THE UNIVERSITY of York

CHE
Centre For Health Economics

Expected Health Benefits of Additional Evidence: Principles, Methods and Applications

CHE Research Paper 83

Expected health benefits of additional evidence: Principles, methods and applications

[1,2]Karl Claxton
[2]Susan Griffin
[3]Hendrik Koffijberg
[2]Claire McKenna

[1]Department of Economics and Related Studies, University of York, UK
[2]Centre for Health Economics, University of York, UK
[3]Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, The Netherlands

April 2013

Background to series

CHE Discussion Papers (DPs) began publication in 1983 as a means of making current research material more widely available to health economists and other potential users. So as to speed up the dissemination process, papers were originally published by CHE and distributed by post to a worldwide readership.

The CHE Research Paper series takes over that function and provides access to current research output via web-based publication, although hard copy will continue to be available (but subject to charge).

Acknowledgements
The research presented in this report was funded by the Patient-Centered Outcomes Research Institute (PCORI) in the United States. We are very grateful to PCORI for the opportunity to develop this work and present it at the PCORI Methodology Workshop for Prioritizing Specific Research Topics, which took place in Washington in December 2012. Hendrik Koffijberg was also supported by the Netherlands Organization for Scientific Research (NWO, VENI grant # 916.11.126). We would especially like to thank Iain Chalmers who over very many years has patiently encouraged us to try and make value of information analysis more easily accessible, with concepts expressed and analysis presented in ways that can engage with a wider audience and remain relevant and useful even in contexts where formal cost-effectiveness analysis may not be accepted as an important starting point for decision about health care technologies. We would also like to thank Ian Roberts for his generous advice when we started to develop our analysis of the CRASH trial prior to this research for PCORI. This initial work was supported by Medical Research Council Programme Grant as part of the Health Services Research Collaboration.

Disclaimer
Papers published in the CHE Research Paper (RP) series are intended as a contribution to current research. Work and ideas reported in RPs may not always represent the final position and as such may sometimes need to be treated as work in progress. The material and views expressed in RPs are solely those of the authors and should not be interpreted as representing the collective views of CHE research staff or their research funders. Specifically, the views expressed in this Research Report do not necessarily represent those of the Patient-Centered Outcomes Research Institute (PCORI).

Further copies
Copies of this paper are freely available to download from the CHE website www.york.ac.uk/che/publications/ Access to downloaded material is provided on the understanding that it is intended for personal use. Copies of downloaded papers may be distributed to third-parties subject to the proviso that the CHE publication source is properly acknowledged and that such distribution is not subject to any payment.

Printed copies are available on request at a charge of £5.00 per copy. Please contact the CHE Publications Office, email che-pub@york.ac.uk, telephone 01904 321405 for further details.

Centre for Health Economics
Alcuin College, University of York
York, UK
www.york.ac.uk/che

Contents

Abstract ii
Executive summary iii
1. Introduction 1
2. Background 2
3. What assessments are needed? 3
3.1 Value of evidence and the value of implementation 3
3.2 Minimum clinical difference 4
3.3 Assessments in different contexts 5
3.4 Variability in patient outcomes and individualized care 5
4. How might these assessments be informed? 7
4.1 Primary endpoint captures key aspects of outcome 7
4.1.1 Value of additional evidence 8
4.1.2 Value of implementation 12
4.1.3 Minimum clinical difference 15
4.2 Primary endpoint linked to other aspects of outcome 16
4.2.1 Value of additional evidence about mortality 17
4.2.2 Value of additional evidence about survival and disability 18
4.2.3 Assessing the value and priority of proposed research 19
4.2.4 Informing research design 22
4.2.5 Impact of commissioned research 22
4.3 Different weights to reflect the relevance of evidence 23
4.3.1 Reflecting the relevance of evidence in meta-analysis 24
4.3.2 Implications for the value of additional evidence 25
4.4 More than two alternative interventions need to be compared 26
4.4.1 Synthesis of evidence with multiple comparisons 27
4.4.2 Implications for the value of additional evidence 30
5. Considerations 32

Abstract

The purpose of this research is to illustrate: i) the principles of what assessments are required when considering the need for additional evidence and the priority of proposed research; and ii) how these assessments might be informed by quantitative analysis based on standard methods of systematic review and meta-analysis.

We briefly outline the principles of what type of assessments are needed when considering research prioritization and commissioning. These are more fully examined through the integration of the principles of value of information analysis with the type of meta-analysis commonly conducted in systematic review and its application to four topics or case studies. The case studies were selected to cover a number of contexts in which these assessments are likely to be required and include: where the primary endpoint in existing studies capture key aspects of outcome; where it can be linked to other important aspects of outcome; when different 'weights' might be used to reflect the relevance and quality of different studies and when more than two alternative interventions need to be compared. Throughout, we distinguish the value of additional evidence and the value of implementing the findings from existing research. We also show how the value of additional evidence and the need for further research depends on the clinical difference in key aspects of outcome that would need to be demonstrated before clinical practice 'should' or is likely to change. We also consider whether the expected health benefits of additional evidence are sufficient to regard a particular research proposal as potentially worthwhile and whether it should it be prioritized over other research topics that could have been commissioned with the same resources. We also set out the implications of this type of analysis for research design, including whether randomised design is likely to be needed, the most appropriate scale of future research and the sequence in which different types of study might be undertaken.

The report demonstrates how making best use of the results of standard meta-analysis can directly inform the questions posed in research prioritisation and commissioning. In principle, this type of analysis could become part of the routine reporting of the findings of systematic review. In addition, it is sufficiently general to be relevant across a range of different types of health care systems, whether or not formal cost effectiveness analysis is explicitly used as part of the decision making process.

Executive summary

This report presents a paper commissioned by PCORI which was presented and discussed at Methodology Workshop for Prioritizing Specific Research Topics in December 2012. As such it is intended to be accessible to a non specialist and non technical audience. It is not intended to be an academic review of value of information or research prioritization methods nor the experiences of its application in health. For this reason and to avoid any appearance of partiality in citation, the referencing is restricted to the sources of evidence and meta-analytic methods required in the Technical Appendices.

We show through application to four case studies how the assessments needed are reflected in the quantitative analysis, so judgement and deliberation can be informed through a relatively simple extension of standard methods of meta-analysis. We have taken it for granted that any topic suggestion or specific research proposal will include a systematic review of existing evidence and, where appropriate, a meta-analysis; since funding additional research without knowledge of existing evidence would seem inappropriate and potentially unethical if an experimental research design is required. Although extending meta-analysis in the way described above is not technically challenging, nor does it pose particular computational problems, there may be an issue of familiarity with the principle that it is the consequences of uncertainty that matters, rather than the precision with which quantities can be estimated.

The following principles of what assessments are required when considering the need for additional evidence and the priority of proposed research, help inform whether the expected health benefits of additional evidence are sufficient to regard a particular research proposal as potentially worthwhile and whether it should it be prioritized over other research topics that could have been commissioned with the same resources. These principles are more practically illustrated and explained in each of the case studies presented in Section 4.

Value of evidence and the value of implementation

Additional evidence is valuable because it can improve patient outcomes by resolving existing uncertainty about the effectiveness of the interventions available, thereby informing treatment choice for subsequent patients. The scale of uncertainty is indicated by the results of systematic review and meta-analysis, so when this is combined with information about baseline risk and incidence, the expected consequences of uncertainty can be expressed in terms of health outcomes. These expected consequences can be interpreted as an estimate of the health benefits that could potentially be gained each year if the uncertainty surrounding treatment choice could be resolved, i.e., it indicates an expected upper bound on the health benefits of further research which would confirm whether or not an intervention is actually more effective than the others currently available.

Health outcomes can also be improved by ensuring that the accumulating findings of research are implemented and has an impact on clinical practice. Indeed, the potential improvements in health outcome by encouraging the implementation of what existing evidence suggests is the most effective intervention may well exceed the potential improvements in health outcomes through conducting further research.

The distinction between these two very different ways to improve patient outcomes is important because additional research is certainly not the only, or necessarily the most effective, way to influence clinical practice. Insofar as there are other mechanisms (e.g., more effective dissemination of existing evidence) or policies which fall within the remit of other bodies (e.g., incentives and sanctions), then continuing to conduct research to influence implementation rather than because there is real value in acquiring additional evidence itself would seem inappropriate. However, the

importance of implementing the findings of proposed research might influence consideration of its priority and research design in a number of ways. If it is very unlikely that findings will be implemented and other mechanisms are unlikely to be effective, then other areas of research where smaller potential benefits are more likely to be realized might be prioritized. If the impact of research on clinical practice is likely to require highly statistically significant results this will influence the design, cost and time taken for research to report and therefore its relative priority (see Section 4.2.4). It maybe that a larger clinical difference in effectiveness would need to be demonstrated before research would have impact on practice (see Section 4.1.3, 4.3.2 and 4.4.2). This will tend to reduce the potential benefits of further research because larger differences are less likely to be found than smaller ones.

Minimum clinical difference

The concept of a minimum clinical difference (MCD) required to change practice (effect size) is used throughout the analysis in Section 4, where estimates of the expected health benefits of additional evidence are calculated for a range of MCD. It shows how the value of additional evidence and the need for further research depends on the clinical difference in key aspects of outcome that would be need to be demonstrated before clinical practice 'should' or is likely to change (see Sections 4.1.3, 4.3.2 and 4.4.2. Although any improvement in health outcomes is valuable (i.e., a minimum clinical difference close to zero) there are a number of reasons why a larger difference might be required:

i. Where the quantitative analysis is restricted to the primary endpoint reported in existing clinical trials but there other important aspects of outcome that are not captured in this endpoint (e.g., adverse events or quality of life impacts).
ii. When there may be an adverse impact on heath system performance; for example, an intervention may increase need for other services or increase out of pocket expenses for patients.
iii. It maybe that in some topics a larger clinical difference in effectiveness would need to be demonstrated before research would have an impact on practice and the findings of proposed research would be widely implemented.

