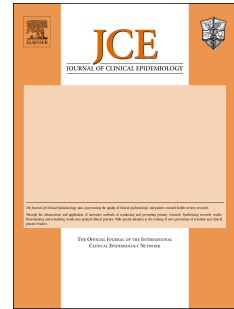# Accepted Manuscript

Certainty ranges facilitated explicit and transparent judgments regarding evidence credibility

Kari A.O. Tikkinen, Samantha Craigie, Holger J. Schünemann, Gordon H. Guyatt

Please cite this article as: Tikkinen KAO, Craigie S, Schünemann HJ, Guyatt GH, Certainty ranges facilitated explicit and transparent judgments regarding evidence credibility, *Journal of Clinical Epidemiology* (2018), doi: 10.1016/j.jclinepi.2018.08.014.

Certainty ranges facilitated explicit and transparent judgments regarding evidence credibility

Kari A.O. Tikkinen[a*], Samantha Craigie[b], Holger J. Schünemann[b,c], Gordon H. Guyatt[b,c]

a. Departments of Urology and Public Health, University of Helsinki and Helsinki University Hospital, Haartmaninkatu 4, 00029 Helsinki, Finland
b. Department of Health Research Methods, Evidence, and Impact, McMaster University, 1280 Main St West, Hamilton, ON, Canada, L8S 4L8
c. Department of Medicine, McMaster University, 1280 Main St West, Hamilton, ON, Canada, L8S 4L8

*Corresponding Author: Kari A.O. Tikkinen. Department of Urology, Helsinki University Hospital and University of Helsinki, Haartmaninkatu 4, 00029 Helsinki, Finland. Email: kari.tikkinen@gmail.com, phone: +358 50 525 0971

Abstract

Objective: The GRADE approach to rating certainty of evidence includes five domains of reasons for rating down certainty. Only one of these, precision, is easily amenable – through the confidence interval – to quantitation. The other four (risk of bias, inconsistency, indirectness, publication bias) are not. Nevertheless, conceptually, one could consider a quantified "certainty range" within which the true effect lies. The certainty range would be at least as wide as the confidence interval, and would expand with each additional reason for uncertainty.

Study Design and Setting: We have applied this concept to rating the certainty of evidence in the baseline risk of venous thromboembolism (VTE) and bleeding in patients undergoing urological surgery. We considered rating up moderate or low quality evidence when the net benefit of VTE prophylaxis was unequivocally positive: that is, when the smallest plausible value of VTE reduction was greater than the largest plausible value of increased bleeding. To establish whether there the net benefit was unequivocally positive, we expanded the range of plausible values by 20% for each of the four non-quantitative domains in which there were serious limitations.

Results: We present how we applied these methods to examples of open radical cystectomy and laparoscopic partial nephrectomy. In high VTE risk laparoscopic partial nephrectomy patients and high- and medium VTE risk open radical cystectomy patients, results proved robust to expanded certainty intervals, justifying rating up quality of evidence. In low risk patients, the results were not robust, and rating up was therefore not appropriate.

Conclusion: This work represents the first empirical application in a decision-making context of the previously suggested concept of certainty ranges and should stimulate further exploration of the associated theoretical and practical issues.

Key words: GRADE; guidelines; quality of evidence; systematic reviews; thresholds; thromboprophylaxis

What is new?

Key findings

- This study represents a first foray into utilizing the concept of the certainty range to place a quantitative estimate on domains of uncertainty (risk of bias, inconsistency, indirectness, publication bias) that are up to now addressed only qualitatively.
- We applied quantitative estimates to the baseline risks of venous thromboembolism and major bleeding in patients undergoing urological surgery and in doing so established whether inferences regarding the net benefits of pharmacologic prophylaxis were secure.

What this adds to what was known?

- The GRADE approach to rating certainty of evidence includes five domains of reasons for rating down certainty. Only one of these, precision, is easily amenable – through the confidence interval – to quantitation. The other four (risk of bias, inconsistency, indirectness, publication bias) are not.
- This work highlights the concept of the uncertainty range and the potential for ultimate quantitation of all domains of uncertainty.

