;336:1106-10.

7 You JJ, Singer DE, Howard PA, Lane DA, Eckman MH, Fang MC, et al. Antithrombotic therapy in atrial fibrillation: ACCP Evidence-Based Clinical Practice Guidelines, 9th edition. *Chest* 2012;141(2_suppl):e531S-75S.

8 Gage BF, van Walraven C, Pearce L, Hart RG, Koudstaal PJ, Boode BS, et al. Selecting patients with atrial fibrillation for anticoagulation: stroke risk stratification in patients taking aspirin. *Circulation* 2004;110:2287-92.

9 Singer DE, Chang Y, Fang MC, Borowsky LH, Pomernacki NK, Udaltsova N, et al. The net clinical benefit of warfarin anticoagulation in atrial fibrillation. *Ann Intern Med* 2009;151:297-305.

10 Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev* 2009;(1):MR000006.

11 Dickersin K, Min YI, Meinert CL. Factors influencing publication of research results. Follow-up of applications submitted to two institutional review boards. *JAMA* 1992;267:374-8.

12 Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence—publication bias. *J Clin Epidemiol* 2011;64:1277-82.

13 Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence—imprecision. *J Clin Epidemiol* 2011;64:1283-93.

14 Bates SM, Greer IA, Veenstra D, Prabulos AM, Vandvik PO. Venous thromboembolism, thrombophilia, antithrombotic therapy, and pregnancy. ACCP Evidence-based Clinical Practice Guidelines, 9th Edition. *Chest* 2012;141(2_suppl):e691S-736S.

15 Mara M, Koryntova D, Rezabek K, Kapral A, Drbohlav P, Jirsova S, et al. [Thromboembolic complications in patients undergoing in vitro fertilization: retrospective clinical study]. *Ceska Gynekol* 2004;69:312-6.

16 Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. *J Clin Epidemiol* 2011;64:1294-302.

17 Gould MK, Garcia DA, Wren SM, Karanicolas PJ, Arcelus JI, Heit JA, et al. Prevention of venous thromboembolism in non-orthopedic surgical patients: ACCP Evidence-based Clinical Practice Guideline, 9th Edition. *Chest* 2012;141(2_suppl):e227S-77S.

18 Danish SF, Burnett MG, Ong JG, Sonnad SS, Maloney-Wilensky E, Stein SC. Prophylaxis for deep venous thrombosis in craniotomy patients: a decision analysis. *Neurosurgery* 2005;56:1286-92, 92-4.

19 Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence—indirectness. *J Clin Epidemiol* 2011;64:1303-10.

The human face of improbabilities

I was recently an attending physician on the general medicine inpatient service for two weeks. During that fortnight I admitted 56 patients, among whom I diagnosed four cases of tuberculosis—pulmonary reactivation, disseminated, peritoneal, and spinal osteomyelitis. Three of them were immigrants (from Mexico and Ecuador), and the other was a native of Chicago. None of them had HIV, and none had other risk factors for tuberculosis.

"What are the odds?" I wondered, so I grabbed an envelope lying on my desk and started scribbling on the back. There were 12 000 cases of tuberculosis diagnosed in the United States in 2009.[1] Based on data from the Healthcare Cost and Utilization Project of the Agency for Health Research and Quality, there were 23 million live discharges from the hospital in 2008 (the most recent data available) after exclusion of pregnant women, children, and patients with mental health related diseases.[2] If every person with newly diagnosed tuberculosis is hospitalised, the probability of a patient being admitted to a US hospital with tuberculosis is 1 in 2000. Assuming that cases of tuberculosis are randomly distributed throughout the year and we apply the binomial distribution equation to these numbers, the probability of admitting four or more patients with tuberculosis out of 56 patients is 1 in 41 million. I should have played the lottery.

But, to be fair, I don't work at a typical US hospital. I work at Cook County Hospital in Chicago. Most of our patients are uninsured, and 30% are foreign born. We diagnose 80 cases of tuberculosis a year (roughly one in five cases in our state). We admit about 15 000 cases to our medicine services annually. From these data, the probability of my admitting a patient with tuberculosis is about 1/200. Putting these numbers into the binomial distribution gives a much more modest probability of 1 in 5300. Still, not too shabby. Maybe I should go to Vegas.

As I looked up from my page of numbers at the patient roster in front of me, I realised that these improbabilities were people. Real people who are really sick. My patient with Pott's disease lost his job from the severe disability of his illness. The man with disseminated tuberculosis may be infertile because he had epididymal involvement. The middle aged woman with peritoneal tuberculosis spent three weeks living with the belief that she had ovarian cancer. These people aren't numbers. And they certainly aren't lucky. My good fortune isn't being exposed to unusual medical phenomena. It stems from being able to care for these people. But those odds are long, aren't they? Maybe I should go to Vegas after all; only I'll take my patients with me.

Josh Baru Cook County Hospital, Chicago Illinois, USA
joshua_baru@rush.edu

Patient consent obtained.

1 Centers for Disease Control and Prevention (CDC). Reported tuberculosis in the United States, 2009. www.cdc.gov/tb/statistics/reports/2009/default.htm.

2 Healthcare Cost and Utilization Project (HCUP). HCUP facts and figures 2008. www.hcup-us.ahrq.gov/reports/factsandfigures/2008/TOC_2008.jsp.

Cite this as: *BMJ* 2012;344:d7353