Specifying an MCD implicitly accounts for the other unquantified aspects of outcome and impacts on system performance through a deliberative process which would need to consider whether proposed research is still a priority at an MCD that is regarded as sufficient to account for these other effects.

Variability in patient outcomes and individualized care

We make a clear distinction between uncertainty, variability and heterogeneity at the outset. Uncertainty refers to the fact that we do not know what the expected effects will be of using an intervention in a particular population of patients (i.e., the effects of an intervention on average). Variability refers to the fact that individual responses to an intervention will differ within the population or even in a sub population of patients with the same observed characteristics. Heterogeneity refers to those individual differences in response that can be associated with differences in observed characteristics, i.e., where the sources of natural variability can be identified and understood.

Patient outcomes can be improved by either acquiring additional evidence to resolve the uncertainty in the expected effects of an intervention, and/or by understanding the sources of variability and dividing the population into finer subgroups where the intervention will be expected to be effective in some but not in others. However, a greater understanding of heterogeneity also has an impact on the value of additional evidence. As more subgroups can be defined the precision of the estimates

of effect is necessarily reduced (the same amount of evidence offers fewer observations in each subgroup) but the uncertainty about which intervention is most effective may be reduced in some (e.g., where it is particularly effective or positively harmful) but increase in others.

We don't include an examination of subgroups in Section 4, but in principle same type of analysis can be applied within each subgroup identified based on existing evidence. Nor do we directly address the potential value of research which might reveal the reasons for variability in outcome; informing which subgroups could benefit most from an intervention, or the choice of the physician patient dyad in selecting care given their symptoms, history and preferences (i.e., individualized care). This type of research may be very different from the type of evaluative research that reduces uncertainty about estimates of effect. For example, it might include: diagnostic procedures and technologies; pharmacogenetics; analysis of observational data and treatment selection; as well as novel trial designs which can reveal something of the joint distribution of effects.

Assessments in different contexts

How these key assessments might be informed by quantitative analysis is examined though application to case studies in the following four contexts that are likely to arise in a body responsible for research prioritisation, considers suggested topics and specific research proposals.

i. Primary endpoint in the meta analysis captures key aspects of outcome (see Section 4.1 and Appendix A)
ii. Primary endpoint in the meta analysis needs to be linked to other aspects of outcome (see Section 4.2 and Appendix B)
iii. Different weights to reflect the relevance and potential bias of existing evidence (see Section 4.3 and Appendix C)
iv. More than two alternative interventions need to be compared (see Section 4.4 and Appendix D)

Four case studies were selected to cover these four different contexts:

i. Early thrombolysis using streptokinase for the treatment of acute myocardial infarction (see Section 4.1 and Appendix A)
ii. Corticosteroids following traumatic head injury (see Section 4.2 and Appendix B)
iii. Probiotics in patients with severe acute pancreatitis (SAP) (see Section 4.3 and Appendix C)
iv. Topotecan, PLDH and paclitaxxel for second-line treatment of advanced ovarian cancer (see Section 4.4 and Appendix D)

Each case study is used to illustrate methods of analysis and appropriate interpretation of the results, including the type of scientific value judgments that are required, in one of these four contexts. The case studies were also selected to cover the different types of meta-analysis that are likely to be required and highlight some of the dangers of: focusing only on a single primary endpoint (see Section 4.2); inappropriate use of a random effect analysis (see Section 4.2 and 4.3); and failing to include the whole network of evidence (see Section 4.4).

Considerations

This work illustrates how explicit and quantitative analysis, based on systematic review and meta analysis, can be used to inform the assessments that need to be made when considering research prioritization and commissioning. Nonetheless, no quantitative analysis, no matter how assiduously conducted or sophisticated, can capture all aspects of scientific and social value relevant to making decisions about research priorities. Not least because both scientific and social value judgments are

quite reasonably disputed. The more relevant question is whether they offer a practical and useful starting point for deliberation and add to the transparency and accountability of the decision making process. It is on this basis that the usefulness of more explicit and quantitative analysis ought to be judged. We believe their potential usefulness in a range of decision making context has been demonstrated, so long as the results are appropriately interpreted and the limitations of the analysis are understood and explored.

The question of their practicality and feasibility within the time and resource constraints of a deliberative process of research prioritization will depend on context. We have taken it for granted that any topic suggestion or specific research proposal will include a systematic review of existing evidence and, where appropriate, a meta-analysis. Although extending meta-analysis in the way described above is not technically challenging, nor does it pose particular computational problems, there is an issue of familiarity with the principle that it is the consequences of uncertainty that matters, rather than the precision with which quantities are estimated. Some of the contexts examined also required more sophisticated forms of meta-analysis, e.g., Bayesian meta-analysis to link multiple endpoints in Section 4.2 and a MTC evidence synthesis in Section 4.4. These methods were required to estimate relative effectiveness even before the value of additional evidence was considered, so would be required in any event. In Section 5 we highlight some specific questions related to practicality and implementation relevant to bodies responsible for research prioritisation and commissioning.

1. Introduction

The purpose of this paper is to illustrate:

i. the principles of what assessments are required when considering the need for additional evidence and the priority of proposed research; and
ii. how these assessments might be informed by quantitative analysis based on standard methods of systematic review and meta-analysis.

We briefly outline the principles of what type of assessments are needed when considering research prioritization and commissioning (see Section 3). These principles are more fully examined through the integration of the principles of value of information analysis with the type of meta-analysis commonly conducted in systematic review and its application to four topics or case studies (see Section 4). The case studies were selected to cover a number of contexts in which these assessments are likely to be required and include: where the primary endpoint in existing studies capture key aspects of outcome (see Section 4.1); where it can be linked to other important aspects of outcome (see Section 4.2); when different 'weights' might be used to reflect the relevance and quality of different studies (see Section 4.3) and when more than two alternative interventions need to be compared (see Section 4.4).

Throughout, we distinguish the value of additional evidence and the value of implementing the findings from existing research. We also show how the value of additional evidence and the need for further research depends on the clinical difference in key aspects of outcome that would need to be demonstrated before clinical practice 'should' or is likely to change. In section 4.2.3 we consider in more detail whether the expected health benefits of additional evidence are sufficient to regard a particular research proposal as potentially worthwhile and whether it should it be prioritized over other research topics that could have been commissioned with the same resources. We also set out the implications of this type of analysis for research design, including whether randomised design is likely to be needed, the most appropriate scale of future research and the sequence in which different types of study might be undertaken.

Finally, Section 5 raises some of the questions posed when considering the practicality of adopting these methods as part of a funding agency's process of research prioritization and commissioning. Although we make clear throughout, it is worth noting at the outset that no quantitative analysis, no matter how assiduously conducted or sophisticated, can capture all aspects of scientific and social value relevant to making decisions about research priorities. Not least because both scientific and social value judgments are quite reasonably disputed. The more relevant question is whether they offer a useful starting point for deliberation and add to the transparency and accountability (to reason, evidence and widely held social value judgments) of the decision making process. It is on this basis that the usefulness of more explicit and quantitative analysis ought to be judged.

2. Background

This paper is intended to be accessible to a non specialist and non technical audience. It is not intended to be an academic review of value of information or research prioritization methods nor the experiences of its application in health.[1] Nonetheless it draws on a well developed and diverse methodological literature published in journals that cover related disciplines, such as: medical statistics, health economics, epidemiology; and operations research, as well as risk and decision science. The principles have a firm foundation in statistical decision theory with closely related concepts and methods in mathematics and financial economics with diverse applications in business decisions, engineering, environmental risk analysis and financial and environmental economics. There are now many applications in health, some commissioned to directly inform policy and others published in specialist as well as general medical and health policy journals.

We show through application to four case studies how each of the types of assessment required can be reflected in quantitative analysis, so judgement and deliberation can be informed through a relatively simple extension of standard methods of meta-analysis. We have taken it for granted that any topic suggestion or specific research proposal will include a systematic review of existing evidence and, where appropriate, a meta-analysis; since funding additional research without knowledge of existing evidence would seem inappropriate and potentially unethical if an experimental research design is required. Although extending meta-analysis in the way described above is not technically challenging, nor does it pose particular computational problems, there may be an issue of familiarity with the principle that it is the consequences of uncertainty that matters, rather than the precision with which quantities can be estimated (and possibly methods of simulation), amongst some of those who commonly conduct systemic reviews. Some of the contexts that are examined in Section 4 also required more sophisticated forms of meta-analysis, e.g., Bayesian meta-analysis in Section 4.2 and a mixed treatment comparison evidence synthesis in Section 4.4. These methods were not made necessary because of the need to consider the expected benefits of additional evidence, but would be required in any event to provide an appropriate estimate of effectiveness based on all the evidence then available.

[1] For this reason and to avoid any appearance of partiality in citation, the referencing is restricted to the essential sources of evidence and meta-analytic methods used in case studies that are reported in the Technical Appendices.

3. What assessments are needed?

The following principles of what assessments are required when considering the need for additional evidence and the priority of proposed research, help inform whether the expected health benefits of additional evidence are sufficient to regard a particular research proposal as potentially worthwhile and whether it should it be prioritized over other research topics that could have been commissioned with the same resources. We only briefly outline these principles here as they are more practically illustrated and explained in each of the case studies presented in Section 4.