What is the implication and what should change now?

- This is the first empirical application in a decision-making context of the previously suggested concept of certainty ranges
- This work should stimulate further exploration of the associated theoretical and practical issues to take these concepts forward.

Introduction

The widely used GRADE approach to rating certainty in evidence (synonyms quality or confidence in evidence) can be applied to a variety of questions in health care, including to alternative management strategies (1) and prognosis (2). In evaluating therapy questions, randomized trials start as high quality evidence; for prognosis, observational studies start as high quality evidence. For both sorts of questions, five domains of limitations may result in rating down certainty.

The uncertainty associated with one of these domains of limitations, imprecision, can be quantitated by examining confidence (for frequentist analysis) or credible (for Bayesian analysis) intervals. The extent of uncertainty associated with the other four domains of limitations – risk of bias, inconsistency, indirectness, and publication bias - is, thus far, not fully amenable to quantitation (3). In this article, we will use the term "certainty range" to characterize uncertainty that considers all these domains.

Conceptually, each of the five limitations extend the range of uncertainty – the range of plausible true effect – around the best estimate of effect. One could therefore picture the certainty range around that best estimate (3). The width of the certainty range would depend on the extent of concerns regarding imprecision – captured in the confidence or credible interval – and the extent of concern regarding the other four domains (Figure 1) (3).

Figure 1 depicts the certainty range – like the confidence interval – as symmetrical around the point estimate. This need not be the case. For instance, if one knew the likely direction of risk of bias, the certainty range could be asymmetrical, skewed in that direction (3). Furthermore, for studies of prognosis or baseline risk – the focus of this article – given that values can range only between 0% and 100%, low probabilities or risks are likely to be skewed to the

4

right (e.g. if the point estimate is 1%, the certainty range can only drop by 1% to 0, while it will plausibly rise to substantially more than 1%).

The extent to which concerns regarding the four, as of yet, non-quantitative domains of uncertainty widen the certainty range is highly speculative. As a result, the notion of the certainty range has heretofore been largely theoretical. In the course of a recently completed project (4), we rated the certainty of evidence regarding the likelihood of thrombosis and bleeding following urological surgery. In doing so, we felt that, despite the speculative nature of the certainty range, it would be worth invoking the concept to help in applying the GRADE certainty of evidence rating. We present the work here because it may be the first scientific publication to empirically apply the certainty range to the rating of GRADE quality of evidence.

We have an important disclaimer: although two of the authors are co-chairs of the GRADE working group (HJS and GHG) this work is not a product of, nor has it been endorsed by, the GRADE working group. Moreover, a number of the concepts presented here, and the way the concepts have been incorporated, go considerably beyond current GRADE guidance. Thus, the current work represents an exploration of possible future directions in thinking about and rating certainty of evidence.

Background of the Project

Patients undergoing surgery are at risk of postsurgical deep venous thrombosis or pulmonary embolism (venous thromboembolism or VTE). VTE can be serious, and indeed fatal. Thus, prophylaxis against VTE with anticoagulants, in particular heparinoids, has become popular.

Unfortunately, pharmacologic prophylaxis is associated with an increased risk of bleeding – always a concern after any surgical procedure - which can also be

serious (in our definition, requiring reoperation) and even fatal. Thus, the decision regarding prophylaxis involves a tradeoff between reduced risk of VTE and increased risk of bleeding. That trade off depends on both the risk of VTE and bleeding in the absence of prophylaxis (which we will call the baseline risk) and the relative decrease in VTE and increase in bleeding with prophylaxis.

As part of a team charged with developing guidelines for prophylaxis after urological surgery (4), we undertook a series of systematic reviews to estimate the baseline risk of both VTE and bleeding (5-7). We interpreted our results in the context of its implications for pharmacologic prophylaxis after major urological procedures.

