Methods Research Report

Evaluation of the Benefits and Harms of Aspirin for Primary Prevention of Cardiovascular Events: A Comparison of Quantitative Approaches

Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • **www.ahrq.gov**

Methods Research Report

Evaluation of the Benefits and Harms of Aspirin for Primary Prevention of Cardiovascular Events: A Comparison of Quantitative Approaches

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract No. 290-2007-10061-I

Prepared by:
Johns Hopkins University Evidence-based Practice Center
Baltimore, MD

Investigators
Milo A. Puhan, M.D., Ph.D.
Sonal Singh, M.D., M.P.H.
Carlos O. Weiss, M.D., M.H.S.
Ravi Varadhan, Ph.D.
Ritu Sharma, B.Sc.
Cynthia M. Boyd, M.D., M.P.H.

AHRQ Publication No. 12(14)-EHC149-EF
November 2013
Updated February 2014

This report is based on research conducted by the Johns Hopkins University Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. Contract No. 290-2007-10061-I). The findings and conclusions in this document are those of the author(s), who are responsible for its content, and do not necessarily represent the views of AHRQ. No statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, and policymakers, among others—make well-informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report may be used, in whole or in part, as the basis for the development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

This report may periodically be assessed for the urgency to update. If an assessment is done, the resulting surveillance report describing the methodology and findings will be found on the Effective Health Care Program Web site at www.effectivehealthcare.ahrq.gov. Search on the title of the report.

Persons using assistive technology may not be able to fully access information in this report. For assistance, contact EffectiveHealthCare@ahrq.hhs.gov.

None of the investigators have any affiliations or financial involvement that conflict with the material presented in this report.

Suggested citation: Puhan MA, Singh S, Weiss CO, Varadhan R, Sharma R, Boyd CM. Evaluation of the Benefits and Harms of Aspirin for Primary Prevention of Cardiovascular Events: A Comparison of Quantitative Approaches. Methods Research Report. (Prepared by Johns Hopkins University Evidence-based Practice Center under contract No. 290-2007-10061-I). AHRQ Publication No. 12(14)-EHC149-EF. Rockville, MD: Agency for Healthcare Research and Quality. November 2013. Updated February 2014. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies and strategies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base in and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by email to epc@ahrq.hhs.gov.

Richard G. Kronick, Ph.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D.
Director, EPC Program
Agency for Healthcare Research and Quality

Parivash Nourjah, Ph.D.
Task Order Officer
Agency for Healthcare Research and Quality

Acknowledgments

The team would like to thank Manisha Reuben for her assistance with the formatting of the report and Dr. Parivash Nourjah for her valuable insight throughout the project. We would like to thank Dr. Eric Bass for his thoughtful review, and Eric Vohr for his editorial contribution.

Technical Expert Panel

Thank you to the following Technical Expert Panel members for their time and expert guidance:

Elie Akl, M.D.
University of New York Buffalo
Buffalo, NY

Wiley Chan, M.D.
Northwest Permanente, Kaiser Permanente
Portland, OR

Robert L. Kane, M.D.
University of Minnesota
Minneapolis, MN

Thomas A. Trikalinos, M.D., Ph.D.
Tufts Medical Center
Boston, MA

Peer Reviewers

Thank you to the following Peer Reviewers for their time and expert guidance:

Scott Braithwaite, M.D., M.Sc.
NYU School of Medicine
New York, NY

Wiley Chan, M.D.
Northwest Permanente
Kaiser Permanente
Portland, OR

Joshua Cohen, Ph.D.
Tufts University School of Medicine
Boston, MA

Robert L. Kane, M.D.
University of Minnesota
Minneapolis, MN

Jonathan Treadwell, Ph.D.
ECRI Institute Evidence-based Practice Center
Plymouth Meeting, PA

Thomas A. Trikalinos, M.D., Ph.D.
Tufts Medical Center
Boston, MA

Fred Wolf, Ph.D.
University of Washington
Seattle, WA

Evaluation of the Benefits and Harms of Aspirin for Primary Prevention of Cardiovascular Events: A Comparison of Quantitative Approaches

Structured Abstract

Background: Prior work has described various quantitative approaches to the assessment of benefits and harms of medical interventions. Researchers rarely use these approaches in the context of a systematic review.

Objective: Our objectives were to illustrate two quantitative approaches to assessing benefits and harms in the context of a systematic review, and to determine the methodological challenges of applying these approaches in a systematic review.

Methods: We compared the number-needed-to-treat (NNT) and number-needed-to-harm (NNH) approach and the Gail/National Cancer Institute (NCI) approach for assessing the benefits (prevention of myocardial infarction [MI] and ischemic stroke) and harms (excess of hemorrhagic stroke and major gastrointestinal [GI] bleeds) of aspirin for primary prevention of cardiovascular events. We based our main analyses for these two approaches on the treatment effects from a meta-analysis of large primary prevention trials, and the incidence rates from observational studies. We focused on observational studies that were most applicable to our target population—aged 50 to 84 years, living in the United States without evidence of cardiovascular disease or stroke. We obtained relative weights denoting the relative importance of different outcomes (required by the Gail/NCI approach) from literature sources. These sources weighted major stroke nearly twice as much as MI and nearly eight times as much as major GI bleeds.

Results: The NNT and NNH for aspirin declined with increasing age because of the increase in baseline incidence rates for all outcomes across age categories as obtained from observational studies. For example, in men aged 45-54, the NNT was 1,786 person-years of treatment to prevent one MI, and the NNH was 1,344 person-years of treatment to induce one major GI bleed (which corresponds to 5.6 MI prevented and 57.4 GI bleeds induced if 1,000 people are treated with aspirin for 10 years, compared with no aspirin use). For men aged 75–84, the NNT was 511 to prevent one MI and the NNH was 202 to induce one major GI bleed. A sensitivity analysis that considered different baseline incidence rates from randomized trials showed a much higher NNH for GI bleeds because the baseline incidence rate of that outcome was 2–3 times lower than in observational studies.

When we used relative weights, the Gail/NCI approach showed that aspirin caused more benefit than harm in all age categories of men and women. When we weighted outcomes equally in a sensitivity analysis, the harm from aspirin was greater compared with the main analysis because of greater relative weight for GI bleeds. When we weighted stroke as a very important outcome (weight of 1), MI as an important outcome (weight of 0.5), and GI bleed as an unimportant outcome (weight of 0), aspirin was associated with net benefit for all sex and age categories.

When comparing the two approaches in terms of estimates for a single outcome, we found comparable results for the number of people who would have a benefit or harm from treatment as long as the baseline incidence rates and the competing risk (all-cause mortality) were small. When the impact of the competing risk was larger, we found substantial differences between the NNT and NNH and Gail/NCI approaches, even though the baseline incidence rates and treatment effects used were identical.

Conclusion: The assessment of benefits and harms requires careful selection and integration of data from disparate sources, including baseline risks of events without treatment, the effects of treatments on various outcomes, and relative weights of these outcomes. We have illustrated that quantitative approaches are feasible in a specific decisionmaking context—using data from a systematic review of aspirin for primary prevention. Quantitative approaches can yield different results even if input data for baseline risks and treatment effects are identical. Quantitative approaches can be particularly valuable in demonstrating how the expected balance of benefits and harms depends on assumptions about the relative weights of different outcomes.

Contents

Background 1

Methods 2
Specification of the Decisionmaking Context 2
Selection of Data Sources 3
Effect Estimates of Aspirin on Benefit and Harm Outcomes 4
Estimates of Incidence Rates Without Aspirin (Baseline Risks) 4
Relative Weights for Outcomes for Gail/National Cancer Institute Approach 5
Assumptions That Apply to Both Quantitative Approaches to Benefit and Harm Assessment 5
Number Needed to Treat and Number Needed to Harm 6
Sensitivity Analysis for Number Needed to Treat and Number Needed to Harm Approach 7
Gail/National Cancer Institute Approach 7
Sensitivity Analysis for Gail/National Cancer Institute Approach 8

Results 9
Number Needed to Treat and Number Needed to Harm 9
Sensitivity Analysis for the Number Needed to Treat and Number Needed to Harm Approach Using Baseline Incidence Rates From the Trials 10
Gail/National Cancer Institute Approach 11
Comparison of Approaches 13

Discussion 18
Summary 18
Methodological Challenges 19
Selection of Data Sources for Treatment Effect and Baseline Risk Estimates 19
Consideration of the Importance of Outcomes (Relative Weights) 20
Conveyance of Statistical and Nonstatistical Uncertainty 20
Combined Effect of Incidence Rates, Treatment Effects, and Relative Weights 20
Discussion of Principles for Quantitative Approaches for Benefit and Harm Assessment 20
Choice of Quantitative Approaches 22
Clinical Implications 22

Limitations 24

Future Research 25

Conclusion 27

References 28

Acronyms/Abbreviations 30

Tables

Table 1. Incidence rates (per 1,000 person-years) without aspiring prevention based on surveillance data....4
Table 2. Assumptions....5
Table 3A. Number needed to treat and number needed to harm for aspirin for primary prevention in men....9
Table 3B. Number needed to treat and number needed to harm for aspirin for primary prevention in women....10
Table 4A. Sensitivity analysis for number needed to treat and number needed to harm for aspirin for primary prevention in men....10
Table 4B. Sensitivity analysis for number needed to treat and number needed to harm for aspirin for primary prevention in women....10
Table 5. Expected number of events without and with aspirin prevention in men....11
Table 6. Expected events without and with aspirin prevention in women....12
Table 7. Benefit harm comparison estimates using the NCI approach....12
Table 8. Comparison of number needed to treat and number needed to harm versus the Gail/NCI approach....15

Appendixes

Appendix A. Details of our Data Sources for Effect Estimates, Baseline Risk of all Four Outcomes, and Relative Weights of Outcomes

Background

Systematic reviews often assess the comparative effectiveness and safety of health care interventions. To be most useful to users, systematic reviews should include estimates of the potential benefits and harms that are important to decisionmakers. Quantitative approaches for the assessment of benefits and harms may enhance, support, and facilitate how decisionmakers use systematic reviews.

Previously, we prepared a report on the challenges and principles of assessing benefits and harms of medical interventions, the influence of values and preference, and the key characteristics of quantitative approaches to benefit and harm assessment.[1] That report identified 16 quantitative approaches for assessing benefits and harms. Researchers and methodologists developed several of these approaches using data from a single study, but these approaches could be used in systematic reviews. Reviewers from the Cochrane Collaboration have routinely used simpler approaches, such as the number needed to treat (NNT) and number needed to harm (NNH). Decisionmaking contexts that have a larger number of relevant benefit and harm outcomes may need more complex approaches. However, we have limited understanding of the comparative strengths and limitations of quantitative approaches to benefit and harm assessment because little work has been done to compare empirical applications of these approaches.

The specific objectives of this report were:

- To illustrate two quantitative approaches to benefit and harm assessment in the context of a systematic review; and
- To evaluate the methodological challenges of applying the two quantitative approaches to benefit and harm assessment in a systematic review.