3.1 Value of evidence and the value of implementation

Additional evidence is valuable because it can improve patient outcomes by resolving existing uncertainty about the effectiveness of the interventions available, thereby informing treatment choice for subsequent patients. For example, the balance of existing evidence might suggest that a particular intervention is expected to be the most effective, but there will be a chance that others are in fact more effective. If treatment choice is based on existing evidence then there will be a chance that other interventions would have improved health outcomes to a greater extent, i.e., there are adverse health consequences associated with uncertainty. The scale of uncertainty is indicated by the results of systematic review and meta-analysis, so when this is combined with information about baseline risk and incidence, the expected consequences of uncertainty can be expressed in terms of health outcomes. These expected consequences can be interpreted as an estimate of the health benefits that could potentially be gained each year if the uncertainty surrounding treatment choice could be resolved, i.e., it indicates an expected upper bound on the health benefits of further research which would confirm whether or not this intervention was actually more effective than the others currently available. These potential expected benefits increase with the size of the patient population whose treatment choice can be informed by additional evidence and the time over which evidence about the effectiveness of these interventions is expected to be useful (see Section 4.2.3).

Health outcomes can also be improved by ensuring that the accumulating finding of research is implemented and has an impact on clinical practice. Indeed, the potential improvements in health outcome by encouraging the implementation of what existing evidence suggests is the most effective intervention may well exceed the potential improvements in health outcomes through conducting further research.

The distinction between these two very different ways to improve patient outcomes is important because, although the results of additional research may influence clinical practice and may contribute to the implementation of research finding, it is certainly not the only, or necessarily the most effective, way to do so. Insofar as there are other mechanisms (e.g., more effective dissemination of existing evidence) or policies which fall within the remit of other bodies (e.g., incentives and sanctions),[2] then continuing to conduct research to influence implementation rather than because there is real value in acquiring additional evidence itself would seem inappropriate, because research capacity could have been used elsewhere to acquire additional evidence in areas where it would have offered greater potential health benefits and such a policy will have negative health effects for those patients enrolled in research who will receive interventions which are expected to be less effective.

Clearly, the potential health benefits of conducting further research will only be realized (patient outcomes actually improve) if the findings of the research do indeed have an impact on clinical

[2] Such bodies may not necessarily be public institutions or government agencies but might also include private for, or not for profit entities as well as professional groups and patient's organisations.

practice. Again, recognition that there are very many ways to influence the implementation of evidence, other than by conducting more research, is important. However, the importance of implementing the findings of proposed research might influence consideration of its priority and research design in a number of ways. If it is very unlikely that findings will be implemented and other mechanisms are unlikely to be effective, then other areas of research where smaller potential benefits are more likely to be realized might be prioritized. If the impact of research on clinical practice is likely to require highly statistically significant results this will influence the design, cost and time taken for research to report and therefore its relative priority (see Section 4.2.4). It maybe that a larger clinical difference in effectiveness would need to be demonstrated before research would have impact on practice (see Section 4.1.3, 4.3.2 and 4.4.2). This will tend to reduce the potential benefits of further research as well (a large difference is less likely to be found than a small one).

3.2 Minimum clinical difference

The concept of an effect size has been central to the design of clinical research and determines the sample size in most clinical trials. The effect size does not represent what is expected to be found by the research, but the difference in outcomes that would need to be detected for the results to be clinically significant and have an impact on clinical practice. The same concept is used throughout the analysis in Section 4, where estimates of the expected heath benefits of additional evidence are calculated for a range of minimum clinical differences (MCD) in outcomes. It shows how the value of additional evidence and the need for further research depends on the clinical difference in key aspects of outcome that would be need to be demonstrated before clinical practice 'should' or is likely to change (see Sections 4.1.3, 4.3.2 and 4.4.2 and the Technical Appendix). Although any improvement in health outcomes is valuable (i.e., a minimum clinical difference close to zero) there are a number of reasons why a non zero MCD might be appropriate.[3] For example larger differences might be required in the following circumstances:

i. Where the quantitative analysis is restricted to the primary endpoint reported in existing clinical trials but there other important aspects of outcome that are not captured in this endpoint (e.g., adverse events or quality of life impacts[4]). In Section 4.2.2 we illustrate how other important aspects of outcome might be more explicitly incorporated into quantitative analysis.
ii. When there may be an adverse impact on heath system performance; for example, an intervention may increase need for other services or increase out of pocket expenses for patients.
iii. It maybe that in some topics a larger clinical difference in effectiveness would need to be demonstrated before research would have an impact on practice and the findings of proposed research would be widely implemented.

Requiring that further research must demonstrate larger differences in effect will tend to reduce its expected potential benefits because large differences are less likely to be found than smaller ones. Specifying an MCD implicitly accounts for the other unquantified aspects of outcome and impacts on system performance through a deliberative process which would need to consider whether proposed research is still a priority at an MCD that is regarded as sufficient to account for these other effects.

[3] A minimum clinical difference less than zero is possible. It would imply that reduced effectiveness on the primary endpoint would be acceptable because it would be compensated for by other types of benefits (e.g., reductions in adverse events, improvement in quality of life, reduced need of other service or a reduction in out of pocket costs for patients.

[4] It might also include some assessment of the relationship between evidence of efficacy and effectiveness which may or may not suggest a smaller effect depending on whether the selection of treatments by patients and clinical exploits information unavailable at a summary level or even recorded in individual patient data.

3.3 Assessments in different contexts

How these key assessments might be informed by quantitative analysis is examined though application to case studies in the following four contexts that are likely to arise as funding agencies consider suggested topics and specific research proposals.

i. Primary endpoint in the meta analysis captures key aspects of outcome (see Section 4.1 and Appendix A)
ii. Primary endpoint in the meta analysis needs to be linked to other aspects of outcome (see Section 4.2 and Appendix B)
iii. Different weights to reflect the relevance and potential bias of existing evidence (see Section 4.3 and Appendix C)
iv. More than two alternative interventions need to be compared (see Section 4.4 and Appendix D)

Each case study is used to illustrate methods of analysis and appropriate interpretation of the results, including the type of scientific value judgments that are required, in one of these four contexts. The case studies were selected to cover these different contexts as well as the different types of meta-analysis that are likely to be required.

3.4 Variability in patient outcomes and individualized care

It is useful to make a clear distinction between uncertainty, variability and heterogeneity at the outset. Uncertainty refers to the fact that we do not know what the expected effects will be of using an intervention in a particular population of patients (i.e., the effects of an intervention on average). This remains the case even if all patients within this population have the same observed characteristics. Additional evidence can reduce uncertainty and provide a more precise estimate of the expected effects in the whole population or within subpopulations that might be defined based on observed characteristics. Variability refers to the fact that individual responses to an intervention will differ within the population or even in a sub population of patients with the same observed characteristics. Therefore, this natural variation in responses cannot be reduced by acquiring additional evidence about the expected or average effect. Heterogeneity refers to those individual differences in response that can be associated with differences in observed characteristics, i.e., where the sources of natural variability can be identified and understood. As more becomes known about the sources of variability (as variability is turned into heterogeneity) the patient population can be partitioned into sub populations or subgroups, each with a different estimate of the expected effect of the intervention and the uncertainty associated with it. Ultimately, as more sources of variability become known the sub populations become individual patients, i.e., individualized care.

This continuum is illustrated in Figure 3.1. It shows that patient outcomes can be improved by either acquiring additional evidence to resolve the uncertainty in the expected effects of an intervention, and/or by understanding the sources of variability and dividing the population into finer subgroups where the intervention will be expected to be effective in some but not in others. However, a greater understanding of heterogeneity also has an impact on the value of additional evidence. As more subgroups can be defined the precision of the estimates of effect is necessarily reduced (the same amount of evidence offers fewer observations in each subgroup). However, the uncertainty about which intervention is most effective may be reduced in some, where it is particularly effective or positively harmful, but increase in others. Therefore, the expected consequences of uncertainty per patient, or value of additional evidence per patient may be higher or lower in particular subgroups. The expected value of evidence across the whole population (the sum across all

subgroups of the population) may rise or fall.[5] However, in the limit as more sources of variability are observed the value of additional evidence will fall. Indeed, if all sources of variability could be observed then there would be no uncertainty at all.[6]

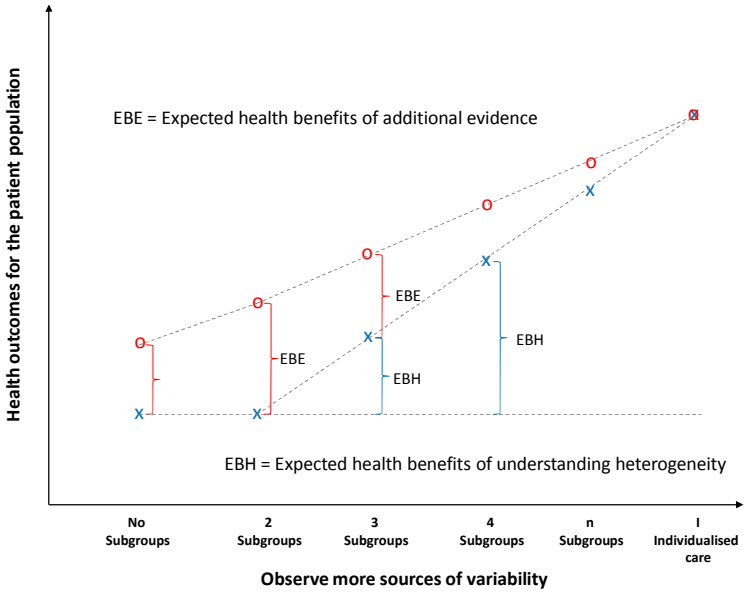


Figure 3.1 The value of resolving uncertainty and the value of understanding heterogeneity

We don't include examination of subgroups in Section 4, but in principle same type of analysis can be applied within each subgroup identified based on existing evidence. However, in these circumstances evidence synthesis might be extended if evidence of effect in one subgroup might inform effects in others and/or if some trials report some subgroup effects but others only report effects for the whole population. Conducting an analysis of the expected health benefits of additional evidence by subgroups is useful because it can indicate which types of patient need to be included in any future research design and others that could be excluded.