Methods and Results: Judging the Certainty of Baseline Risk Estimates

Readers will find details of our methods in other articles (4-7). In brief, we used rigorous systematic review methods to identify, evaluate, and summarize observational studies addressing the risk of VTE and bleeding requiring reoperation following urological surgery in the absence of VTE prophylaxis. Our evaluation included a risk of bias assessment for each individual study, including consideration of representativeness of the patient population, thromboprophylaxis documentation, data source, whether a majority of patient recruitment years were earlier or later than 2000, clear specification of duration of follow-up, and study type. In addition, we identified risk factors for VTE and classified patients as at low, medium, and high risk.

We were interested in VTE and bleeding risk at 4 weeks, and if studies reported VTE risk at some other interval we modeled the VTE risk at 4 weeks on basis of data from large-scale population-based observational studies (8, 9) and bleeding risk on the basis of a large randomized trial (10). In doing so, we used a previously published approach (5) that demonstrates an approximately constant hazard of VTE up to 4 weeks (8, 9); bleeding risk, by contrast, is concentrated in

6

the first 4 days (10). The modeling was required to offer the most trustworthy estimates of the benefits and risks of anticoagulation over a period of four weeks. We applied the GRADE rating of certainty of evidence, and interpreted our results in the context of the decision regarding administration of prophylaxis. Best evidence suggests that heparinoids decrease the relative risk of VTE by approximately 50%, and increase the relative risk of bleeding by 50% (6-7). We rated the certainty of the evidence regarding these relative risks as high, and in applying the quantitative estimates assumed no error. A reasonable alternative would have applied some estimate of uncertainty (for instance, confidence intervals) to the relative effects. Doing so would have widened all certainty ranges shown in the following presentation.

The desirability of quantitating uncertainty for the 4 previously listed non-quantitative domains arose when we found a possible large gradient between benefits (VTE reduction) and harm (bleeding requiring reoperation). Let us say, for instance, that the baseline risk of VTE for a procedure was 10%. Applying the relative risk reduction with pharmacologic prophylaxis of 50%, we calculate an absolute reduction in VTE of 5%. Let us say the associated baseline risk of bleeding is 1%, with a relative increase of 50% with prophylaxis, and thus an absolute increase in risk of 0.5%. Even applying a judgment we made that a bleed has twice the importance (disutility) of a VTE, this appears to be a situation in which prophylaxis is clearly indicated and recommended (benefit of 5%, importance-adjusted harm of only 1%).

One solution to expressing the large net benefit might be to rate up the low certainty evidence to moderate, or moderate to high. This approach would be consistent with GRADE's definition of certainty of evidence relating to the extent to which the evidence supports a recommendation. It goes, however, beyond current GRADE guidance in applying this definition to judgments regarding certainty of evidence.

Before coming to the conclusion that benefits of prophylaxis clearly outweigh harm we must, however, consider the uncertainty regarding the baseline risks. We reasoned that the conclusion that thromboprophylaxis was clearly warranted would require that, even assuming the lowest plausible benefit of pharmacologic prophylaxis and the highest plausible harm from bleeding, there would still be a net benefit of VTE prophylaxis. Our challenge then was, for any surgical procedure, to provide estimates of the lowest plausible benefit of prophylaxis and the associated highest risk of bleeding.

Consider, for instance, patients at high risk of VTE undergoing open radical cystectomy – our estimate was 11.6% (the median of the available studies, after adjusting for high risk patient category). Because, in this context, we were skeptical of pooled estimates, rather than using a pooled estimate and the associated confidence interval, we quantitated imprecision as the range of VTE in the available studies – in this case 5.4% to 18.5%.

In terms of other sources of uncertainty, we had no concerns about risk of bias, inconsistency, indirectness, or publication bias. We did, however, lack confidence in both the risk stratification, and the model used to infer VTE risk when studies did not report the 4-week outcome that was our focus[1]. In qualitative terms, these sources of uncertainty that relate to indirectness led us to rate down overall certainty in the evidence regarding VTE from high to moderate.

To judge whether the net benefit of thromboprophylaxis is unequivocally positive we need to determine if the smallest plausible benefit in VTE reduction is greater than the largest plausible harm for bleeding. Let us begin with estimating the smallest plausible benefit in VTE reduction. To make this estimate, we must first establish the uncertainty in baseline risk associated with imprecision alone – in

---

[1] How to classify these sources of uncertainty in the GRADE framework is not altogether clear. Risk of bias and indirectness would both be candidates.

this case suggesting a lowest plausible baseline risk of VTE of 5.4% (the Table presents this, and all subsequent calculation results).