Methods

To select the clinical question to illustrate the approaches to benefit and harm assessment, we reviewed systematic reviews from the www.effectivehealthcare.org Web site and investigators' reference libraries.[2-9] These systematic reviews addressed a wide variety of clinical questions. We selected the systematic review of the benefits and harms of low-dose aspirin for primary prevention of cardiovascular events. We chose this complex decisionmaking scenario because it needs to balance the multiple outcomes for benefit (prevention of myocardial infarction [MI], ischemic stroke) and harm (excess of gastrointestinal [GI] bleeds and hemorrhagic stroke).[9] Such a complex scenario allowed us to explore how variations in baseline risk across these outcomes (risk of an outcome over a particular time period without treatment, i.e., without aspirin) and variations in their relative importance and relevance to patients (i.e., preferences) may affect the comparison of benefits and harms of aspirin. We also considered it a pragmatic choice because data were available for both benefit and harm outcomes from meta-analyses.

Although one can potentially apply several quantitative approaches to a key question, we illustrated two quantitative approaches for the assessment of benefits and harms of aspirin for primary prevention of cardiovascular events: NNT and NNH approach, and the Gail/National Cancer Institute (NCI) approach. We outlined the selection process for choosing a particular quantitative approach in a previous report.[1] We based this decision on the number of outcomes available and the need for a benefit and harm comparison metric. We also wanted to compare a more commonly-used method which does not provide a benefit and harm comparison metric with another advanced method that provides a benefit and harm comparison metric to understand the assumptions, similarities, and differences in methods. Also, we selected the NNT and NNH approach because these are frequently-used metrics to judge the balance of benefits and harms, and therefore may often be available to decisionmakers. Systematic reviews conducted by the Cochrane Collaboration routinely report NNT and NNH separately. The NNT and NNH approach provides a natural point of reference for more complex approaches. We selected the Gail/NCI approach because it can consider multiple outcomes, it offers a benefit and harm comparison metric,[*] it can consider competing risks, and it can take into account relative weights for various outcomes.[10]

Specification of the Decisionmaking Context

We specified the decisionmaking context as the perspective of a hypothetical guideline-maker assessing the benefits and harms of aspirin in the primary prevention of cardiovascular events. We defined the target population as age 50 to 84 years, living in the United States without evidence of cardiovascular disease or stroke. We assessed the use of low dose (75-100 mg) aspirin compared with placebo using a time horizon of 10 years. The outcomes of interest were defined in terms of potential benefits of treatment in preventing MI and ischemic stroke, and potential harms of aspirin treatment including excess hemorrhagic stroke and major GI bleeds. We defined the unit of NNT as the number of person-years of exposure to treatment needed to prevent one event. We considered death from any cause as a competing risk for the Gail/NCI approach,11 which accounted for the fact that people who die (from any cause) cannot experience any of the events of interest (MI, stroke, or GI bleed) later on.

[*] A benefit and harm comparison metric provides explicit quantitative information on the benefit and harm outcomes, such as by putting benefit and harm on the same scale (e.g., quality-adjusted life-years (QALYs), probability scale, risk scale, NNT/NNH ratio, etc.), resulting in a benefit and harm comparison estimate.[1]

Selection of Data Sources

The data inputs needed for a quantitative approach to benefit and harm assessment included:

a. The estimated effects of aspirin on the outcomes of MI, ischemic stroke, hemorrhagic stroke, and GI bleed;
b. The baseline risk of these outcomes without aspirin for a time horizon of 10 years; and
c. The relative weights for these outcomes. (Certain quantitative approaches for benefit and harm assessment, such as the NNT and the NNH, do not require an explicit weight for these outcomes and rely on decisionmakers to weigh the outcomes.)

There is general scientific consensus that high-quality randomized controlled trials (RCTs) and meta-analyses provide a reliable source of evidence on treatment effects of interventions, including benefits and harms. However, sometimes data on harm may not be available from RCTs, and researchers will need to retrieve data from additional sources including observational studies. Retrieving and selecting the appropriate estimates for the probabilities of benefits and harms without treatment may require additional data sources beyond RCTs. This may be particularly true when the target population is dissimilar from the trial population with respect to characteristics that influence baseline risks (e.g., age or comorbidities). An important reason to go beyond RCTs and to look to observational studies is that baseline risks, as observed in RCTs, are often not applicable to the target population of a benefit and harm assessment because of restrictive eligibility criteria. Additional data sources may include large observational studies or national surveillance data. Occasionally, when the trial populations reflect the baseline risk (for both benefit and harm outcomes) of the target population of intended users, RCTs can provide a useful source of information on baseline risk without treatment. Researchers often call any measure of the occurrence of an outcome "risk" even when they have calculated a rate. In this report we also use the term baseline risk. For all calculations in this report we refer to the actual metric used, which were primarily incidence rates (number of events per person-time) rather than cumulative incidence (i.e., the cumulative number of participants with an event over a particular period of time divided by all participants that the study observed).

We followed a transparent selection process. We prespecified that the benefit and harm estimates should be applicable to the population of interest. We relied on a previously conducted meta-analysis of RCTs for the effect estimates of aspirin on the outcomes of MI, ischemic stroke, hemorrhagic stroke, and GI bleeds.[9] Additionally, we searched for data sources that described the incidence rates of cardiovascular events and GI bleeds among the target population in cohort studies.[12] We expected that incidence rates may be different in the RCTs of the meta-analysis as compared with those in the observational studies. For example, if people with previous GI bleeds were excluded from RCTs, the incidence rates as observed in placebo groups of RCTs would underestimate the incidence rates in the target population (age 50 to 84 years, living in the United States without evidence of cardiovascular disease or stroke).

Some quantitative approaches for benefit and harm assessment explicitly use relative weights for various outcomes.[1] RCTs do not provide selection of relative weights for these outcomes (i.e., how important the outcomes are relative to one another as perceived by patients). Additional evidence from surveys using conjoint analysis or other preference-eliciting techniques can provide information on relative weights of outcomes. We searched for studies that elicited patient preferences for MI, stroke (ischemic and hemorrhagic) and GI bleeds to weight the outcomes. No established search filters exist to identify studies on patient preferences and this literature is not indexed consistently. We searched MEDLINE with the PubMed interface using the medical subject headings "patient preference," "aspirin," and "cost-benefit analysis," and

used the "related articles" function to search for potential studies. We selected a study that measured the preferences of participants that had not experienced these events (primary prevention population). We relied on a study conducted in the primary prevention population,[13] because the relative weights assigned by participants who have not experienced events may be different from those who have experienced such events.

The Appendix shows the details of our data sources for effect estimates, baseline risk of all four outcomes, and relative weights of outcomes.

Effect Estimates of Aspirin on Benefit and Harm Outcomes

We used the effect estimates from an updated meta-analysis of aspirin for primary prevention for our relevant decisionmaking context.[9] This study reported a relative risk (RR) of 0.86 (95% confidence interval [CI], 0.74-1.00) for MI, a RR of 0.87 (95% CI, 0.73-1.02) for ischemic stroke, a RR of 1.35 (95% CI, 1.01-1.81) for hemorrhagic stroke, and a RR of 1.62 for GI bleeds (95% CI, 1.31-2.00) for aspirin compared with placebo or controls using a random effects meta-analysis. Since we did not have access to individual time-to-event patient data, we approximated incidence rate ratios (NNT and NNH approach) and hazard ratios (Gail/NCI approach) using the RRs listed above. A previous meta-analysis of a smaller dataset by the same authors formed the basis of recommendations by the U.S. Preventive Services Task Force on the benefits and harms of aspirin.[14]

Estimates of Incidence Rates Without Aspirin (Baseline Risks)

We used data from the Atherosclerosis Risk in Communities Study, which is a prospective epidemiologic study conducted in four U.S. communities, to obtain baseline incident rates of MI and stroke without treatment (1987-2001).[12] We chose the Atherosclerosis Risk in Communities Study over other large cohorts such as the Framingham Study or the Cardiovascular Health Study because of its random population-based sampling. Since we were unable to find appropriate population-based estimates of GI bleeds in the U.S. population, we relied on the year 2000 estimates from the General Practice Research Database from the United Kingdom, and the Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria from Spain.[15] That report provided estimates for the prevalence of risk factors for severe GI bleeds (age, sex, and prior GI history, i.e., GI pain, mild or severe ulcer) together with incidence rates for severe GI bleeds for each age, sex, and GI history category. Based on those data, we calculated the incidence of severe GI bleeds for the four age categories of the benefit harm assessment for aspirin (45–54, 55–64, 65–74, and 75–84 years) while assuming that people did not take nonsteroidal and anti-inflammatory drugs. To estimate the risk of death we relied on 2007 estimates from the U.S. Center for Disease Control.[16] Treatment information is not available from the death reports of the Center for Disease Control, so we assumed that the risks of death were for people without treatment. Table 1 shows the baseline incidence rates for untreated men and women. For all outcomes, incidence rates differ according to age categories, and between men and women. Since we did not have hazard rates available, we used the incidence rates (as shown in Table 1) as an approximation of (discrete-time) hazard rates.

Table 1. Incidence rates (per 1,000 person-years) without aspirin based on surveillance data (baseline incidence rates)

Outcomes	Baseline Incidence Rates in Men for Four Age Categories				Baseline Incidence Rates in Women for Four Age Categories			
Age Categories	45-54	55-64	65-74	75-84	45-54	55-64	65-74	75-84
MI	4.0	6.2	9.3	14.0*	1.2	3.0	4.7	8.2
Major ischemic stroke	1.2	2.5	5.6	10.8	0.9	2.0	3.6	7.5
Major hemorrhagic stroke	0.2	0.4	0.8	1.6	0.1	0.3	0.5	1.1
Major GI bleeds	1.2	2.5	4.9	8.0	0.6	1.2	2.3	3.7
All-cause mortality	5.0	10.0	25.0	67.0	3.0	7.0	16.0	48.0

GI = gastrointestinal; MI = myocardial infarction

*No reliable data was available from the Atherosclerosis Risk in Communities Study for men in age category 75-84 years. To estimate the incidence rate we assumed a 50 percent increase from age category 65-74 years based on similar increases in incidences in age category 65-74 to 75-84 in the Framingham Heart Study and the Cardiovascular Health Study.

Relative Weights for Outcomes for Gail/National Cancer Institute Approach

Any assumption about preferences may have limitations given the paucity of data on preferences. We identified a single study of 42 participants that reported on patients' relative weights for various health outcomes with aspirin for primary prevention using a visual analog scale.[13] The participants were free of cardiovascular disease and represented subjects from a primary prevention setting. To assign relative weights for various outcomes, we used data on relative weights for these outcomes from this small sample of participants.[13] The relative weights for stroke (ischemic and hemorrhagic), MI, and GI bleeds were 0.89, 0.45, and 0.20 respectively and we list them in Appendix Table A-3. We considered the average weight (mean) that participants assigned to these outcomes but, for simplicity, we did not consider the variability of assigned relative weights across participants. We solicited electronic input from our Technical Expert Panel on the incorporation of relative weights for the outcomes identified above. We specifically asked the experts to comment on whether they felt the utility estimates of these outcomes were reasonable.