In this paper we do not directly address the potential value of research which might not resolve uncertainty but instead reveal the reasons for variability in outcome; informing which subgroups could benefit most from an intervention, or the choice of the physician patient dyad in selecting care given their symptoms, history and preferences (i.e., individualized care). This type of research may be very different from the type of evaluative research that reduces uncertainty about estimates of effect. For example, it might include: diagnostic procedures and technologies, pharmacogenetics; analysis of observational data and treatment selection as well as novel trial designs which can reveal something of the joint distribution of effects. Much methodological and applied work has been conducted in this rapidly developing area. There is an opportunity to explore ways of estimating the potential value of such research (the expected benefits of heterogeneity) based only on existing evidence and the results of standard meta-analysis. At present the upper bound that could, in principle, be specified (the expectation over all permutations of the joint distribution consistent with existing evidence) is computationally expensive and may be so high as to be relatively uninformative. However, should this be possible it would provide a very useful complement to the methods illustrated in Section 4.

[5] This is true for the overall value of additional evidence that could resolve all sources of uncertainty for all subgroups. However, insofar as resolving uncertainty for one subgroup might provide information relevant to others, a subgroup specific estimate of value will underestimate the true value of evidence about effectiveness in that subgroup.

[6] For illustration Figure 3.1 is drawn as if all sources of variability could be observed at an individual level – at which point there would be no uncertainty. However, there are fundamental limits to how much of natural variability can be observed and turned into heterogeneity no matter how much effort might be made.

4. How might these assessments be informed?

How the assessments required might be usefully informed is explored through the simple application of the principles of value of information analysis to the type of random or fixed effects meta-analysis commonly conducted as part of a systematic review. This type of analysis, the interpretation of the results and the other considerations that are relevant to research prioritization decisions are illustrated through application to four case studies:

i. Early thrombolysis using streptokinase for the treatment of acute myocardial infarction (see Appendix A)
ii. Corticosteroids following traumatic head injury (Appendix B)
iii. Probiotics in patients with severe acute pancreatitis (SAP) (see Appendix C)
iv. Topotecan, PLDH and paclitaxxel for second-line treatment of advanced ovarian cancer (see Appendix D)

The case studies were selected to cover the different contexts likely to arise for funding agencies (see Section 3.3). As made clear previously, we take it for granted that any topic suggestion or specific research proposal will include a systematic review of existing evidence and, where appropriate, a meta-analysis, before it is prioritized and further research commissioned.

4.1 Primary endpoint captures key aspects of outcome

The purpose of this and the following sections is to illustrate:

i. the principles of what assessments are required when considering the need for additional evidence and the priority of proposed research;
ii. how these assessments might be informed by quantitative analysis based on standard methods of systematic review and meta-analysis.

In this section we also show how evidence accumulates as a sequence of clinical trials, addressing a particular clinical question, reported over time. As a consequence, the need for additional evidence ultimately declines but the benefits of implementing the research findings of the cumulating evidence grow. Eventually the potential benefits of conducting more research is exceeded, not only by the benefits of implementing what existing evidence suggests is the most effective intervention, but also by the negative health consequences of continuing to enroll more patients to less effective interventions in subsequent trials.

A classic example is the sequence of trials that investigated early thrombolysis using streptokinase (SK) for the treatment of acute myocardial infarction (MI). Although many of the later trials in this sequence also informed the choice between other interventions to deliver early thrombolysis, for ease of exposition this re-analysis focuses on the choice between SK and control (no thrombolysis).[7]. We also restrict attention to the primary endpoint reported in the trials and initially assume that this captures the key aspects of health effects associated with the intervention. We also recognize that although mortality following acute MI maybe the appropriate primary outcome, it is not necessarily the only relevant outcome and certainly stroke and its consequences are also very relevant, especially when considering the later interventions for early thrombolysis. Specifying a minimum clinical difference required to change clinical practice is one way to incorporate such concerns about potential adverse events and other consequences of recommending a more effective intervention (e.g., adverse impacts on health system performance), albeit implicitly. In Section 4.2 we illustrate

[7] We illustrate how this type of analysis can be extended to multiple alternatives in Section 4.4 using the case study in ovarian cancer.

how important aspects of outcome which are not captured by the primary endpoint in the trials might be more explicitly incorporated into quantitative analysis.

The sequence of trials can be represented by a standard Forest plot reported in Figure 4.1 (a). The nature and context of these trials, as well as formal tests for heterogeneity, suggest that a random rather than fixed effect meta-analysis would appear to be more appropriate (see Appendix A for details).[8] Of course, the combined evidence of all these trials now strongly favors SK compared to control, whether or not this evidence is synthesized using random or a fixed effect meta-analysis. The more interesting questions, however, include:

i. How did the estimated effectiveness of SK and the uncertainty associated with it change as this body of evidence evolved?
ii. When might it have been reasonable to conclude that the evidence was sufficient to recommend SK over control?
iii. At what point was it unlikely that the additional research that was conducted would be worthwhile?
iv. Could health outcomes have been improved most by acquiring additional evidence or by encouraging the implementation what existing evidence suggests is the most effective intervention?
v. How might the value of additional evidence about this topic have been judged relative to the need for additional evidence in other clinical areas?

Importantly, we do not use hindsight to address these questions but ask how quantitative analysis might have informed these assessments at the time, with the evidence that was then available.

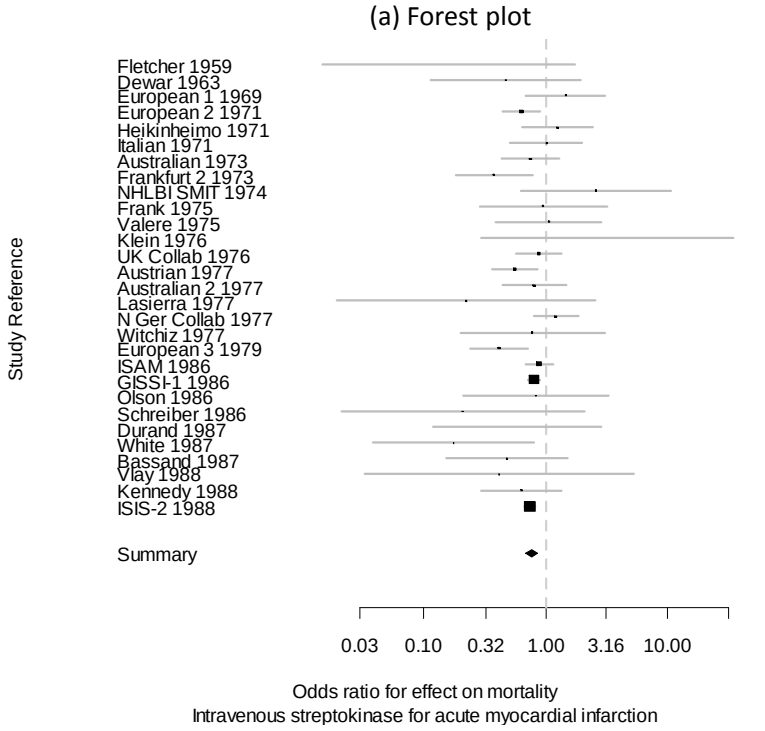
This sequence of trials can also be re-analyzed as a cumulative meta-analysis, where the estimate of the effect on mortality and the uncertainty associated with it is updated as each subsequent trial reports. This cumulative meta-analysis is also illustrated in Figure 4.1(b) and shows that the balance of evidence very quickly favored SK. The uncertainty associated with whether SK is more effective than control ultimately declined and once ISIS 2 reported it was entirely resolved. However, it was not until the European 3 trial in 1979 that a single trial reported a statistically significant result that favored SK *and* the cumulated evidence also suggested a statistically significant result, i.e., the probability that SK was more effective than control exceeded 0.95 in the trial and the random effect meta-analysis.[9]

4.1.1 *Value of additional evidence*

Restricting attention to whether or not a result of a trial or a meta-analysis is statistically significant provides only a partial summary of the uncertainty associated with the effectiveness of an intervention, nor does it indicate the importance of the uncertainty for patient outcomes or the potential gains in health outcomes that might be expected from acquiring additional evidence that could resolve it.

[8] The assumption underlying the fixed effect model is that each trial result is estimating a common unknown pooled effect, while any variation in the estimated effect sizes across studies is due to sampling error. If this assumption is considered to be too restrictive, a random effects model can be used, which allows the unknown pooled effect to vary between studies. There are particular dangers of inappropriately using random effect to take account of what are believed to be differences in the relevance and potential for bias. This is especially acute if the smaller trials are believed to be more vulnerable to bias and less relevant to the target population.

[9] The uncertainty associated with the effectiveness of SK resolves more quickly based on a fixed effects meta-analysis, reaching a statistically significant result in 1971, following European 2 which itself reported the first statistically significant result in favor of SK - see Appendix A for analysis based on fixed effects.