Next – and here we come to the point of this article - we need to quantify the uncertainty associated with risk stratification and model inferences, and thus establish an uncertainty range. We specified, arbitrarily, that each non-quantitative reason for rating down quality of evidence would (relatively) widen the certainty range by 20%.

There could be a number of ways of applying this 20% inflation; our approach was as follows. We took the difference between the point estimate of baseline risk of VTE (in this case 11.6%) and the lower limit based on imprecision alone (in this case 5.4%) and calculated the difference – here, 6.2%. We increased this difference by 20%, multiplying by 1.2 (20% increase to from 6.2% to 7.4%)[2]. To obtain the new lower limit of plausible baseline risk we subtracted this value – 7.4% - from the point estimate (11.6% - 7.4%); the new lower limit of baseline risk in this case was therefore 4.2%. Since pharmacologic prophylaxis reduces the relative risk of VTE by 50%, the smallest possible benefit for high VTE risk patients undergoing cystectomy is 2.1%.

Applying the same logic to bleeding requiring reoperation undergoing cystectomy, our best estimate of bleeding was 0.3%, with a range of 0 to 1.6%. We once again rated down for model uncertainty. To do so quantitatively, we took the difference between the point estimate and upper confidence boundary (1.6% - 0.3% = 1.3%), multiplied by 1.2 to 1.6%, and added this to 0.3% for an upper boundary of 1.9%. However, as approximately 50% of major bleeds occur between surgery and the next morning but cumulative risk of VTE during the first four weeks post-surgery is almost constant, and as – because of mentioned differences in the timing of risks of VTE and bleeding - our recommendations for pharmacological prophylaxis were based on a starting time of the morning after

---

[2] Calculations are, where possible, rounded to a single decimal place.

surgery, the baseline risk of bleeding was only half of that 1.9%. Therefore, as the increase in bleeding with prophylaxis is half of this 1.0%, but because bleeding has double the disutility of VTE, we can consider the negative value of the bleeding 1.0% - this however, is still less than the value of VTE reduction of 2.1% (Table).

Thus, the smallest possible net benefit (value of the smallest reduction in VTE of 2.1% - value of the largest increase in bleeding of 1.0%) is still positive, 1.1%. Given that this is the case, we assumed that we had high certainty evidence for the benefit of thromboprophylaxis in high VTE risk patients undergoing open radical cystectomy because the concerns about baseline risk estimates that led us to rating down for indirectness are mitigated by this sensitivity analysis. Thus we decided to retain the certainty of evidence regarding baseline risk of both VTE and bleeding at high.

The Table presents the results of the high-risk cystectomy example we have just worked through, as well as low and moderate risk of cystectomy, and another example, laparoscopic partial nephrectomy, in which we applied the approach. Figure 2 depicts the expansion of the lower boundary of laparoscopic partial nephrectomy. Even after rating down twice (expanding the range 20% both times) the net benefit remains positive in the high-risk group, thus increasing our conviction that the benefits of thromboprophylaxis outweigh the harms.

In the work we described above, we focused exclusively on issues of the certainty of baseline risk, assuming that we had no concerns regarding relative effect estimates. Thus, we made no attempt to address the integration of baseline and relative risk judgments. In the following we speculate on how GRADE might approach such integration.

Judging Certainty of Evidence of Intervention Effects

Our discussion to this point has focused on the impact of the certainty of the evidence regarding baseline risks on net benefit. GRADE requires a rating of certainty regarding absolute effects of treatment – the absolute effects are influenced both by the relative effects of treatment and by the baseline risk (also referred to as the control event rate). Prior GRADE discussions of making judgments about certainty of absolute effects has focused on the certainty of the estimates of relative effect, with very little discussion of the certainty of baseline risk, the focus of this article. Moreover, the GRADE working group has not yet written about incorporating the uncertainties regarding baseline risk with the uncertainties regarding relative effects. The following, therefore, represents speculation regarding how GRADE might in future address the issue.