Assumptions That Apply to Both Quantitative Approaches to Benefit and Harm Assessment

The implementation of the NNT and NNH approach and the Gail/NCI approach required some assumptions. Table 2 describes assumptions that apply to both approaches, except relative weights are not required for the NNT and NNH approach.

Table 2. Assumptions that apply to the quantitative approaches to benefit and harm assessment

Subject	Assumptions*
Heterogeneity of Treatment Effects	We assumed that the effect of aspirin on benefit and harm outcomes was the same in all groups (e.g., men and women) on a relative scale based on the results of the 2011 Berger meta-analysis.[9]
Effects by Gender	We assumed that the effects of aspirin on benefit and harm outcomes did not vary by gender based on the results of the 2011 Berger meta-analysis.[9]
Effects of Aspirin over Time	We assumed that the relative risk reductions or relative risk increases of aspirin on benefit and harm outcomes did not change over the time horizon of 10 years.
Baseline Incidence Rates over Time	We assumed that the incidence rate of all outcomes without treatment did not change over the time horizon of 10 years.
Measures of Severity	We assumed that all myocardial infarctions were of the same severity, that all strokes were severe, and that all major bleeds were either severe upper or lower GI bleeds to address inconsistent reporting across the trials.
Risk Profiles	We considered different risk profiles of the population, based on characteristics including age and sex. We did not consider additional characteristics such as race, blood pressure, or cholesterol levels because all outcome incidence rates were not available for such profiles.
Applicability of Effects	We assumed RR reductions in benefits or relative risk increases in harms, from the trials, were applicable to the source populations from which estimates of the baseline risks originated.
Relative Weighting of Outcomes (applies only to Gail/National Cancer Institute approach)	To assign relative weights for various outcomes, we used data on preferences for these outcomes from a sample of participants considering aspirin for primary prevention.[13] We assumed that the average weight assigned to these outcomes reflected the relative weights of our target population.

GI = gastrointestinal; RR = relative risk

*The assumptions above are common to both approaches, except for weighting of outcomes, which does not apply to the NNT and NNH approach.

Number Needed to Treat and Number Needed to Harm

We calculated the NNT for MI and ischemic stroke and the NNH for hemorrhagic stroke and GI bleeds (95% CI) with aspirin, by applying the RR for ischemic stroke, hemorrhagic stroke, and GI bleeds from the above meta-analysis[9] to the population event rates for these respective outcomes from the observational studies.[12 15] (Table 1) We used Visual Rx, version 3.0, an online NNT and NNH calculator provided by the editor of the Airways Group of the Cochrane Collaboration.[17,18] We estimated NNTs and NNHs for various age and sex specific categories.

The NNT is the number of person-years of treatment with aspirin needed for one patient to be protected from MI or ischemic stroke, when compared with no aspirin treatment. The NNH is the number of person-years of treatment with aspirin needed for one additional patient to be harmed by a bleeding event, when compared with no aspirin treatment. It can be argued that the terms "number needed to treat to benefit" and "number needed to treat to harm" may be more accurate. We decided to use the more conventional term NNT when we refer to outcomes that are benefits, as opposed to harms. We estimated the NNT and NNH with aspirin in patients with different baseline incidence rates of cardiovascular events and bleeds, as the NNT and NNH varies when aspirin is used in a general population versus highly-selected trial participants.[19] To better account for time at risk, we defined the unit of NNT or NNH as the number of person-years of exposure to treatment needed to prevent one event.

Sensitivity Analysis for Number Needed to Treat and Number Needed to Harm Approach

In a sensitivity analysis, we varied the source of data for baseline incidence rates of cardiovascular events and GI bleeds. We used sex-specific incidence rates from the control groups of trials for MI, stroke, and GI bleeds instead of surveillance data from observational studies (Appendix Table A-4).[14]

Gail/National Cancer Institute Approach

The Gail/NCI approach considers multiple patient-important outcomes of a medical intervention and provides profile-specific estimates of the benefit and harm balance. For example, for a patient of a certain age, sex, and with presence or absence of risk factors for the patient-important outcomes, the Gail/NCI approach provides a benefit and harm comparison estimate that can inform decisionmakers (patients, health care providers, policymakers, payers) about whether treatment will increase or decrease patient-important outcomes over a defined period of time, as compared with receiving no treatment.

In a first step we calculated the number of events for each of the four outcomes per 1,000 subjects over 10 years, using baseline incidence rates as described above, and stratified for age and sex. We calculated the number using equation 1:

(1) $Nx,p = 1{,}000*\{Ix /(Ix+ M)\}*[1-\exp\{-10(Ix+ M)\}]$

In this equation Nx, p is the number of events (N) for a specific outcome (x) per 1,000 subjects over 10 years in subjects without aspirin (p), Ix is the baseline incidence rate of the event (x), and M is all-cause mortality, which we treated as a competing risk and assumed to be equal for both groups (i.e., RR of 1.0 for aspirin versus placebo).

We then calculated the number of events with aspirin for each of the four outcomes per 1,000 subjects over 10 years, again stratified for age and sex. We used equation 2:

(2) $Nx,t = 1{,}000*\{RRx*Ix /(RRx*Ix+ M)\}*[1-\exp\{-10(RRx*Ix+ M)\}]$

In this equation Nx,t is the number of events (N) for a specific outcome (x) per 1,000 subjects over 10 years in subjects with aspirin (t) and RRx represents the RR of aspirin for a specific outcome (x) (derived from the Berger meta-analyses[9]).

We calculated the difference, Nx, in the number of events (N) for a specific outcome (x) per 1,000 over 10 years in subjects with a certain profile between aspirin users and nonusers using equation 3:

(3) Difference $Nx = Nx,p - Nx,t$

Finally, we put all four outcomes on a single scale using equation 4:

(4) $\text{Index}(W1, W2, W3) = W1 \Sigma Nx_1 + W2 \Sigma Nx_2 + W3 \Sigma Nx_3$

In this equation the index represents the benefit and harm comparison metric as the sum (Σ) of differences in events (Nx) for each outcome (1, 2, 3, etc.), using relative weights W, which

represent the importance patients attach to certain outcomes in relation to one another. The number of weights depends on how many different weights are used and does not need to equal the number of outcomes. In the original paper about the Gail/NCI approach, for example, the investigators used three different relative weights: 1.0 for very important outcomes, 0.5 for important outcomes, and 0.0 for unimportant outcomes. The index has a negative value if more events are expected with a treatment (in the case of aspirin it would mean more harm than benefit), compared with no treatment or an alternative treatment. The index has a positive value if the treatment reduces the number of events (e.g., more benefit than harm in the case of aspirin). In the main analysis we used the relative weights as described above, followed by sensitivity analyses.

Sensitivity Analysis for Gail/National Cancer Institute Approach

One critique of quantitative approaches using relative weights is that their selection is arbitrary to some extent, unless there is strong evidence on relative weights from relevant studies (e.g., from primary prevention populations). To address this, we used alternative relative weights. We compared the results using the alternative relative weights to the results incorporating relative weights derived from a study of benefit and harm outcomes of aspirin for primary prevention.[13]

In the first sensitivity analysis, we assigned equal weights to all outcomes (i.e., "equal preference" for all four outcomes). In the second sensitivity analysis, we assigned relative weights just as the Gail/NCI approach did in their example on tamoxifen for the prevention of breast cancer. In this second analysis, we assigned a weight of 1 for very important outcomes (ischemic or hemorrhagic stroke), 0.5 for important outcomes (MI), and 0 for unimportant outcomes (GI bleeds). Thus, the weights in the second sensitivity analysis mean that we ignored GI bleeds as a harm outcome.

In the third sensitivity analysis, we varied the source of data for baseline incidence rates and used incidence rates from the placebo groups of trials for MI, stroke, and GI bleeds instead of surveillance data. Since a 2006 version of the 2011 Berger meta-analysis reported sex-specific incidence rates, we used this earlier publication as the source of baseline incidence rates.[14] We had to use a single baseline incidence rate for each gender since this version did not report baseline incidence rates for different age categories.

Results

Number Needed to Treat and Number Needed to Harm

Tables 3A and 3B show the NNT and NNH associated with use of low-dose aspirin in men and women respectively. Both the NNT and NNH for aspirin consistently declined with increasing age because of the increase in baseline incidence rates for all outcomes across age categories. Apart from the effect of aspirin on outcomes (which we assume is constant on the relative scale here), the increasing baseline incidence rate across increasing age categories affected estimates of NNT and NNH. Among men in the oldest age category, a 35 percent relative increase in the risk of hemorrhagic stroke with aspirin (Appendix Table A-1) still resulted in a low absolute risk of such events (NNH 1,786) because of the low baseline incidence rate of hemorrhagic stroke.

Among women age 45-54 years, a 62 percent relative increase in the risk of major gastrointestinal (GI) bleeds with aspirin (Appendix A) resulted in an NNH of 2,688. The NNH for major GI bleeds for other age and sex specific categories were even smaller because of the relatively higher baseline incidence rates of GI bleeds across these categories.

Since the estimates of some of the rare harmful effects, such as hemorrhagic stroke, were imprecise, we had more uncertainty around the NNT and NNH estimates. The 95% confidence interval (CI) for ischemic stroke included the possibilities of both benefit and harm (relative risk [RR] 0.87; 95% confidence interval [CI], 0.73-1.02) and thus we reported the NNT as well as the NNH, respectively. This illustrates the challenge of conveying sampling uncertainty to end users of the analysis.

Table 3A. Number needed to treat and number needed to harm in person-years for primary prevention with low-dose aspirin in men for four age categories

Age category in years	**45-54**	**55-64**	**65-74**	**75-84**
Benefit outcomes NNT (95% CI)				
MI	1,786 [962 - NA]	1,153 [621 - NA]	769 [414 - NA]	511 [275 - NA]
Major ischemic * stroke	6,411 [NNT 3,087 - NNH 41,667]	3,077 [NNT 1,482 - NNH 20,000]	1,374 [NNT 662 - NNH 8,929]	713 [NNT 343 - NNH 4,630]
Harm outcomes NNH (95% CI)				
Major hemorrhagic stroke	14,286 [6,173 - 500,001]	7,143 [3087 - 250,001]	3,572 [1,544 - 125,001	1,786 [772 - 62,501]
Major GI bleeds	1,344 [833 – 2,688]	645 [400 – 1,290]	329 [204 - 658]	202 [125 - 403]

CI = confidence interval; GI = gastrointestinal; MI = myocardial infarction; NA = not applicable because CIs for RR approximate 1; NNH = number needed to harm; NNT = number needed to treat; RR = relative risk

*The 95% CIs for ischemic stroke (RR 0.87; 95% CI, 073 -1.02) include the possibilities of both benefit and harm and thus we report NNT as well as NNH respectively. The NNT is lower because it reflects the lower limit of the CI, and the NNH is higher because it reflects the upper limit of the CI.