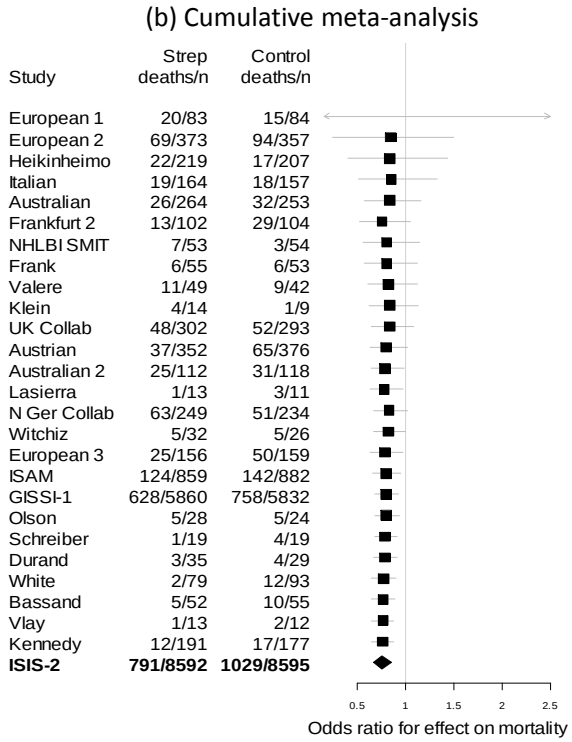


Figure 4.1 The sequence of trials of early thrombolysis using streptokinase

For example, following European 3 the cumulated evidence suggests that SK can confidently be expected to reduce mortality (the chance that SK is more effective than control is 98.41%). Nonetheless, this means that there remains a small chance (an 'error' probability) of 1.59% that mortality will be higher if SK is used following acute MI, i.e., there is a small chance that a decision to

use SK for MI will be 'wrong' and, if it is, there will be consequences in terms of higher mortality in the patient population.

Translating the chance of error at different points in this sequence of trials into a distribution of the scale of consequences associated with the uncertainty in patient outcomes requires applying the uncertain estimate of the relative effect of SK (the odds ratios reported in Figure 4.1(b)) to an estimate of baseline mortality risk following acute MI, which will also be estimated with uncertainty. In addition an estimate of the incidence of patients who face a choice between SK or control following acute MI is also required.[10] The combined effect of uncertainty in relative effect and baseline risk can be characterized by taking repeated random samples from their distributions, which are already estimated in the meta-analysis. Each random sample, or simulated value, can be interpreted as one possible realization of uncertainty, i.e., one possible 'true' value or way things might turn out given the information that was then available.[11]

This analysis can provide a distribution of the health consequences of uncertainty, which is illustrated in Figure 4.2. Commonly in this case there are no consequences, because the use of SK is most likely to be the correct decision, i.e., the balance of evidence favors SK (e.g., following European 1, 69% of the simulated values from the meta-analysis favor SK and are associated with zero consequences). However, there is a real possibility that SK is not effective (error probability of 0.31) so there are consequences, if SK is used for acute MI, in terms of deaths that could have otherwise been averted. There is a greater chance of more limited consequences (e.g., a 20% chance of consequences between zero and 15,000 deaths per year) and a smaller chance of larger consequences (e.g., a 11% chance of consequences greater than 15,000 deaths per year). The average over this distribution provides an estimate of the expected consequences of uncertainty, which is 6,264 deaths per year following European 1.

These expected consequences can be interpreted as an estimate of the health benefits that could potentially be gained each year if the uncertainty surrounding this decision could have been resolved at that time, i.e., it indicates an expected upper bound on the health benefits of further research which would confirm whether or not SK was actually more effective than control. These potential expected benefits increase with the size of the patient population whose treatment choice can be informed by additional evidence and the time over which evidence about the effectiveness of SK was expected to be useful (see Section 4.2).

Importantly, Figure 4.2 also illustrates that both uncertainty and its potential consequences declines as evidence accumulates. For example, following NHLBI SMIT in 1974 the chance that SK is more effective than control has risen to 91% (there is 91% chance of zero adverse MI mortality consequences of using SK). The chance of more limited consequences is lower (8.77% chance of consequences between zero and 15,000 deaths per year) and the chance of greater consequences is almost eliminated (0.21% chance of consequences greater than 15,000 deaths per year). As a result the expected consequences of the uncertainty surrounding the effectiveness of SK (the average over this distribution) fall to 306 deaths per year. By the time European 3 reports there is much less uncertainty about the effectiveness of SK (98% chance that SK is more effective with zero adverse consequences). The chance of more limited consequences is lower (1.58% chance of consequences between zero and 15,000 deaths per year) and there is no measureable chance of consequences

[10] An estimate of the size of the patient population whose treatment choice can benefit from the information provided by further research is required. Therefore, in a chronic condition an estimate of prevalence as well as incidence would be required.

[11] In this analysis we use the control arms of the trials as an estimate of baseline risk directly from the output of the meta-analysis, preserving the correlation in their estimation. However, baseline risk could be based only on the control arms of those trials regarded as most relevant to the target population (see Section 4.3) or from external evidence. The estimates reported here are also based on estimates of UK incidence - see Appendix A for details.

greater than 15,000 deaths per year. As a result the expected consequences or the upper bound of the expected value of additional evidence falls to 27 deaths per year. At this point the potential health benefits of subsequent research appear modest compared to the potential benefits of research that might have been conducted using the same resources in other clinical areas.

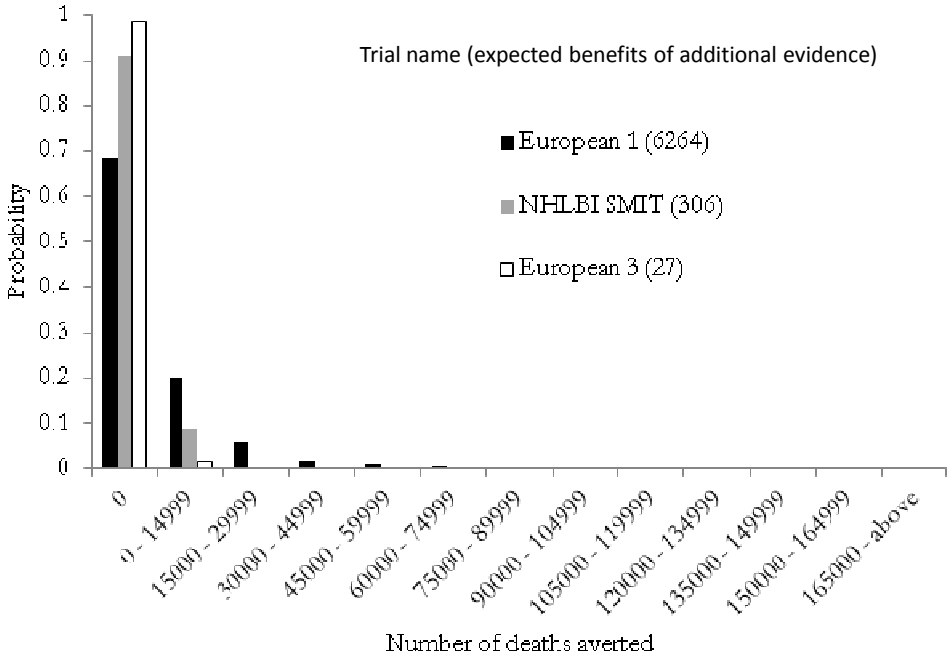


Figure 4.2 Distribution of the consequences of uncertainty (SK for acute MI)

This decline in the potential value of additional evidence is also illustrated in Figure 4.3. The steep decline in potential health benefits of additional evidence suggests that the later trials in the sequence may not have been necessary to inform the question of whether SK was more effective than control in acute MI, although they may have been valuable in informing the choice between other interventions. It also illustrates the negative health impact of subsequent trials which randomly allocated patients to a control arm that was expected to be less effective on the balance of the then existing evidence.

After European 3 (the 19th study) the health costs of the next trial exceeds the upper bound on the expected health benefits they offered each year. Of course, in coming to a view about when the evidence was sufficient and when further research in this area should not have been regarded as a priority some consideration needs to be given to:

i. The period over which evidence might be valuable, i.e., relevant to and able to inform clinical practice (see Section 4.2).
ii. The potential benefits of research in other areas that could have been conducted using these resources (see Section 4.2.3).
iii. Whether health outcomes could be improved most by acquiring additional evidence or by encouraging the implementation of what existing evidence suggests is the most effective intervention (see Section 4.1.2).
iv. Whether there are other aspects of outcome or health system performance that are relevant (see Section 4.1.3).

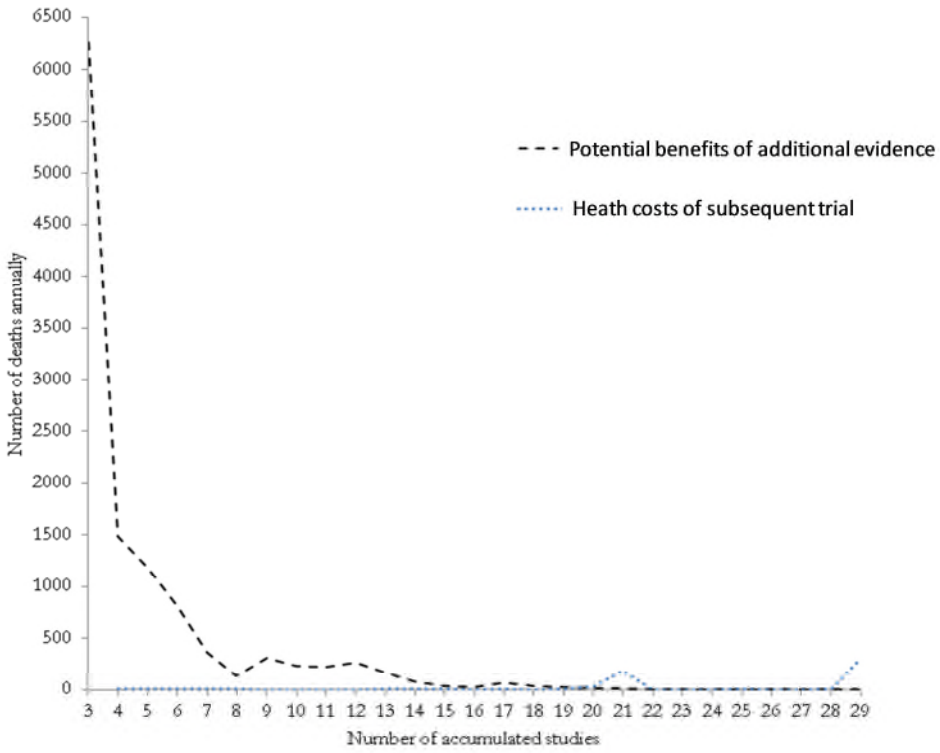


Figure 4.3 Expected health benefits of additional evidence (SK for acute MI)