How should one integrate certainty ratings of baseline risk with certainty ratings of relative effect? One of the GRADE domains applied to rating certainty of relative effects is indirectness that includes indirectness of the population, the intervention, the comparator and the outcome, as well as indirect comparisons. One might consider uncertainty regarding the baseline risk as an issue of indirectness of the population (Figure 3). In this conceptualization, uncertainty about baseline risk leaves us unsure about the population to whom the baseline risk we have generated applies (if indeed, it applies to any population – it may, for instance, be biased, and though we may still use that baseline risk, we would do so with reservations).

Using this approach, one might, in making judgments regarding certainty of the absolute effect, rate down for indirectness of the population if the certainty of the baseline risk was low or very low, and not rate down if the certainty of baseline risk was moderate or high. Applying such an approach to the current discussion, when certainty prior to considering the extent of overlap between certainty ranges of value-adjusted certainty of thrombosis and bleeding was moderate or low, and the certainty ranges suggest clear benefit or harm, one would not rate down certainty of the evidence regarding the absolute effect of prophylaxis

11

because of uncertainty of baseline risk. In other situations (very low quality evidence, or low quality and certainty ranges do not indicate clear benefit or harm), one would rate down the certainty of the evidence regarding the absolute effect of treatment because of indirectness of the population.

Discussion

This study represents our first foray into utilizing the concept of the certainty range to place a quantitative estimate on domains of uncertainty (risk of bias, inconsistency, indirectness, publication bias) that, within the GRADE framework, are up to now addressed only qualitatively (3). We applied these quantitative estimates to the baseline risk of VTE and bleeding risk in patients undergoing urological surgery and in doing so established whether inferences regarding the net benefits of pharmacologic prophylaxis were secure. In some instances we found they were secure (Table), and in others they were not.

Strengths of this work includes the rigorous systematic review methodology applied to baseline risk estimates of VTE and bleeding in urological surgery and the application of the logic of uncertainty ranges to these real life examples. The practical application, despite the arbitrariness of the magnitude and distribution assumptions of the certainty ranges, presents an important strength of this work, which has previously not been done in any detail. Further, the results have been incorporated in clinical practice guidelines for the use of thromboprophylaxis in urological surgery (4). Additional strengths include the grounding of our approach within the widely used GRADE framework and, as a result, the consideration of all major domains of uncertainty.

The primary limitation of our work is inherent in the uncertainty range approach at this point (3). There is a great deal of arbitrariness in the decision regarding the quantitation of uncertainty of domains other than precision, in particular risk of bias, indirectness, and publication bias. We could have, but did not, address this

12

uncertainty to some extent through more extensive sensitivity analysis. For instance, we could have examined the impact on our inferences of increasing the range of plausible truth for each domain of uncertainty beginning at 10% and increasing to 50% or even more. Some might find our examples more compelling had we used data amenable to conventional meta-analysis, and addressed an issue of relative effects rather than baseline risk. Our primary purpose, however, was to illustrate the essential concept, and our data serves that purpose well.

In terms of prior work, one could argue that those using random effect models, in which confidence intervals are inflated by between-study differences in effect, are applying a quantitative measure of uncertainty related to inconsistency. GRADE's position on the matter has been that this widening of the confidence intervals does not fully address inconsistency.

Formal decision analytic models sometimes do address issues of imprecision, inconsistency and indirectness of populations quantitatively through sensitivity analysis. Thus far they have not, however, done so with respect to indirectness and, to a limited degree, with respect to risk of bias, nor have they expressed results explicitly using the concept of the uncertainty range.

Our primary purpose in presenting this work is to further highlight the concept of the uncertainty range and the potential for ultimate quantitation of all domains of uncertainty (3). We look forward to others' work in taking these concepts forward.

Conflicts of interest: 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 conflicts of interest.