Table 3B. Number needed to treat and number needed to harm in person-years for primary prevention with low-dose aspirin in women for four age categories

Age Category in years	45-54	55-64	65-74	75-84
Benefit outcomes NNT (95% CI)				
MI	5,953 [3,206 - NA]	2381 [1,283 - NA]	1,520 [819 - NA]	872 [470 - NA]
Major ischemic stroke*	8,548 [NNT 4,116 - NNH 55,556]	3487 [NNT 1852 - NNH 25,000]	2,137 [NNT 1,029 - NNH 13,889]	1,026 [NNT 494 - NNH 6,667]
Harm outcomes NNH [95% CI]				
Major hemorrhagic stroke	28,572 [12,346 - 1,000,001]	9,524 [4,116 - 333,334]	5,715 [2,470 - 200,000]	2,598 [1,123 - 90,910]
Major GI bleeds	2,688 [1,667 – 5,376]	1,344 [833 – 2,688]	701 [435 – 1,403]	436 [270 - 872]

* The 95% CIs for ischemic stroke (RR 0.87; 95% CI, 073 -1.02) include the possibilities of both benefit and harm and thus we reported NNT as well as NNH respectively. The NNT is lower because it reflects the lower limit of the CI, and the NNH is higher because it reflects the upper limit of the CI.

CI = confidence interval; GI = gastrointestinal; MI = myocardial infarction; NA = not applicable because CIs for RR approximate 1; NNH = number needed to harm; NNT = number needed to treat; RR = relative risk

Sensitivity Analysis for the Number Needed to Treat and Number Needed to Harm Approach Using Baseline Incidence Rates from the Trials

Below we present the results of the sensitivity analysis for the NNT and NNH approach using baseline incident rates from the trials. Since the baseline incidence rates in the trials were lower than that of the observational studies (Appendix Table A-4), the respective NNTs and NNHs in this sensitivity analysis were higher for all four outcomes for both men and women (Tables 4A and 4B).

Table 4A. Sensitivity analysis for number needed to treat and number needed to harm in person-years for primary prevention with low-dose aspirin in men

Outcomes	NNT(95% CI)
MI	1,299 [700 - NA]
Major ischemic stroke*	3,664 [NNT 1,764 - NNH 23,810]
	NNH(95% CI)
Major hemorrhagic stroke	9,524 [4,116 - 333,334]
Major GI bleeds	1,793 [1,112 – 3,585]

* The 95% CIs for ischemic stroke (RR 0.87; 95% CI, 073 -1.02) include the possibilities of both benefit and harm and thus we reported NNT as well as NNH respectively. The NNT is lower because it reflects the lower limit of the CI, and the NNH is higher because it reflects the upper limit of the CI.

NA= Not applicable because CIs for RR approximate 1.

NNT = Number needed to treat, NNH = Number needed to harm, CI = confidence interval, MI = Myocardial infarction, GI = Gastrointestinal, RR = Relative risk

Table 4B. Sensitivity analysis for number needed to treat and number needed to harm in person-years for primary prevention with low-dose aspirin in women

Outcomes	NNT (95% CI)
MI	7,143 [3,847 - NA]
Major ischemic stroke*	6,994 [NNT 3,368 - NNH 45,455]
	NNH (95% CI)
Major hemorrhagic stroke	14,286 [6,173 - 500,000]
Major GI bleeds	3,226 [2,000 - 6,452]

* The 95% CIs for ischemic stroke (RR 0.87; 95% CI, 073 -1.02) include the possibilities of both benefit and harm and thus we report NNT as well as NNH respectively. The NNT is lower because it reflects the lower limit of the CI, and the NNH is higher because it reflects the upper limit of the CI.

NA= Not applicable because CIs for RR approximate 1.

NNT = Number needed to treat, NNH = Number needed to harm, CI = confidence interval, MI = Myocardial infarction, GI = Gastrointestinal, RR = Relative risk

Our estimates for the NNT and NNH in person-years differ from those reported by Berger et al. for the outcomes of major adverse cardiovascular events and bleeds in the trials.[9] They reported a NNT of 253 (95% CI, 163-568) to prevent one major cardiovascular event and NNH of 261 (95% CI, 182-476) to cause one major bleed with aspirin over the mean duration of the trials (6.9 years). An important difference is that they used cumulative incidences as the metric for baseline risks (i.e., number of persons with an event during the entire followup period divided by the total number of persons) and did not use incidence rates (number of events per person-time) as we did. Thus, their results are not directly comparable to ours. Also, they used the baseline risk entirely from RCT data (as we did in a sensitivity analysis), which excluded participants with a history of GI events. Finally, they implicitly assumed cardiovascular events and major bleeds were of equal weight, stating that for 1,000 people treated with aspirin over 5 years three cardiovascular events would be prevented and three major bleeds would be induced, compared with no aspirin use. We chose to report separate NNTs for different outcomes, as we considered them to be too dissimilar in terms of their importance to participants.

Gail/National Cancer Institute Approach

Tables 5 and 6 show the absolute numbers of expected events over 10 years without or with low-dose aspirin for 1,000 men or women, respectively. Take, for example, men age 55 to 64 years. The expected number of MIs per 1,000 untreated men over 10 years was 57. This was a little less than what would be calculated by simply multiplying the incidence rate of 6.2 MIs per 1,000 person-years by 10 years (see Table 1) because we considered death as a competing risk that prevents MIs from occurring in those who die early. As age increased, the impact of death became larger. For example, the expected number of GI bleeds in untreated men age 75 to 84 years was 56 over 10 years with death as a competing risk. This is considerably lower than the 480 that would be expected without considering death as a competing risk (Table 1).

Table 5. Expected number of events without and with low-dose aspirin in men *

Outcomes	Number of Expected Events Over 10 Years Per 1,000 Men							
	Without aspirin				With aspirin			
Age categories in years	45-54	55-64	65-74	75-84	45-54	55-64	65-74	75-84
MI	38	57	79	96	33	49	68	83
Major ischemic stroke	14	27	55	84	12	24	48	74
Major hemorrhagic stroke	2	4	8	12	3	5	11	17
Major GI bleeds	12	24	42	56	19	38	67	89

GI = gastrointestinal; MI = myocardial infarction; RR = relative risk

*We considered all-cause mortality a competing risk that was equal for both groups (i.e. RR of 1.0).

Tables 5 and 6 also allow for a comparison between the numbers of expected events without and with aspirin in men and women, respectively. For example, the difference in the number of MIs for men age 65 to 74 years was 11 (79 minus 68) over 10 years. This means that 11 MIs were prevented by aspirin, based on a RR reduction of 14 percent from the meta-analysis by Berger et al.[9] Aspirin reduced major ischemic strokes to a similar extent (13 percent RR reduction). At the same time, aspirin increased the number of major hemorrhagic strokes (35 percent RR increase, e.g., excess of three hemorrhagic strokes for men age 65 to 74 years) and the number of major GI bleeds (62 percent RR increase, e.g., excess of 25 major GI bleeds for men age 65 to 74 years) over 10 years.

Table 6. Expected events without and with low-dose aspirin in women *

Outcomes	Number of expected events over 10 years per 1,000 women							
	Without aspirin				With aspirin			
Age categories in years	45-54	55-64	65-74	75-84	45-54	55-64	65-74	75-84
MI	12	29	42	63	10	25	37	54
Major ischemic stroke	10	22	38	66	9	20	33	58
Major hemorrhagic stroke	1	3	6	9	1	4	7	13
Major GI bleeds	6	12	21	29	10	19	34	46

*All-cause mortality is considered as a competing risk and equal for both groups (i.e. relative risk of 1.0).
MI = Myocardial infarction, GI = Gastrointestinal

Table 7 shows the Gail/NCI index for men and women for the main analysis and all three sensitivity analyses. When we used relative weights as reported in the literature[13] (e.g., major stroke was weighted about twice as much as MI and about 8 times as much as GI bleeds), aspirin caused slightly more benefit than harm in all age categories of men and women (i.e. positive Gail/NCI index). When we weighted the outcomes equally (sensitivity analysis I), the harm from aspirin was greater than the benefit (i.e., negative Gail/NCI index). In sensitivity analysis II, where we weighted strokes as very important with a weight of 1, MIs as important with a weight of 0.5, and GI bleeds as unimportant with a weight of 0, aspirin provided net benefit for all sex and age categories.

Table 7. Gail/NCI index to estimate the relative weighted number of harm minus benefit outcomes

	Gail/NCI benefit harm comparison index per 1,000 persons treated over 10 years for four age categories*							
	Men				Women			
Age (years)	45-54	55-64	65-74	75-84	45-54	55-64	65-74	75-84
Main analysis using empirically derived relative weights	2*	3	3	5	1	2	3	5
Sensitivity analysis I, equal preference of outcomes	-1†	-5	-10	-14	-1	-1	-4	-4
Sensitivity analysis II, only very important and important outcomes‡	4§	6	9	13	2	4	6	9
Sensitivity analysis III, based on control group event rates from aspirin RCTs	4**				1			

*Negative values of index= aspirin harmful (harm events [hemorrhagic stroke and severe GI bleeds] outweigh benefit events [myocardial infarctions and severe ischemic stroke])
†The values for the index are higher in sensitivity analysis I than in the main analysis and sensitivity analyses II and III because severe GI bleeds have more weight in sensitivity analysis I (i.e. relative weight of 1.0) compared to the other analysis (0.2 in the main analysis and sensitivity analysis III, and 0 in sensitivity analysis II).
‡Weight of 1 for very important outcomes such as severe stroke, 0.5 for important outcomes such as myocardial infarction and 0 for gastrointestinal bleeds
§Positive values of index = aspirin beneficial (prevented benefits outweigh harm events)
**Mean age across RCTs 53 to 65 years for men and 55 to 65 years for women. Since no age-specific incidence rates were available (Table 1), benefit and harm comparison estimates are not available for the four age categories.

Finally, when we based the analyses on baseline incidence rates of the placebo groups in the trials (sensitivity analysis III), aspirin use was associated with more benefit than harm. For sensitivity analysis III, we did not make age-specific benefit and harm comparisons because trials only reported baseline incidence rates for the entire male or female trial populations. Baseline incidence rates in the trials were comparable to those from surveillance data, with the important exception of GI bleeds (see next paragraph). The baseline incidence rate of MI in the trials was 5.6 per 1,000 person-years for men and 1.0 per 1,000 person-years for women. Compared with the baseline incidence rates in Table 1, the trial-based incidence rates compared

well with those observed in the same age categories (45 to 54 and 55 to 64 years). We observed similar age-specific baseline incidence rates for major ischemic strokes (2.1 per 1,000 person-years for men and 1.1 per 1,000 person-years for women), major hemorrhagic strokes (0.3 per 1,000 person-years for men and 0.2 per 1,000 person-years for women), and for death (7.9 per 1,000 person-years for men and 3.5 per 1,000 person-years for women). Thus, the difference in the expected number of MIs, major strokes, and deaths for 1,000 men and women with or without aspirin treatment was similar, regardless of whether we based baseline incidence rates on surveillance data or trial data.

The baseline incidence rates for GI bleeds differed when comparing trial data with population-based surveillance data. In the trials, the baseline incidence of GI bleeds was lower (1 per 1,000 person-years for men and 0.5 per 1,000 person-years for women) than in the population-based surveillance data (by a factor of about 2). As a consequence, using the baseline incidence rates of GI bleeds from trials, the number of GI bleeds caused by aspirin was lower than found in population-based surveillance data; and, as a result, the benefit and harm comparison estimates were positive and thus favored aspirin.