4.1.2 *Value of implementation*

The cumulating evidence in Figure 4.1(b) very quickly suggested that, on balance, the use of SK in acute MI would be expected to reduce mortality (the expected or mean odds ratio in Figure 4.1(b) is less than 1). Therefore, health outcomes can be improved by ensuring that the accumulating findings of research are implemented and have an impact on clinical practice. Indeed, the potential improvements in health outcome by encouraging the implementation of what existing evidence suggests is the most effective intervention may well exceed the potential improvements in health outcomes through conducting further research. This is illustrated in Figure 4.4 where, from European 1 onwards the value (in terms of deaths that could be averted each year), through the use of SK exceed the potential health benefits of acquiring additional evidence about its effectiveness. Of course, over time as evidence accumulated, and especially after European 3 (the first trial that reported a statistically significant result *and* a random effect analysis of the cumulated evidence also suggested a statistically significant result in favor of SK), clinical practice did respond to this evidence with widespread use of SK for early thrombolysis. Therefore, the value of implementation efforts by other bodies declines as much of clinical practice respond to evidence and has already implemented the findings.[12]

[12] We have not attempted to reconstruct the historic growth in utilisation of SK in the UK over the period when this sequence of trials reported - in part due to limited historic data sources in the UK. Instead, we explore a number of scenarios of how utilisation may have changed to illustrate the impact on the value of implementation efforts and the value of both information and implementation (see Appendix A).

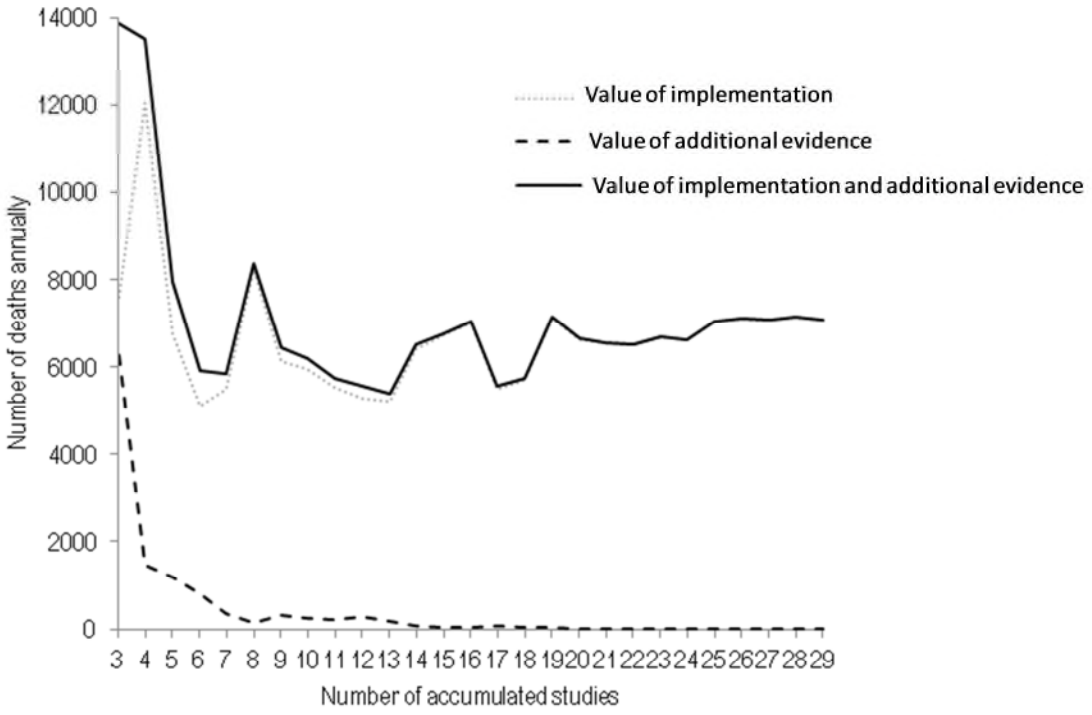


Figure 4.4 The value of implementation and the value of additional evidence (SK for acute MI)

Figure 4.4 illustrates the important distinction between these two very different ways to improve patient outcomes, which have two important implications:

i. Although the results of additional research may influence clinical practice and may itself contribute to the implementation of research finding, it is certainly not the only or necessarily the most effective way to do so. Insofar as there are other mechanisms (e.g., more effective dissemination of existing evidence) or policies which fall within the remit of other bodies (e.g., incentives and sanctions),[13] then continuing to conduct research to influence implementation rather than because there is real value in acquiring additional evidence itself (e.g., beyond study 19 in Figure 4.4) would seem inappropriate for two reasons. The limited research capacity and the resources to fund the research could have been used elsewhere to acquire additional evidence in areas where it was genuinely needed and would have offered greater potential health benefits. Secondly, such a policy will have negative health effects for those patients enrolled in research and allocated to interventions which are expected to be less effective.

ii. It also illustrates that the potential health benefits of acquiring additional evidence through conducting further research will only be realized (patient outcomes actually improve) if the findings of the research do indeed have an impact on clinical practice. Again, recognition that there are very many ways to influence the implementation of existing evidence, other than by conducting more research, is important. However, implementation of the findings of proposed research might influence consideration of its priority and research design in a number of ways. For example, if it is very unlikely that the findings will be implemented and other mechanisms and policies are unlikely to be used or be effective, then other areas of research which might offer lower potential benefits but where these benefits are more likely to be realized might be prioritized. If the impact of research on clinical practice is likely to require highly statistically significant results this will influence the design, cost and time

[13] Such bodies may not necessarily be public institutions or government agencies but might also include private for, or not for profit entities as well as professional groups and patient organisations.

taken for research to report and therefore its relative priority. It maybe that a large clinical difference would also need to be demonstrated before research would have impact on practice. This also tends to reduce the potential benefits of further research as well (see Section 4.1.3).

The analysis illustrated in Figure 4.1 and 4.4 starts to suggest that always waiting to implement research findings until the traditional rules of statistical significance are achieved may well come at some considerable cost to patient outcomes. For example, it was not until 1979 that a single trial reported a statistically significant result that favored SK *and* a random effect analysis of the cumulated evidence also suggested a statistically significant result (the 19th study, European 3). However, prior to this trial reporting the expected benefits of implementing SK for acute MI were already substantial (5,696 deaths averted per year) and the balance of previous evidence had already suggested similar positive benefits for many years.

However, there are a number of issues that need to be considered before implementation based on the balance of accumulated evidence, rather than the need for additional evidence, should be the focus of policy:

i. If earlier implementation means that the type of research required to generate the evidence needed is impossible or more difficult to conduct[14] then the uncertainty and need for additional evidence needs to be considered alongside the expected benefits of early implementation.
ii. Insofar as widespread use of an intervention will be very difficult to reverse if subsequent research demonstrates that it is not effective, then account must be taken of the consequences of this possibility (the possibility of research finding that SK is not effective is represented by the error probabilities reported in Section 4.1.1 and the consequences are the realized differences in mortality which are also available from this analysis).
iii. If an intervention needs to offer a substantial improvement in effectiveness (a minimum clinical difference – see Section 4.1.3) to justify implementation and if there remains uncertainty about whether such an improvement in effectiveness will be realized, then there are circumstances (e.g., in chronic and stable conditions) were it might be better to the delay the use of the intervention until additional evidence is available.
iv. There is a common and quite natural aversion to iatrogenic effects, i.e., health lost through adopting an intervention not in widespread use tends to be regarded as of greater concern than the same health lost through continuing to use existing interventions that are less effective than others available. However, it should be noted that the consequences for patients are symmetrical and this 'aversion' also depends entirely on which intervention just happened to have diffused into common clinical practice first.

These considerations can inform an assessment of whether more health might be gained through efforts to implement the findings of existing research or by acquiring more evidence to inform which intervention is most effective. It should be noted that these considerations are likely to differ between topic areas and certainly do not necessarily lead to a single 'rule' based on the statistical significance of the results of a particular study or a meta-analysis of existing studies.

[14] Experimental research (RCTs), for example, might be regarded as unethical and in any event might struggle to recruit. The difficulties might include additional resource costs, a greater risk of bias in estimation, or taking more time to report. Of course, there are also circumstances when evidence becomes easier to acquire once an intervention is in use as long as there is sufficient variation in treatment assignment, adequate routine data collection and availability of valid instruments to adequately account for selection bias.

4.1.3 *Minimum clinical difference*

We have restricted attention to the mortality endpoint reported in the trials and implicitly assumed that this captures the key aspects of health effects associated with early thrombolysis. Although mortality following acute MI maybe the appropriate primary outcome, it is not necessarily the only relevant outcome, for example stroke and its consequences are also very relevant as well as survival and the type of health experienced in the additional years of life associated with mortality effects. Specifying a minimum clinical difference required to change clinical practice is one way to incorporate concerns about potential adverse events and other consequences of recommending a more effective intervention (e.g., adverse impacts on health system performance), albeit implicitly (see Section 4.2.2 for an analysis which does so quantitatively.[15] A larger clinical difference in effectiveness may also be needed to be demonstrated before research will have an impact and the finding of proposed research will be widely implemented and the potential benefits of additional evidence realized.