References

1. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383-94.
2. Iorio A, Spencer FA, Falavigna M, Alba C, Lang E, Burnand B, et al. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. BMJ. 2015;350:h870.
3. Schunemann HJ. Interpreting GRADE's levels of certainty or quality of the evidence: GRADE for statisticians, considering review information size or less emphasis on imprecision? J Clin Epidemiol. 2016;75:6-15.
4. Tikkinen KA CR, Gould MK, Naspro R, Novara G, Sandset PM, Violette PD, Guyatt GH. Thromboprophylaxis in Urological Surgery. European Association of Urology Guideline Office. 2017.
5. Tikkinen KA, Agarwal A, Craigie S, Cartwright R, Gould MK, Haukka J, et al. Systematic reviews of observational studies of risk of thrombosis and bleeding in urological surgery (ROTBUS): introduction and methodology. Syst Rev. 2014;3:150.
6. Tikkinen KAO, Craigie S, Agarwal A, Violette PD, Novara G, Cartwright R, et al. Procedure-specific risks of thrombosis and bleeding in urological cancer surgery: systematic review and meta-analysis. Eur Urol. 2018;73(2):242-251.
7. Tikkinen KAO, Craigie S, Agarwal A, Siemieniuk RAC, Cartwright R, Violette PD, et al. Procedure-specific risks of thrombosis and bleeding in urological non-cancer surgery: systematic review and meta-analysis. Eur Urol. 2018;73(2):236-241.
8. Amin AN, Lenhart G, Princic N, Lin J, Thompson S, Johnston S. Retrospective administrative database study of the time period of venous thromboembolism risk during and following hospitalization for major orthopedic or abdominal surgery in real-world US patients. Hosp Pract (1995). 2011;39(2):7-17.
9. Sweetland S, Green J, Liu B, Berrington de Gonzalez A, Canonico M, Reeves G, et al. Duration and magnitude of the postoperative risk of venous thromboembolism in middle aged women: prospective cohort study. BMJ. 2009;339:b4583.
10. Devereaux PJ, Mrkobrada M, Sessler DI, Leslie K, Alonso-Coello P, Kurz A, et al. Aspirin in patients undergoing noncardiac surgery. N Engl J Med. 2014;370(16):1494-503.

15

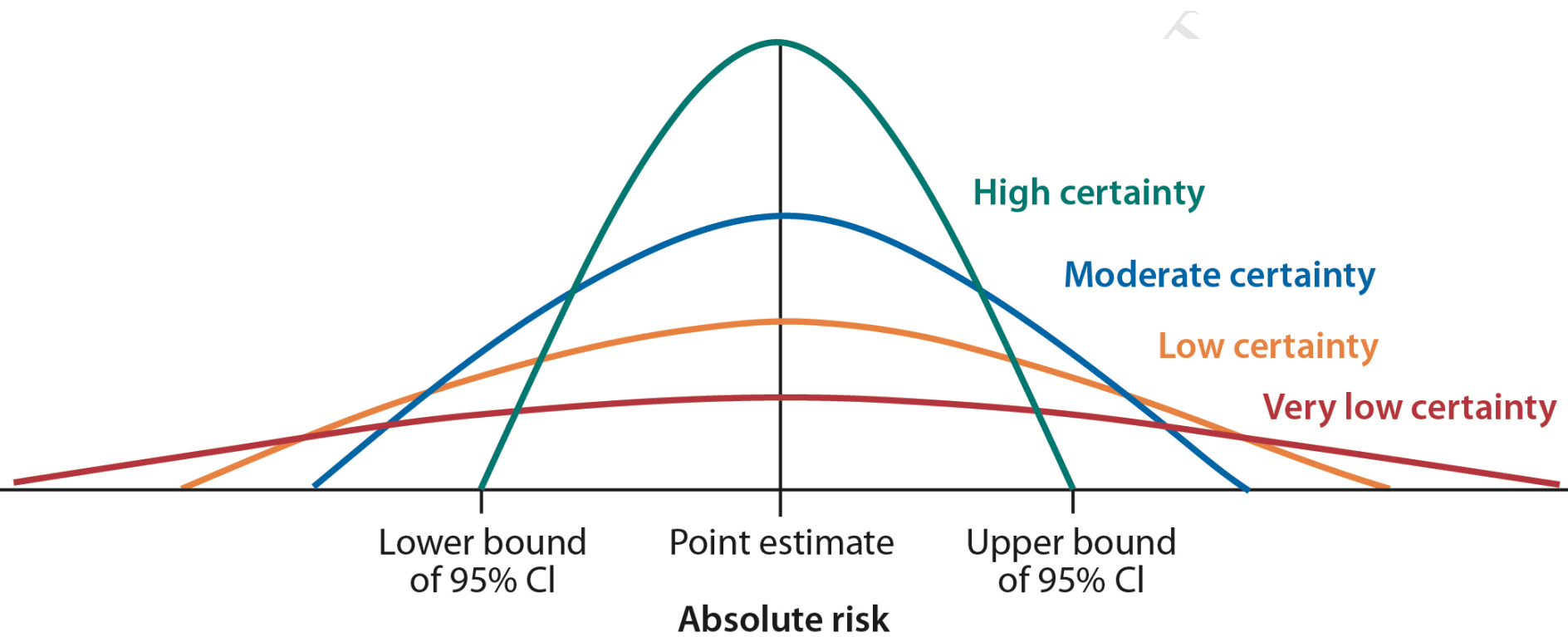**Table 1.** Calculations resulting in certainty ranges.