Comparison of Approaches

Table 8 summarizes the characteristics of the NNT and NNH approach and the Gail/NCI approach and allows for a comparison between the two approaches. We show examples where we used baseline incidence rates for men aged 45-54 years (example 1) and men aged 75-84 years (example 2). The NNT approach estimates the NNT directly using equation 6 (see Table 8). The Gail/NCI approach first estimates the number of expected events while considering mortality (or any other outcome) as a competing risk (equations 1 and 2 above and in Table 8), and then takes the difference in events between treated and untreated subjects (equation 3). One can easily transform the NNT based on incidence rates to express the number of events per person-time or per a certain number of persons treated over a specific time periods, respectively (equation 7), and thereby facilitate comparison with estimates from the Gail/NCI approach. Thus, although the target estimates researchers commonly use differ markedly between the NNT and NNH approach and the Gail/NCI approach, equation 7 for the NNT and NNH approach (Table 8) shows that we can easily make the results comparable. The numerical comparison of the NNT and NNH approach with the Gail/NCI approach is restricted to a single outcome since NNT only deals with one outcome at a time whereas the Gail/NCI approach provides an index to summarize the effects of treatment across all outcomes. Thus, the numerical examples presented in Table 8 do not compare the NNT and NNH approach with the full Gail/NCI approach.

Looking at example 1, where both the baseline incidence for MI and the risk for mortality is low (among men age 45-54 years), the number of MIs prevented per 1,000 men treated with aspirin over 10 years, compared with no aspirin, is 5.6 according to the NNT approach and 5.3 according to the Gail/NCI approach. The impact of the competing risks is small (difference of 5.6 and 5.3, respectively). Therefore, the NNT and NNH approach, which does not consider competing risks, provides a good approximation of the number of events prevented or in excess for treated versus untreated subjects. In contrast, example 2 illustrates the effect of a larger competing risk, in the setting of a still relatively low baseline incidence rate for MIs. In men age 75-84 years, the difference between the NNT approach and Gail/NCI approach is 6.9 (19.6 minus 12.7) prevented MIs per 1,000 men treated with aspirin over 10 years compared with no aspirin.

As mentioned above, the NNT and NNH approach provides separate estimates for each outcome. In contrast, the Gail/NCI index provides an index that summarizes the (weighted) number of events prevented or in excess by treatment across outcomes (equation 4 in Table 8) per a certain number of persons treated over a specific time horizon. Some of the input parameters needed (e.g., treatment effect and baseline risk estimates) are the same for both approaches. However, only the Gail/NCI approach considers relative weights reflecting the importance of the different outcomes. The assumptions underlying both approaches are similar; we noted the differences in Table 8.

Table 8. Comparison of the number needed to treat and number needed to harm approach and the Gail/National Cancer Institute approach*

Characteristics to be Compared	Number Needed to Treat and Number Needed to Harm (Unit= Person Years)	Gail/National Cancer Institute
Model: mathematical equations	Based on constant incidence rate model Equation 5: (NNT general formula) (5) NNT = 1/Absolute Risk Reduction Equation 6: NNT formula for incidence rate (number of person-years of treatment needed to prevent one event) (6) NNT = 1/[Ix – Ix*RRx)] Equation 7: Difference in the number of expected events per 1,000 subjects without and with aspirin over 10 years (7) Difference Nx = 1,000/{1/[10*Ix – 10*Ix*RRx)]} Ix = Baseline incidence rate for a specific outcome (x) RRx = RR of aspirin for a specific outcome (x) Nx = Difference in the number of events (N) for a specific outcome (x) per 1,000 subjects without and with aspirin over 10 years)	Based on constant hazard rate exponential model Equation 1: (Number of expected events per 1,000 subjects without aspirin over 10 years) (1) Nx,p = 1,000*{Ix /(Ix+ M)}*[1-exp{-10(Ix+ M)}] Please note that the numbering of equations in this table is consistent with the numbering of equations in the text Equation 2: (Number of expected events per 1,000 subjects with aspirin over 10 years) (2) Nx,t = 1,000*{RRx*Ix /(RRx*Ix+ M)}*[1-exp{-10(RRx*Ix+ M)}] Equation 3: (Difference in the number of expected events per 1,000 subjects without and with aspirin over 10 years) (3) Difference Nx =Nx,p - Nx,t Equation 4: (Sum of differences in events per 1,000 subjects without and with aspirin over 10 years) (4) Index(W1, W2, W3)=W1 Σ Nx_1 + W2 Σ Nx_2 + W3 Σ Nx_3 Nx, p = number of events (N) for a specific outcome (x) per 1,000 subjects over 10 years in subjects without aspirin (p) Ix = Baseline incidence rate for a specific outcome (x), approximates hazard rates M = all-cause mortality Nx, t = number of events (N) for a specific outcome (x) per 1,000 subjects over 10 years in subjects with aspirin (t) RRx = RR of aspirin for a specific outcome (x), approximates hazard ratio Nx = Difference in the number of events (N) for a specific outcome (x) per 1,000 subjects without and with aspirin over 10 years)

Table 8. Comparison of the number needed to treat and number needed to harm approach and the Gail/National Cancer Institute approach* (continued)

Characteristics to be Compared	Number Needed to Treat and Number Needed to Harm (Unit= Person Years)	Gail/National Cancer Institute
Example 1	**Low baseline incidence rate for MI in men age 45-54 years and low competing mortality risk** MI: Ix = 0.004 per person-year; RRx aspirin vs. placebo: = 0.86; All-cause mortality: M = 0.005 per person-year	
	Equation 7: (Difference in the number of expected events per 1,000 subjects without and with aspirin over 10 years) Difference Nx = 1,000/{1/(10*Ix – 10*Ix*RRx)} = 1,000/178.6 = 6 events (5.6 with one decimal)	Equation 3: (Difference in the number of expected events per 1,000 subjects without and with aspirin over 10 years while considering death as competing risk) Difference Nx =Nx,p -Nx,t = 38-33 = 5 events (5.3 with one decimal)
Example 2	**Low baseline incidence rate for MI in men age 75-84 years and moderate competing mortality risk** MI: Ix = 0.014 per person-year; RRx aspirin vs. placebo: = 0.86; All-cause mortality: M = 0.067 per person-year	
	Equation 7: (Difference in the number of expected events per 1,000 subjects without and with aspirin over 10 years) Difference Nx = 1,000/{1/(10*Ix – 10*Ix*RRx)} = 1,000/51.0 = 20 (19.6 with one decimal)	Equation 3: (Difference in the number of expected events per 1,000 subjects without and with aspirin over 10 years while considering death as competing risk) Difference Nx =Nx,p -Nx,t = 96-83 = 13 events (12.7 with one decimal)
Target estimate (target of inference)	The NNT is the number of person-years of treatment with aspirin, rather than with comparator, for one patient to be protected from MI or ischemic stroke. The NNH is the number of person-years of treatment with aspirin, rather than with placebo or comparators, for one additional patient to be harmed by an adverse bleeds event	Sum of differences in the (weighted) number of harm (GI bleeds or hemorrhagic stroke) and benefit (MI and ischemic stroke) outcomes for aspirin compared with placebo per 1,000 men/women over 10 years
Input parameters: Baseline incidence rate for benefit and harm outcomes	Baseline incidence rates (per 1,000 person-years) without aspirin based on surveillance data in Table 1 in report	Baseline incidence rates (per 1,000 person-years) without aspirin based on surveillance data in Table 1 in report
Input parameters: Treatment effect	MI: RRx = 0.86 [0.74,1.00] Ischemic stroke: RRx = 0.87 [0.73,1.02] Hemorrhagic stroke: RRx =1.35 [1.01,1.81] GI bleed: RRx= 1.62 [1.31,2.00]	MI: RRx = 0.86 [0.74,1.00] Ischemic stroke: RRx = 0.87 [0.73,1.02] Hemorrhagic stroke: RRx =1.35 [1.01,1.81] GI bleed: RRx= 1.62 [1.31,2.00]
Input parameters: Relative Weights/utilities	NA	MI: Relative weight = 0.45 Ischemic stroke: Relative weight= 0.89 Hemorrhagic stroke: Relative weight= 0.89 GI bleed: Relative weight=0.20

Table 8. Comparison of the number needed to treat and number needed to harm approach and the Gail/National Cancer Institute approach* (continued)

Characteristics to be Compared	Number Needed to Treat and Number Needed to Harm (Unit= Person Years)	Gail/National Cancer Institute
Input parameters: Competing risks	NNT and NNH is unable to account for competing risk	Competing risk of mortality based on population-based mortality rates stratified for age and gender.
Input parameter: Time	Person-years of exposure	Estimated for 10 years
Assumptions: baseline risk over time	Incidence rate of all outcomes without treatment did not change over time.	Incidence rate of all outcomes without treatment did not change over time.
Assumptions: Treatment effect over time	Aspirin associated RR reductions or RR increases for benefit and harm outcomes did not change over time.	Aspirin associated RR reductions or RR increases for benefit and harm outcomes did not change over time. RR assumed to approximate hazard ratio.
Assumptions: weight/utilities	NNT and NNH allow end users to put respective relative weights on outcomes.	Relative weights for outcomes came from a small sample of participants considering aspirin for primary prevention[13]
Assumptions: competing risks	NNT and NNH cannot account for competing risk.	Competing risks are assumed to reflect the risk of an untreated population.
Assumptions: Applicability	RR reductions in benefits or RR increase in harms from the trials were applicable to the source populations from which estimates of the baseline risks originated.	RR reductions in benefits or RR increase in harms from the trials were applicable to the source populations from which estimates of the baseline risks originated.
Assumptions: Risk profiles	This approach considered different risk profiles of the population, based on characteristics including age and sex. It did not consider additional characteristics such as race, blood pressure, or cholesterol levels because not all outcome incidence rates were available.	This approach considered different risk profiles of the population, based on characteristics including age and sex. It did not consider additional characteristics such as race, blood pressure, or cholesterol levels because not all outcome incidence rates were available.

CI = confidence interval; GI = gastrointestinal; MI = myocardial infarction; NA = not applicable; NNH = number needed to harm; NNT = number needed to treat; RR = relative risk

*The numbering of equations in this table is consistent with the numbering of equations in the text.

Discussion

Summary

Users of systematic reviews are challenged in many ways to make optimal use of complex information. The critical pieces of information in a systematic review are the relative benefits and harms of the treatment options that are compared. Most reviews summarize these separately, and leave it to the user to make sense of it all. Reviewers can either use a qualitative assessment of benefits and harms or consider quantitative approaches. Quantitative approaches can include simpler approaches such as the number needed to treat (NNT) and number needed to harm (NNH) approach. Systematic reviewers can also consider a quantitative combination of benefits and harms, such as the Gail/National Cancer Institute (NCI) approach. Our objectives were to illustrate two quantitative approaches and to evaluate the methodological challenges of applying the two quantitative approaches to benefit and harm assessment in a systematic review.