The expected benefits of additional evidence for a range of minimum clinical differences (MCD) are illustrated in Figure 4.5 for the same three points in the sequence as were illustrated in Figure 4.2. When the MCD is close to zero the expected health benefits of additional evidence are the same as those reported in Figure 4.2, i.e. higher at points earlier in the sequence of these trials. Requiring that further research must demonstrate larger differences in effect reduces these expected health benefits because large differences are less likely to be found than smaller ones, therefore, the potential benefits of further research decline as a greater MCD is required. The rate of decline depends on the distribution of effects estimated from the meta-analysis (combining uncertainty in relative effect and baseline risk). This explains why at a MCD of a 0.07 reduction in odds of death, the expected benefits of additional evidence appears to be higher following European 3 than NHLBI. The same analysis can be reported when MCD is measured on a relative scale (e.g., odds ratio) or on an absolute scale (odds or proportionate change in deaths) or as the absolute number of deaths per year (see Appendix A), depending on which scale is more useful when considering what differences are likely to be required.

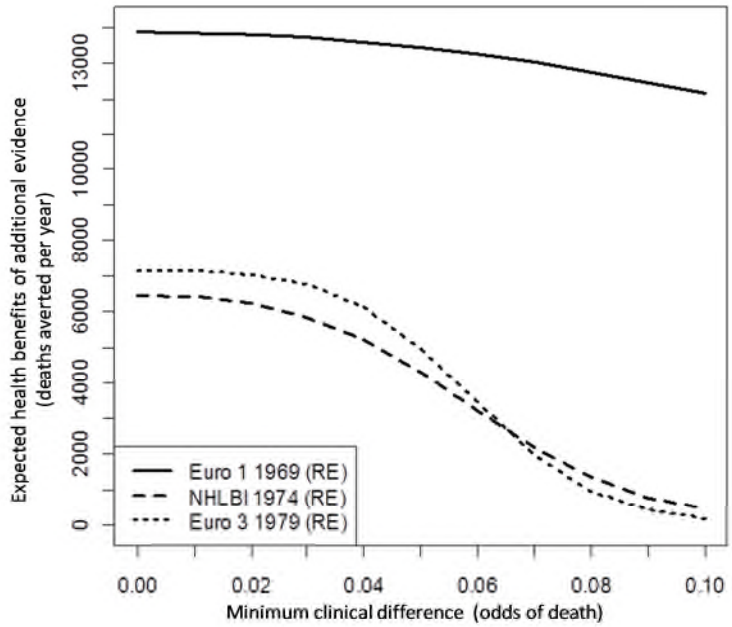


Figure 4.5 Value of additional evidence and minimum clinical difference required (SK for MI)

[15] A minimum clinical difference less than zero is possible. It would imply that reduced effectiveness on the primary endpoint would be acceptable because it would be compensated for by other types of benefits (e.g., reductions in adverse events, improvement in quality of life, reduced need of other service or a reduction in out of pocket costs for patients.

4.2 Primary endpoint linked to other aspects of outcome

Prior to the CRASH trial (Corticosteroid Randomisation After Significant Head injury) the effects of corticosteroids (CS) on death and disability following traumatic head injury was unclear, despite 19 trials conducted between 1972 and 1995. The CRASH trial, which first reported in 2004, was stopped early after enrolling 10,008 adults with traumatic head injury (THI). It reported a higher risk of death or severe disability associated with the use of corticosteroids (CS). As a consequence of this definitive, and to some extent, unexpected result clinical practice changed dramatically, resulting in many thousands of iatrogenic deaths averted around the world (prior to CRASH CS was used in 64% of patients with THI in the US and 12% in the UK). The global value of the CRASH trial appears, with hindsight, self evident. A more interesting question, however, is whether the type of analysis described in Section 4.1 would have identified this topic as of particular value and would have supported the UK Medical Research Council's (MRC) decision to fund one of the largest clinical trials it has commissioned.

The trials comparing the use of corticosteroids to placebo or no treatment in acute THI available prior to CRASH are illustrated as a standard Forest plot in Figure 4.6.

These trials were of varying study quality, length of follow-up, steroids administered, doses and time to administration. A standard meta-analysis of these trials suggests substantial uncertainty about the effectiveness of CS in THI, whether or not a fixed or random effects analysis is used.[16] For example, the random effect analysis of the mortality endpoint favors CS with an expected odds ratio for death of 0.93, but there is a 26% chance that the use of CS is in fact harmful (an odds ratio greater than 1).

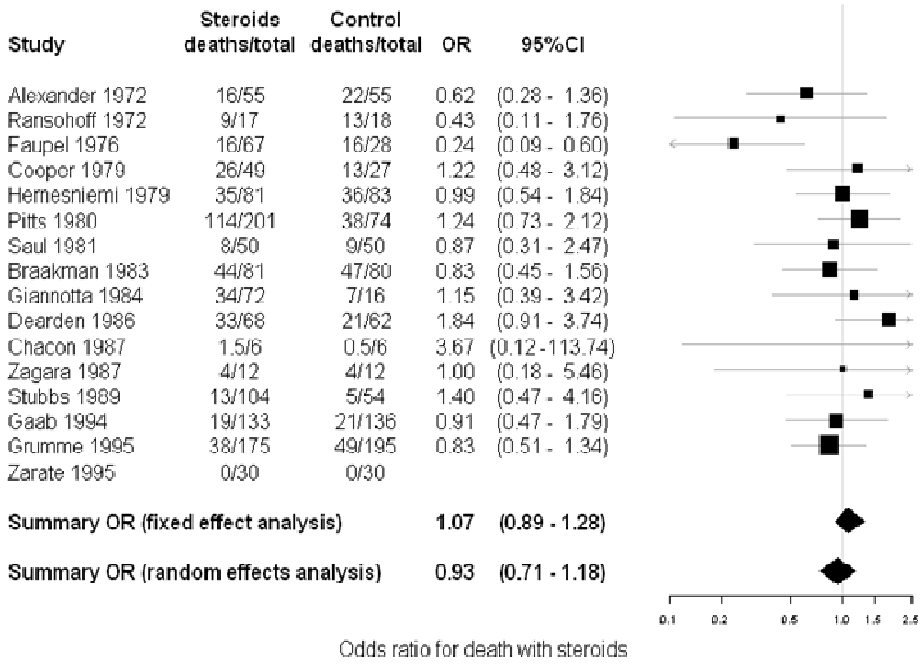


Figure 4.6 The trials of corticosteroids for traumatic head injury

[16] Note that the random effect analysis on the mortality endpoint favours CS but the fixed effect does not because it is generally the smaller trials that favour CS. This illustrates some of the dangers of using a random effect if the potential for bias is associated with study size. See Appendix B for the details of this re-analysis and the results of the original analysis published in the BMJ by those who were to become the CRASH investigators in 1997.

4.2.1 *Value of additional evidence about mortality*

Just as in Section 4.1, translating the chance that decisions about the use of CS in THI based on existing evidence will be 'wrong' into the consequences for patient outcomes requires applying the uncertain estimate of the relative effect of CS (the odds ratios reported in Figure 4.6) to an estimate of base line mortality risk in THI, which will also be estimated with uncertainty. An estimate of the incidence of severe THI is also required.[17] The combined effect of uncertainty in relative effect and baseline risk is characterized in the same way as in Section 4.1, by taking repeated random samples from the distributions estimated in the meta-analysis.[18]

The distribution of the health consequences of uncertainty is illustrated in Figure 4.7 based on the random effect meta-analysis. If the use of CS in the UK prior to CRASH had been based on the random effect analysis of evidence that favors their use, then there is a 74% chance of no adverse mortality consequences (the probability that CS is effective is 0.74). There is, however, a chance (26%) that using CS in THI will not be effective and causes iatrogenic deaths, including: a greater chance of more limited iatrogenic consequences (19% chance of greater than zero and 200 deaths per year); and a smaller chance of larger consequences (7% chance of greater than 200 deaths per year). The expected iatrogenic deaths due to this uncertainty is 40 per year (the average over the distribution illustrated in Figure 4.7), which is also an estimate of the expected health benefits that could potentially be gained each year if the uncertainty surrounding the use of CS could have been resolved at that time. It represents an estimate of the upper bound on the expected health benefits of additional evidence that would confirm whether or not CS reduces or increases mortality.

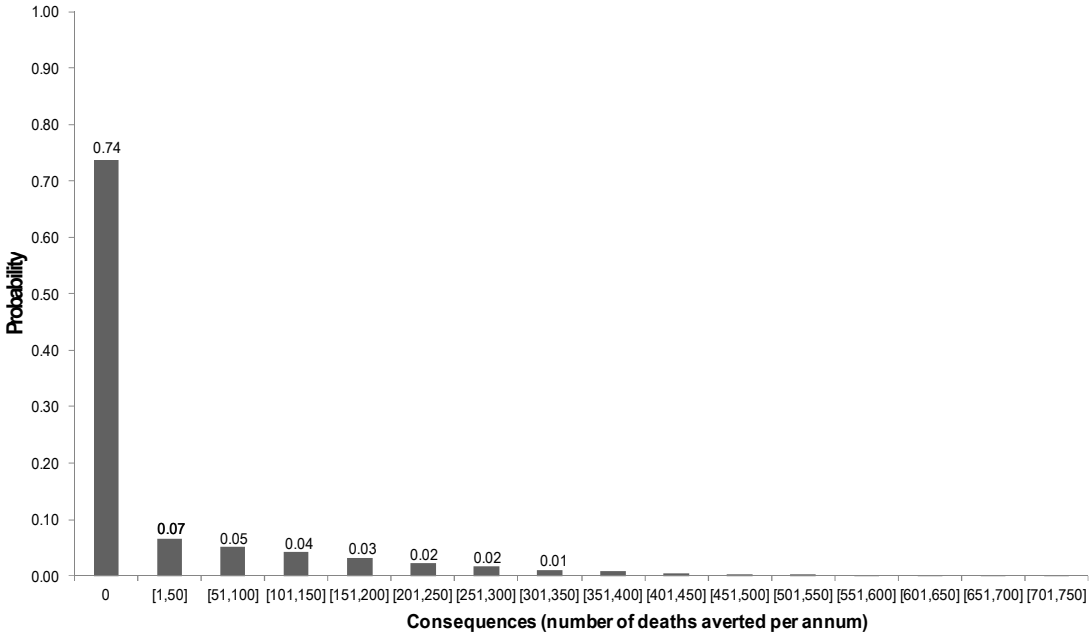


Figure **4.7** **Distribution of the consequences of uncertainty (CS for severe THI)**

[17] An estimate of the size of the patient population whose treatment choice can benefit from the information provided by further research is required. Therefore, in a chronic condition an estimate of prevalence as well as incidence would be required.