Procedure	Number of quality domains with serious problems	Initial lowest estimate of the baseline VTE risk	The lowest estimate of the baseline VTE risk adjusted for uncertainty*	Benefit of VTE prophylaxis (decrease in VTE)†	Initial highest estimate of the baseline bleeding risk	Highest estimate of the baseline bleeding risk adjusted for uncertainty*	Harm of VTE prophylaxis (increase in bleeding)†	Lowest plausible net benefit‡	Decision re-rating up§
Open radical cystectomy									
Low risk	One	1.3%	1.0%	0.5%	1.6%	1.9%	0.5%	-0.4%	Do not rate up
Medium risk	One	2.7%	2.1%	1.0%	1.6%	1.9%	0.5%	0.1%	Rate up
High risk	One	5.4%	4.1%	2.1%	1.6%	1.9%	0.5%	1.1%	Rate up
Laparoscopic partial nephrectomy									
Low risk	Two	0.9%	0.8%	0.4%	2.3%	2.6%	0.6%	-0.9%	Do not rate up
Medium risk	Two	1.8%	1.6%	0.8%	2.3%	2.6%	0.6%	-0.5%	Do not rate up
High risk	Two	3.2%	3.0%	1.5%	2.3%	2.6%	0.6%	0.2%	Rate up

* Value after widening certainty interval by 20% for each domain (risk of bias, inconsistency, indirectness, publication bias) with serious problems warranting rating down the certainty of the evidence.
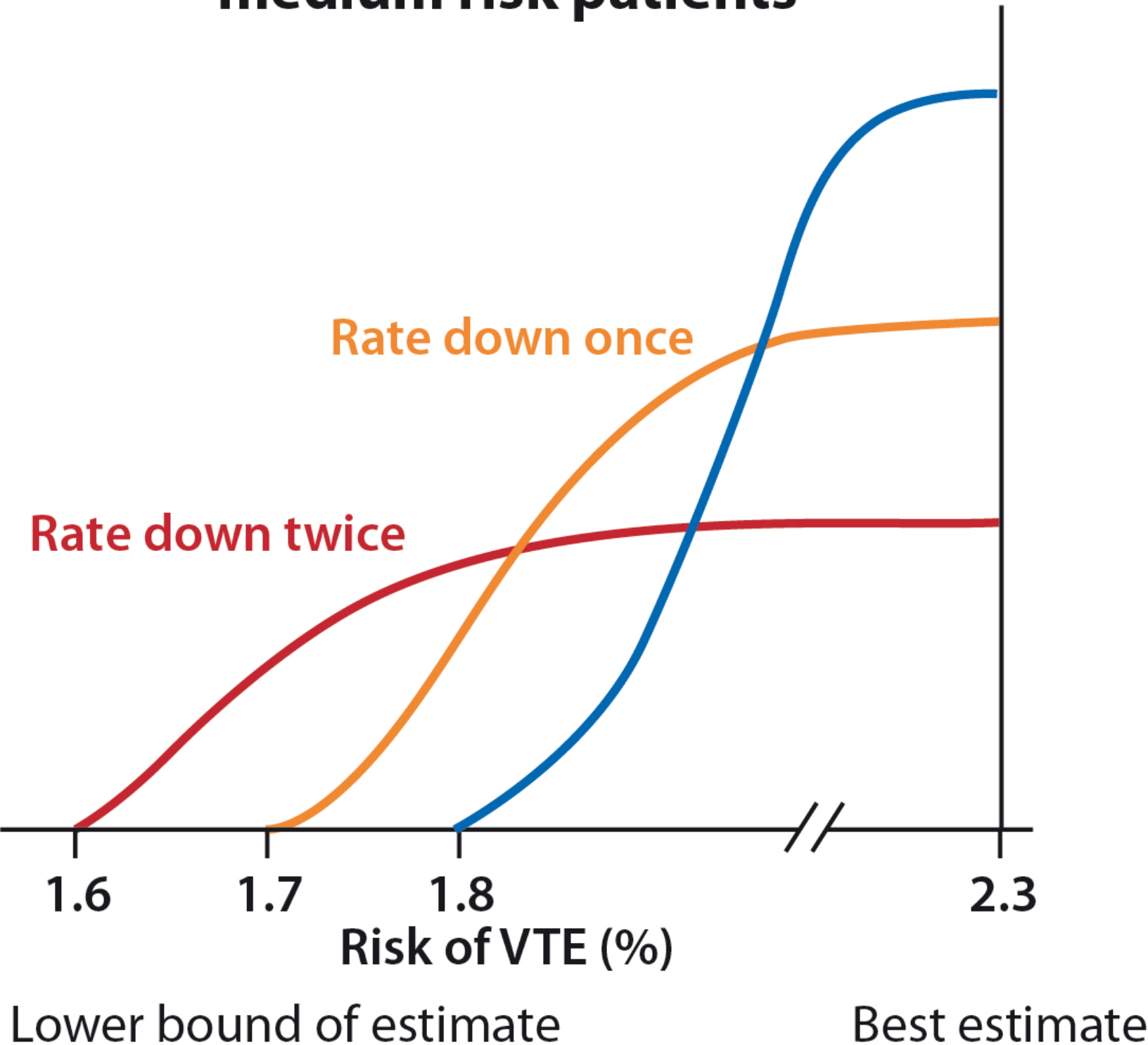