We illustrated that it is feasible to conduct quantitative benefit and harm assessment in the context of a systematic review of aspirin for primary prevention. Both quantitative approaches showed a comparable excess of gastrointestinal (GI) bleeds caused by aspirin, compared with the number of myocardial infarction (MI) and ischemic strokes prevented. The results appeared slightly favorable for aspirin for primary prevention in the primary analyses and most sensitivity analyses. These results were driven by the relatively higher baseline risks for MI and GI bleeds compared with the lower baseline risks for ischemic stroke and hemorrhagic stroke. The results also were affected by the relatively higher magnitude of treatment effect on the outcomes of hemorrhagic stroke and GI bleeds compared with the smaller effect on the outcomes of MI and ischemic stroke.

The sensitivity analyses that used baseline risks obtained from trials showed that the incidence of some outcomes (GI bleeds) in trials can be different from the incidence observed in observational studies, and that this affects the results of quantitative approaches. Our analyses also showed that since patient-profile-specific baseline risks vary, the characteristics of the target population, such as age and gender, can substantially modify the results of quantitative approaches to benefit and harm assessment. Finally, the sensitivity analyses for the Gail/NCI approach showed that the relative weights used to reflect the importance of different outcomes can have an impact on the results of a quantitative approach to benefit and harm assessment. An important conclusion from these findings, which we will further discuss below, is that assessors of benefits and harms need to carefully choose the three different sources of data, i.e. estimates for treatment effects, baseline risks, and relative weights, and their choices could affect the conclusions drawn from systematic reviews of the corresponding evidence.

The comparison of the two approaches we have chosen for illustration (NNT and NNH, and Gail/NCI) showed that the results for single outcomes (see examples in Table 8) may or may not differ. To ensure comparability, we used identical baseline incidence rates, treatment effect estimates, time horizons, and number of persons treated. We also made the assumptions for both approaches that baseline risks and treatment effects do not change over time. Also, we transformed the typical NNT to the metric used by Gail/NCI (difference in number of events between treated and untreated subjects over a certain period of time) to increase comparability of the actual results.

However, the two approaches have two inherent differences, namely the Gail/NCI approach includes competing risks and patient preferences (relative weights) for outcomes. The examples in Table 8 showed that even if input data for baseline risks and treatment effects are identical, the

inclusion of competing risks has a substantial impact on the results when the competing risk is frequent in absolute and relative terms compared with the outcomes of interest (see example 2 in Table 8). We illustrated the impact of different patient preferences in the sensitivity analyses of the Gail/NCI approach. If the relative importance of the relevant types of outcomes differ (as may be the case for GI bleeds and severe stroke), the relative weights used have an impact on the results of a benefit and harm assessment. For the NNT and NNH approach and the Gail/NCI approach, we found little difference in the number of prevented events between the aspirin and placebo groups when the incidence rate for the outcome of interest (e.g., MI) and the competing risk were low. If the competing risk was moderate to high, the simpler NNT and NNH approach did not provide a good approximation of the more complex Gail/NCI approach to estimate the number of prevented or excess events per person year.

Methodological Challenges

In the following, we discuss important methodological challenges inherent to quantitative benefit and harm assessment in the context of systematic reviews.

Selection of Data Sources for Treatment Effect and Baseline Risk Estimates

The conduct of a quantitative approach to an assessment of benefits and harms requires integration of data from disparate sources, including individual risks of events without treatment, the effects of treatments on various outcomes, and the relative weights of the outcomes. [1] The latter is discussed in the next methodological challenge. Our results illustrate how the selection of data sources can drive the results of the quantitative approaches.

For instance, if we want to evaluate the balance of benefits and harms for a target population that is quite different from the trial participants, then baseline risks from the placebo arm might not be appropriate. This is illustrated by the effect of varying the baseline risk for GI bleeds in the sensitivity analyses for both the Gail/NCI and NNT and NNH approaches. Some trials excluded subjects with prior GI bleeds and thus minimized the risk for trial participants assigned to placebo. As a consequence, the incidence rates for GI bleeds in the trials were at least 2–3 times lower than in the observational studies that were most representative of the target population of our benefit and harm assessment. In reality, GI bleeds are even more frequent than we assumed in our analysis since some people take nonsteroidal anti-inflammatory drugs, which greatly increase the risk of GI bleeds. Patient characteristics are another key driver of baseline risks. For all four outcomes we considered in our report, there were substantial differences across age and gender categories. Our results illustrate that it is important to stratify the benefit and harm analyses whenever valid estimates for baseline risks for different subgroups are available. This, of course, requires large observational studies or registries. When conducting a quantitative approach to benefit and harm assessment in the context of a systematic review, it will be necessary to go beyond systematic reviews of randomized controlled trials (RCTs) because these RCTs may not provide estimates of baseline risks that are applicable to the target population of the benefit and harm assessment.

We assumed that the treatment effects from trials applied to our target populations. However, one must be careful that, in the presence of heterogeneity of treatment effects, the average treatment effect estimate for the trials may not be entirely applicable to the target population. The 2006 meta-analysis by Berger et al., for example, showed some differences in effects of aspirin

in men and women.[14] Although there has been some debate about whether gender modifies the effects of aspirin, we did not consider potential effect modification. Another solution would be to conduct another sensitivity analysis to explore whether the conclusions change when we consider different effect estimates for different subgroups.

Consideration of the Importance of Outcomes (Relative Weights)

The differences between the main results and sensitivity analyses of the Gail/NCI approach illustrate how the assigned relative weights can also have a major impact on results. From the results of the Gail/NCI approach, aspirin is beneficial only when we obtain incidence rates for GI bleeds from the placebo groups of trials or when we ignore GI bleeds (assign a weight of zero). This highlights the importance of considering the source of relative weights that are used in a quantitative approach, and conducting sensitivity analyses to investigate the effect of these relative weights in quantitative approaches.

The issue of relative weights of multiple outcomes is a challenging one for systematic reviewers and end users, whether they report results using a qualitative approach or a quantitative one. For example, considering use of information from NNT and NNH analyses in decisionmaking contexts, weighting many outcomes of varying importance may also be cognitively challenging for decisionmakers. Relative weights may vary across patient profiles. This also highlights the importance of using sensitivity analyses to investigate the effect of these relative weights in quantitative approaches.

Conveyance of Statistical and Nonstatistical Uncertainty

It is difficult to convey statistical and nonstatistical uncertainty when the intervention may result in either a net harm or a net benefit to the population. With regard to the NNT and NNH approach, the confidence interval (CI) refers to a NNT at one end and to a NNH at the other end. Since the type of quantity referred to by the CI changes, it may be difficult for audiences to understand this information. The communication of statistical uncertainty, such as the CIs around ischemic stroke, which reflect the possibility of both benefit and harm, is challenging. Some have argued that the 95% CIs for NNT or NNH are difficult to describe and should not be reported when there is uncertainty around the possibility of both benefit and harm.[20] Considering the Gail/NCI approach, methodological modifications such as probabilistic simulation could incorporate some statistical uncertainty.[10] Another challenge to appropriately estimating statistical uncertainty is, as discussed in our previous report in detail, the correlation among outcomes (joint distribution). Most approaches assume marginal distributions (i.e. statistical independence), which is unlikely to be appropriate in many situations. However, little research has been done so far to estimate the impact of joint distributions of outcomes on the estimates of uncertainty (e.g., 95% CI). Finally, nonstatistical sources of uncertainty for benefit and harm assessments exist, including the validity of estimates on treatment effects, baseline risks, and relative weights, or the extent to which the results from scientific studies can be applied to the target population of a benefit and harm assessment.

Discussion of Principles for Quantitative Approaches for Benefit and Harm Assessment

We compared these quantitative approaches in light of principles that are potentially relevant to the conduct of quantitative approaches for benefit and harm assessment in the context of

systematic reviews.[1] We summarized these in Table 9. We report our assumptions and the underlying decisionmaking context for both approaches. We used an information preserving approach to present our results using the NNT and NNH person-year approach, and presented results using the number of events prevented or in excess for the Gail/NCI approach. We also showed how assessors of benefits and harms can make the results from the two approaches comparable so that end users can appreciate the differences more easily. We could not adhere to certain principles because of the limitations of underlying data. Neither approach could account for the joint distribution of benefits and harms because of the lack of data. The NNT and NNH approach did not account for competing risk. We were unable to assess the strength of evidence around the benefit and harm comparison metric, or convey uncertainty around such assessment because methods for assessing the strength of evidence around a benefit and harm comparison metric have not been developed.

Table 9. Principles for quantitative approaches for benefit and harm assessment

Principle	Adherence With Number Needed to Treat and Number Needed to Harm Approach	Adherence With National Cancer Institute Approach
Identify the Key Potential Benefits and Harms	We identified benefit and harm outcomes from an updated systematic review. We considered all outcomes commonly considered important in guidelines, trials, and observational studies. We did not interview subjects from a primary prevention setting to ask about outcomes.	We identified benefit and harm outcomes from an updated systematic review. We considered all outcomes commonly considered important in guidelines, trials, and observational studies. We did not interview subjects from a primary prevention setting to ask about outcomes
Report the Characteristics and Assumptions of the Selected Quantitative Approaches	We stated assumptions on page 5, Table 2.	We stated assumptions on page 5, Table 2.
State Whether Preferences Were Considered in the Benefit and Harm Assessment, and if so, Describe how These Were Ascertained, and how Variation in Preferences Would Affect the Assessment	NNT and NNH allow end users to put weights on outcomes.	We used weights from a study for aspirin among primary prevention and considered alternative weights (equal weights). We also used weights of 1.0/0.5/0 as done in the original publication of the Gail/NCI model.
Preserve Information When Reporting on Benefits and Harms	We presented results using the person-year approach. We were unable to account for joint distribution of benefits and harms. We were unable to assess competing risk.	We presented results using excess events. We were unable to account for joint distribution of benefits and harms. This considers competing risk of mortality.
Convey Statistical Uncertainty and Uncertainty in the Strength of the Evidence	Table 3 and Table 4 convey statistical uncertainty on outcomes through 95% CI.	There are no estimates of uncertainty (such as 95% CI), but we could assess the probability of the index being positive or negative using Markov modeling.[10]
State how Decisions About Comparisons, Outcomes, Baseline Risks, and Time Horizons Were Made to Increase Transparency and Traceability	See the decisionmaking context on page 2, data sources in Appendix.	See the decisionmaking context on page 2, data sources in Appendix.

CI = confidence interval; NCI = National Cancer Institute; NNH = number needed to harm; NNT = number needed to treat

Choice of Quantitative Approaches

Overall, it is easier to use simpler approaches directly from the results generated from meta-analysis. However these simpler approaches may also have strong assumptions, and therefore defer the task of weighting outcomes for decisionmakers. In contrast, more complex approaches which generate a benefit and harm comparison metric require additional data on baseline risk and relative weights. Both complex and simpler approaches may include the inherent assumption about the applicability of these data to the intended target population. One could base the selection of a quantitative approach on the specific decisionmaking context, the number of relevant outcomes, and the desire for a single benefit and harm comparison metric. One could also pursue a more data-driven approach where one interprets the results for the different outcomes separately (e.g., using NNT and NNH) and only combines them using a more complex approach if the benefit and harm comparison does not clearly favor one of the treatment options. However, such a data-driven approach may confront researchers with arbitrary decisions because in many instances there may not be a clear winner, particularly if one considers patient profiles.