† Our recommendations for pharmacological prophylaxis were based on a starting time of the morning after surgery (4). Approximately 50% of major bleeds occur between surgery and the next morning. In contrast, cumulative risk of VTE during the first four weeks post-surgery is constant (5).

‡ Net benefit is equal to absolute reduction in VTE risk minus absolute increase in bleeding risk (with twice the weight for major bleeding as for VTE). The net benefit is positive when the value of reduced VTE is greater than increased bleeding.

§ Decision to rate up only if lowest plausible net benefit positive; that is, if difference after valuation greater than 0.

High certainty

Moderate certainty

Low certainty

Very low certainty

Lower bound
of 95% CI

Point estimate

Upper bound
of 95% CI

Absolute risk

Laparoscopic partial nephrectomy, medium risk patients

Rate down once

Rate down twice

Risk of VTE (%)

1.6   1.7   1.8   2.3

Lower bound of estimate   Best estimate

Relative Effect Uncertainty

- Risk of bias
- Publication bias
- **Indirectness**
- Imprecision
- Inconsistency

Indirectness:
- Indirect comparison
- Indirect intervention
- **Indirect population**
- Indirect outcome

Indirect population:
- Population differs (e.g. older versus younger) and relative effect may differ across subpopulations

Baseline Risk Uncertainty

- Risk of bias
- Publication bias
- Indirectness
- Imprecision
- Inconsistency