The Gail/NCI approach shows a potential advantage over the NNT and NNH approach in that it not only combines the information from the three data sources but also combines the results in a single number, which we call a benefit and harm comparison metric. If one agrees with the inputs (estimates on treatment effects, baseline risks, and relative weights), the interpretation is relatively straightforward. In our example, a positive summary index means that harms exceed benefits, and a negative number means that benefits exceed harms.

With the NNT and NNH approach, it is more difficult to draw a conclusion because it does not weight the outcomes and it lacks a benefit and harm comparison metric to put the results on a single scale. For the example of aspirin, the NNH estimates are so much lower than the NNT estimates that one may still conclude that aspirin is likely to cause more harm than benefit. If the estimates for NNH and NNT were more similar, the interpretation would become more challenging.

Finally, the examples in Table 8, where we directly compared the two approaches, showed that different approaches may or may not yield similar results. When, for example, competing risks are moderate or high (as in our example 2, where the risk for mortality was about 5 times as high as the incidence rates for MI) one should choose those approaches that consider competing risks.

When planning a quantitative approach to a benefit and harm assessment for a specific decisionmaking context, it is critically important to carefully consider potential data sources, the selection of the quantitative approach, and the planning of sensitivity analyses. One could consider more complex approaches like the Gail/NCI approach if there are multiple outcomes, if there are important competing risks, and if the relative importance of outcomes is likely to differ. The Gail/NCI way of combining the three sources may be easier to describe than the NNT and NNH approach. Interpreting several NNTs and NNHs for decisionmaking requires an informal combination of NNT and NNH. Finally, planning for specific quantitative approaches in the context of systematic reviews has implications for both time and resources (financial and personnel with different types of expertise).

Clinical Implications

Our findings demonstrate the feasibility and utility of such methods, but are not clinically directive. However, we believe a brief discussion, comparing our findings to clinically focused

work, may be useful. Earlier analysis by the United States Preventive Services Task Force (USPSTF) in 2009 showed a benefit of aspirin for people with moderate to high cardiovascular disease risk.[21] However, the analysis assumed a relative risk reduction of about 32 percent in MI from a 2006 meta-analysis[21] of six primary prevention trials.[14] The updated meta-analysis (same authors) reported a smaller relative risk reduction for MI of about 14 percent[9] using the relative risk method. A major difference between the work by the USPSTF and our work is the specific decisionmaking context, and thus the aim of the benefit and harm assessment. A distinction needs to be made between benefit and harm assessment on a population level versus a benefit and harm assessment for individual subjects. Our target of inference was a United States population, age 50 to 84 years without evidence of cardiovascular disease or stroke. We did not intend to make any recommendations, but assessed the benefits and harms of aspirin on a population level. Therefore, we chose to use population-based outcome data from the population-based Atherosclerosis Risk in Communities Study. The USPSTF statement, however, provides recommendations for individuals in practice settings where risk prediction models provide better guidance than population-based statistics. The USPSTF proposed to use risk prediction equations to estimate the absolute 10-year risks of MI.

For clinical practice, the suggestions of the USPSTF to use risk prediction models is attractive because one can estimate the baseline risk for individual subjects. An important caveat, however, is that the performance of risk prediction models depends largely on the specific population where they are used.[11] Even a widely established model like the Framingham prediction model has limited applicability across different populations unless the model is recalibrated carefully.[22] In summary, the differences between the benefit and harm assessments made by the USPSTF and our work highlights the importance of clarity of purpose when conducting a benefit harm and assessment because it affects the selection of appropriate sources of evidence. Furthermore, although the USPSTF analysis used similar baseline risk estimates, it did not consider death as a competing risk or incorporate relative weights and could not explore the effect of a gradient in relative weights across outcomes.

Limitations

Our work has several important limitations. As one would expect, these limitations overlap with the methodological challenges we described above.

A quantitative approach to the assessment of benefits and harms is limited by the quality of reported data. We obtained our estimates for the treatment effects for both benefit and harm outcomes from a meta-analysis of large RCTs of aspirin for primary prevention, which increases our confidence in these estimates. However, the updated meta-analysis included some patients with diabetes and vascular disease who may be at higher risk of cardiovascular events. We could not consider additional characteristics that affect baseline risk such as race, blood pressure, or cholesterol levels because baseline risk for such data were not available from observational studies. Trial incidence rates were comparable for MI and stroke but not for gastrointestinal (GI) bleeds. The observational studies reported a 2–3 times higher risk of GI bleeds than the trials. The possible reasons for this include less rigorous ascertainment of GI bleeds in the trials because it was not a primary outcome in the trials. Heterogeneity in definitions of GI bleeds is possible across trials. Finally, the trials excluded participants at high risk for GI bleeds (e.g., previous GI bleeds due to aspirin or nonsteroidal anti-inflammatory drugs) but the surveillance data included such patients.

For the Gail/NCI approach, we used relative weights obtained from a sample of participants considering aspirin for primary prevention. As described in our previous report,[1] it is possible to use NNT and NNH to incorporate weights from different stakeholders to generate relative value NNTs and relative value NNHs, but we did not do this. For the Gail/NCI approach, we assumed that these perceptions of the importance of MI, stroke, and GI bleeds, as expressed on a visual analog scale, reflected their relative weights for these health states. As reflected by the relative weights we used, irreversible outcomes such as stroke are usually assigned higher weights than reversible outcomes such as GI bleeds. Alternative preference elicitation techniques may lead to different valuations, depending on the patient population (primary vs. secondary prevention), the techniques used (time tradeoff vs. standard gamble vs. conjoint analysis), and the outcomes evaluated. Preference elicitation using either the analytic hierarchy process or conjoint analysis may be more suitable for exploring the heterogeneity of treatment preferences, but was beyond the scope of this project. Benefit and harm estimates remain sensitive to assumptions about the heterogeneity of treatment preferences in the population.[23] Real heterogeneity in patient preferences would be important to inform the decisionmaking process, but frequently we do not know how much variability among preferences exists. Therefore, we conducted sensitivity analyses using equal relative weights for outcomes and different relative weights for outcomes.

Several other assumptions and inputs merit discussion, and relate to the uncertainty of our results. We did not investigate the possibility that benefit and harm balance varies over time, due to lack of individual-level data. Although researchers have proposed methods for estimating NNT and NNH from survival analysis data, access to only summary data from the meta-analysis prevented such estimation.[24] Although we evaluated low-dose aspirin, we included studies of higher doses that researchers tested in large primary prevention trials.

Future Research

Our report suggests several areas for future research. Our prior report suggests a framework for choosing a quantitative approach or quantitative approaches to benefit and harm assessment. Deciding whether a choice of an approach for a specific decisionmaking context was the right one will require further research and perhaps consensus criteria. Also, end users of systematic reviews and end users of the results of quantitative approaches for benefit and harm assessment should participate in methodological research and evaluations of quantitative benefit and harm assessments to ensure that the evidence generated meets their needs. Future studies should evaluate the reliability and consistency of these quantitative approaches to benefit and harm assessment. Researchers should conduct end-user evaluations to determine the comparative utility and additional advantages of quantitative approaches versus a qualitative assessment of the evidence.

While the literature clearly defines methods for selecting randomized controlled trials (RCTs) for inclusion in a systematic review, the appropriate selection of additional data sources, beyond what is typically included in systematic reviews, needs further development. Future research should address how to select the most valid and applicable incidence data,[25] and how to select and rate the quality of preference assessment studies.[26] Future studies could improve the reporting and assessment of heterogeneity of treatment effects in systematic reviews.[27]

Research to understand when a quantitative approach is sufficiently patient-centered is necessary. Although we considered estimates stratified by age and sex, we did not consider additional characteristics such as race, blood pressure, or cholesterol levels, because stratified outcome incidence rates were unavailable for these variables. The challenge will be to balance the need for finely granulated data (e.g., incidence rates stratified for four to five variables) and accurate estimates of incidence. One could consider reporting incidence rates from surveillance studies stratified for these variables. However, limited sample sizes may become a challenge, even in large studies such as the Atherosclerosis Risk in Communities Study. An alternative would be to estimate baseline risks based on risk prediction models such as the Framingham Risk Index. However, outcome predictions may be poorly calibrated if the models were not developed or updated in the population of interest.

It is uncertain how these approaches will perform when outcome data on harms are more sparse and heterogeneous, as is typical in many systematic reviews. Much uncertainty exists about estimates for harms with new, and even newly approved, therapies. For this illustrative example, we did not have a situation in which robust evidence is available for benefit outcomes but little evidence is available for harms, which is typical for many other clinical questions. Such a situation would add a layer of complexity and uncertainty to quantitative approaches to benefit and harm assessments, and is beyond the scope of this report. Future research should investigate methods such as optimal information size to assess explicitly whether statistical power is adequate to detect harms. Future research on quantitative approaches to benefit and harm assessment should investigate how to deal with surrogate outcomes (such as forced expiratory volume in 1 second in chronic obstructive pulmonary disease or glycated hemoglobin in the context of type 2 diabetes mellitus) in benefit and harm assessments. An evaluation of quantitative approaches in the context of surrogate outcomes is needed. In these situations, systematic reviewers and investigators will either have to elicit data on preferences for intermediate outcomes (surrogate outcomes) or make assumptions about linkages between intermediate outcomes and health outcomes (patient-important outcomes).

Future research is needed regarding which methods of preference assessment are most appropriate. We did not evaluate which preference assessment methods are most appropriate; rather we used available relative weights. Future work comparing the elicitation of relative weights for various outcomes using methods such as the analytic hierarchy process or conjoint analysis would also be informative. In the absence of a gold standard method for elicitation of preferences, such research should assess the concordance or discordance of relative weights generated using various methods. Similarly, an important issue is to consider whose preferences a study assesses, and whether there is variability in these preferences across important subgroups or patient profiles.

Future research on these and other quantitative approaches (such as probabilistic simulation and multicriteria decision analysis) should consider appropriate methods of capturing and conveying the uncertainty around the benefit and harm assessment. This uncertainty relates to many of the identified methodological challenges and future research directions described here. Future evaluation should include a comparison of a larger number of quantitative approaches for benefit and harm assessment other than the two presented in this report. In the absence of a gold standard, reliability and consistency in results across various quantitative approaches for assessing benefits and harms may increase our confidence in their results.

Conclusion

The assessment of benefits and harms requires careful selection and integration of data from disparate sources, including baseline risks of events without treatment, the effects of treatments on various outcomes, and relative weights of these outcomes. We have illustrated that quantitative approaches are feasible in a specific decisionmaking context—using data from a systematic review of aspirin for primary prevention. Quantitative approaches can yield different results even if input data for baseline risks and treatment effects are identical. Quantitative approaches can be particularly valuable in demonstrating how the expected balance of benefits and harms depends on assumptions about the relative weights of different outcomes.

References

1. Boyd CM, Singh S, Varadhan R, et al. Methods for Benefit and Harm Assessment in Systematic Reviews. Reviews. Methods Research Report. (Prepared by Johns Hopkins University Evidence-based Practice Center under contract No. 290-2007-10061-I). AHRQ Publication No. 12(13)-EHC150-EF. Rockville, MD: Agency for Healthcare Research and Quality; November 2012.

2. Bennett WL, Wilson LM, Bolen S, et al. Oral Diabetes Medications for Adults With Type 2 Diabetes: An Update. AHRQ Publication No. 11-EHC038-EF. Rockville, MD: Agency for Healthcare Research and Quality; March 2011.

3. Matchar DB, McCrory DC, Orlando LA, et al. Comparative Effectiveness of Angiotensin-Converting Enzyme Inhibitors (ACEIs) and Angiotensin II Receptor Antagonists (ARBs) for Treating Essential Hypertension. AHRQ Publication No. 08-EHC003-EF. Rockville, MD: Agency for Healthcare Research and Quality; November 2007.

4. Bravata DM, McDonald KM, Gienger AL, et al. Comparative Effectiveness of Percutaneous Coronary Interventions and Coronary Artery Bypass Grafting for Coronary Artery Disease. AHRQ Publication No. 08-EHC002-EF Rockville, MD: Agency for Healthcare Research and Quality; October 2007.

5. Coleman CI, Baker WL, Kluger J, et al. Comparative Effectiveness of Angiotensin Converting Enzyme Inhibitors or Angiotensin II Receptor Blockers Added to Standard Medical Therapy for Treating Stable Ischemic Heart Disease. AHRQ Publication No. 10-EHC002-EF. Rockville, MD: Agency for Healthcare Research and Quality; October 2009.

6. Bolen S, Wilson L, Vassy J, et al. Comparative Effectiveness and Safety of Oral Diabetes Medications for Adults With Type 2 Diabetes. AHRQ Publication No. 07-EHC010-EF. Rockville, MD: Agency for Healthcare Research and Quality; July 2007.

7. Bennett WL, Maruthur NM, Singh S, et al. Comparative effectiveness and safety of medications for type 2 diabetes: an update including new drugs and 2-drug combinations. Ann Intern Med 2011; 154(9):602-13.

8. Chong J, Poole P, Leung B, et al. Phosphodiesterase 4 inhibitors for chronic obstructive pulmonary disease. Cochrane Database Syst Rev. 2011 May 11;(5):CD002309.

9. Berger JS, Lala A, Krantz MJ, et al. Aspirin for the prevention of cardiovascular events in patients without clinical cardiovascular disease: a meta-analysis of randomized trials. Am Heart J. 2011;162(1):115-24.E2.

10. Gail, MH, Costantino JP, Bryant J, et al. Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer. J Natl Cancer Inst. 1999 Nov 3; 91(21):1829-46.

11. Varadhan R, Weiss CO, Segal JB, et al. Evaluating health outcomes in the presence of competing risks: a review of statistical methods and clinical applications. Med Care. 2010 Jun;48(6 Suppl):S96-105.

12. National Institutes of Health: National Heart, Lung, and Blood Institute. Incidence and Prevalence: 2006 Chart Book on Cardiovascular and Lung Diseases. http://www.nhlbi.nih.gov/resources/docs/06a_ip_chtbk.pdf. Accessed October 22, 2012.

13. Man-Son-Hing M, Laupacis A, O'Connor AM, et al. Patient preference-based treatment thresholds and recommendations: a comparison of decision-analytic modeling with the probability-tradeoff technique. Med Decis Making. 2000 Oct-Dec;20(4):394-403.

14. Berger JS, Roncaglioni MC, Avanzini F, et al. Aspirin for the primary prevention of cardiovascular events in women and men: a sex-specific meta-analysis of randomized controlled trials. JAMA. 2006 Jan 18;295(3):306-13.

15. Hernandez-Diaz S, Garcia Rodriguez LA. Cardioprotective aspirin users and their excess risk of upper gastrointestinal complications. BMC Medicine 2006; 4(22).

16. Xu J, Kochanek KD, Murphy SL, et al. Deaths: final data for 2007. National Vital Statistics Reports. 2010; 58(19).

17. Cates CJ. Simpson's paradox and calculation of NNT from meta-analysis. BMC Med Res Methodol. 2002;2:1.

18. Cates CJ. Dr Chris Cates' EBM Web site. http://www.nntonline.net . Accessed 14 October 2012.

19. McQuay HJ, Moore RA. Using numerical results from systematic reviews in clinical practice. Ann Intern Med.1997;126(9):712.

20. Altman DG. Confidence intervals for the number needed to treat. BMJ 1998; 317:1309.

21. U.S. Preventive Services Task Force. Aspirin for the prevention of cardiovascular disease: U.S. Preventive Services Task Force Recommendation Statement. Ann Intern Med. 2009 Mar 17;150(6):396-404.

22. Eichler K, Puhan MA, Steurer J, et al. Prediction of first coronary events with the Framingham Score: a systematic review. Am Heart J. 2007 May; 153(5): 722-31, 731.E1-8 .

23. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008; 336(7650):924-6.

24. Altman DG, Andersen PK. Calculating the number needed to treat for trials where the outcome is time to an event. BMJ1999;319 (7223): 1492-5.

25. Shamliyan T, Kane RL, Ansari MT, et al. Development of the Quality Criteria To Evaluate Nontherapeutic Studies of Incidence, Prevalence, or Risk Factors of Chronic Diseases: Pilot Study of New Checklists. Methods Research Report. AHRQ Publication No. 11-EHC008-EF. Rockville, MD: Agency for Healthcare Research and Quality; January 2011. http://effectivehealthcare.ahrq.gov/.

26. MacLean S, Mulla S, Akl EA, et al. Patient values and preferences in decision making for antithrombotic therapy: a systematic review: antithrombotic therapy and prevention of thrombosis, 9th ed.: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest. 2012 Feb;141(2 Suppl):e1S-23S.

27. West SL, Gartlehner G, Mansfield AJ, et al. Comparative Effectiveness Review Methods: Clinical Heterogeneity. Methods for Effective Health Care. AHRQ Publication No. 10-EHC070-EF. Rockville (MD): Agency for Healthcare Research and Quality; September 2010.

Acronyms/Abbreviations

Acronym	Definition
AHRQ	Agency for Healthcare Research and Quality
CI	Confidence interval
Gail/NCI	Gail/National Cancer Institute
GI	Gastrointestinal
MI	Myocardial Infarction
NNH	Number needed to harm
NNT	Number needed to treat
RCTs	Randomized controlled trials
RR	Relative Risk

Appendix A. Details of Our Data Sources for Effect Estimates, Baseline Risk of All Four Outcomes, and Weights of Outcomes

Table A-1. Treatment effects of aspirin for the primary prevention of cardiovascular events

Treatment Effect	RR [95% CI][1*]
MI	0.86 [0.74,1.00]
Ischemic stroke	0.87 [0.73,1.02]
Hemorrhagic stroke	1.35 [1.01,1.81]
Major bleeds	1.62 [1.31,2.00]

***** This meta-analysis included the nine major primary prevention trials: British Doctors' Trial (BMD), Physicians' Health Study (PHS), Thrombosis Prevention Trial (TPT), Hypertension Optimal Treatment (HOT) study, Primary Prevention Project (PPP), Women's Health Study (WHS), Atherosclerosis Trial [AAAT], Japanese Primary Prevention of Atherosclerosis with ASA for Diabetes, and the Prevention of progression of arterial disease and diabetes (POPADAD) trial. The mean age of the participants ranged from 54.6 years in the Women's Health Study to 65 years in the JPAD trial. Most of the trials used low-dose daily aspirin except for the British Doctors study which evaluated 500 mg of daily aspirin and PHS which evaluated 325 mg of aspirin every alternate day and WHS which evaluated 100 mg every alternate day. The mean followup duration ranged from 3.6 years in the Primary Prevention Project to 10.1 years in the Womens Health Study.) CI = confidence interval, MI = Myocardial infarction, RR = Relative risk

Table A-2. Baseline risk

Indicator (Incidence Rate)	Source (Year)
MI	ARIC[2] Community Surveillance Component (1987-89, 1990-92, 1996-98) Baseline *n*=4,000 aged 45-64 (total sample: 15,792). Hospitalized MI incidence currently documented in men and women aged 35-84. If population-based estimates are unavailable, data extracted from primary care setting
Major ischemic stroke	ARIC Cohort Component[2]
Major hemorrhagic stroke	ARIC Cohort Component[2]
Major GI bleeds[4]	GPRD: population-based database in the U.K. (approx. 3 million patients)[3]
All-cause mortality	U.S. Vital Statistics[5]

MI = Myocardial infarction, RR = Relative risk, GI = Gastrointestinal, U.K. = United Kingdom, U.S. = United States, ARIC = Atherosclerosis Risk in Communities, GPRD = General Practice Research Database

Table A-3. Relative weights for outcomes[6]

Weights (descending severity)	Perceived severity	Complement (Weights used)
Major stroke	0.11	0.89
MI	0.55	0.45
Major GI bleeds	0.80	0.20

MI = Myocardial infarction, GI = Gastrointestinal

Table A-4. Source of incidence rates (per 1,000 person-years) without aspirin based on control event rates in the trials for Sensitivity analysis[7]

	Incidence rates in Men	Incidence rates in Women
MI	5.6	1.0
Major ischemic stroke	2.1	1.1
Major hemorrhagic stroke	0.3	0.2
Major GI bleeds	0.9	0.5
All-cause mortality	7.9	3.4

The incidence rate was estimated by dividing the number of events in the control arm by the person years of followup in the control arm which equals the total number of participants in the control arm and the mean length of followup for the trial= (n/ N * Mean followup duration of the trial in years)
MI = Myocardial infarction, GI = Gastrointestinal

References

1. Berger JS, Lala A, Krantz MJ, et al. Aspirin for the prevention of cardiovascular events in patients without clinical cardiovascular disease: a meta-analysis of randomized trials. Am Heart J. 2011 Jul;162(1):115-24.e2.

2. ARIC/NHLBI Surveillance Data, Incidence and prevalence: 2006 chart book on cardiovascular and lung diseases, Table 2.1, p.6 and Ibid, Table 4.6, p. 42.

3. Garcia Rodriguez LA, Perez Gutthann S. (1998). Use of the uk general practice research database for pharmacoepidemiology. Br J Clin Pharmacol. 1998 May;45(5):419-25.

4. Hernández-Díaz S, García Rodríguez LA. Cardioprotective aspirin users and their excess risk of upper gastrointestinal complications. BMC Med. 2006 Sep 20;4:22.

5. National Vital Statistics System, 2012. Retrieved from Centers for Disease Control and Prevention Web site: http://www.cdc.gov/nchs/nvss.htm.Accessed October 22, 2012.

6. Man-Son-Hing M, Laupacis A, O'Connor AM, et al. Patient preference-based treatment thresholds and recommendations: a comparison of decision-analytic modeling with the probability-tradeoff technique. Med Decis Making. 2000 ;20(4):394-403.

7. Berger JS, Roncaglioni MC, Avanzini F, et al. Aspirin for the primary prevention of cardiovascular events in women and men: a sex-specific meta-analysis of randomized controlled trials. JAMA. 2006 Jan 18;295(3):306-13.