[18] In this analysis we use the control arms of the trials as an estimate of baseline risk, directly from the output of the meta-analysis; preserving the correlation in their estimation. However, an estimate of baseline risk could be based only on the control arms of those trials regarded as most relevant to the target population (see Section 4.3) or from external evidence. The estimates reported here are also based on estimates of UK incidence - see Appendix B for details.

These expected benefits might appear relatively modest compared to the value of additional evidence early in the sequence of trials of SK for acute MI in Section 4.1.[19] However, when assessing the potential value of CRASH based on this analysis it is important to consider that:

i. The subsequent trials that confirmed the effectiveness of SK for acute MI provided very valuable information that ultimately avoided many thousands of deaths, i.e., CRASH may have been less valuable than the trials of SK following European 1 but is nonetheless still valuable and was correctly regarded as a priority by the UK MRC.
ii. The value of the information that CRASH provided extends beyond the UK so the global value of this, and also the sequence of SK trials, is much greater than estimates restricted to expected benefits to UK patients.
iii. Mortality is only one aspect of outcome. Combining the impact of CS on disability and mortality changes both the estimate of its effectiveness and the uncertainty associated with its use (see Section 4.2.2)

Prior to CRASH clinical practice in the UK did not reflect the balance of evidence that favors the use of CS (based on random effect analysis), partly due to the substantial uncertainty about its effectiveness. At the time approximately 12% of patients with THI received CS. Therefore, the value of both implementing the uncertain findings of existing research *and* acquiring additional evidence that would resolve this uncertainty is greater (180 deaths per year – see Appendix B for details of this analysis which incorporates evidence about utilization).

4.2.2 *Value of additional evidence about survival and disability*

Mortality is only one aspect of outcome in THI because the impact on disability and subsequent survival is also important. Since most of the trials also report effects on vegetative and severely disabled states, described in the Glasgow Outcomes Scale (GOS), the meta-analysis can be extended to include effects on both mortality and disability. This analysis also exploits other evidence about the distribution of GOS states following THI, movements between them (i.e. the possibility of recovery or deterioration over time) and life expectancies given survival in a particular GOS state, as well as estimates of the quality of life associated with different levels of disability described in the GOS (see Appendix B for details of this extension to standard meta-analysis). This allows the different aspects of health outcomes in THI to be combined into the equivalent years of full health, i.e., the impact on life years expected to be lived due to the effects on mortality risk adjusted for the quality in which they are likely to be lived.

Interestingly, linking the primary mortality endpoint to these other important aspects of outcome changes both the estimate of the effectiveness of CS and the uncertainty associated with it. Based on a random effect analysis, the existing evidence no longer favors CS. This is because the evidence from the trials suggested that the use of CS was also associated with greater proportion of patients surviving in a vegetative or severely disabled state, i.e., on balance CS appeared to increase the risk of severe disability.

The distribution of the health consequences of the uncertainty associated with a more complete picture of the effectiveness of CS is illustrated in Figure 4.8.

This analysis of the effects of CS on a more complete measure of important aspects of health outcome suggests that on balance CS was not expected to be effective and its effects were more uncertain. If the use of CS in the UK had been based on this evidence (no CS use for THI) then there

[19] One reason why the expected benefits are lower is that the size of the patient population that can benefit from information about the use of CS for THI is lower than acute MI.

would be a 63% chance of no avoidable lost years lived in full health (probability CS is not effective is 0.63). There is, however, a chance (37%) that using CS in THI would have been effective and improve health outcomes, including: a greater chance of more limited potential gains from the use of CS (23% chance of between zero and 3000 equivalent years of full health gained per year) and a smaller chance of larger potential benefits (14% chance of more than 3000 equivalent years of full health gained per year). The expected consequences of this uncertainty or the potential health gains of resolving it through additional evidence is 1,067 equivalent years of full health gained each year in the UK.

At the time 12% of patients with THI received CS in the UK. Therefore, the value of both implementing the uncertain findings of existing research (not to use CS for THI) *and* acquiring additional evidence that would resolve this uncertainty is greater (1,264 equivalent years of full health gained each year – see Appendix B).

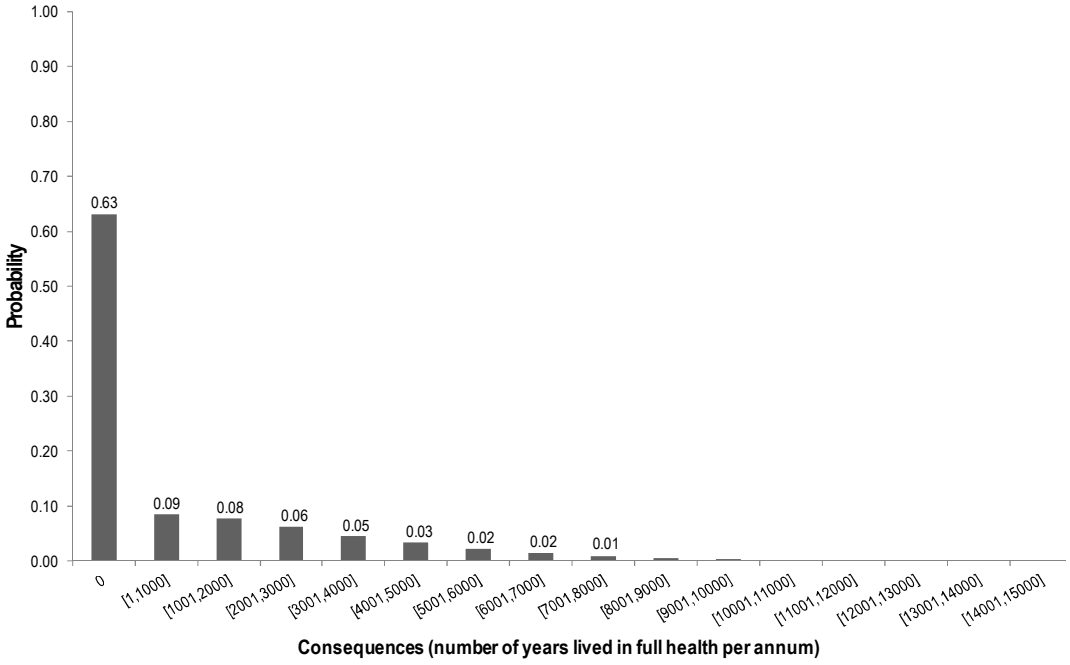


Figure 4.8 Distribution of the consequences of uncertainty (CS for severe THI)

4.2.3 *Assessing the value and priority of proposed research*

Two questions are posed when considering whether the decision to prioritize and commission CRASH was appropriate:

i. Are the expected health benefits of additional evidence of 1,067 equivalent years of full health each year sufficient to regard CRASH as potentially worthwhile?
ii. Should it be prioritized over other research topics that could have been commissioned with the same resources?

These assessments require some consideration of the period of time over which the additional evidence generated by research is likely to be relevant and can inform clinical practice; as well as the time likely to be taken for proposed research to be commissioned, conducted and report.

The information generated by research will not be valuable indefinitely, because other changes occur over time, which will have an impact on the future value of the information generated by research that can be commissioned today. For example, over time new and more effective interventions become available which will eventually make current comparators obsolete, so information about their effectiveness will no longer be relevant to future clinical practice. Other information may also become available in the future which will also impact on the value of the evidence generated by research that can be commissioned today. For example, other evaluative research might be (or may already have been) commissioned by other bodies or health care systems, that may resolve much of the uncertainty anyway. Also, this research or other more basic science may fundamentally change our understanding of disease processes and effective mechanisms. Finally, as more information about individual effects is acquired through greater understanding of the reasons for variability in patient outcomes, the value of evidence that can resolve uncertainty in expected or average effects for the patient population and/or it's sub populations will decline (see Section 3.4). For all these reasons there will be a finite time horizon for the expected benefits of additional evidence, i.e., there will be a point at which the additional evidence that can be acquired by commissioning research today will no longer be valuable.

The actual time horizon for a particular research proposal is unknown, because it is a proxy for a complex, and uncertain process of future changes. Nonetheless some judgment, whether made implicitly or explicitly, is unavoidable when making decisions about research priorities. Some assessment is possible based on historical evidence and judgments about whether a particular area is more likely to see future innovations, other evaluative research and the development of individualized care (e.g., where diagnostic technologies, application of genomics, and the development of evidence based algorithms are rapidly developing). Information can also be acquired about trials that are already planned and underway around the world (e.g., various trial registries) and future innovations from registered patents and/or phase I and II trials as well as licensing applications, combined with historic evidence on the probability of approval and diffusion. For these reasons, an assessment of an appropriate time horizon may differ across different clinical areas and specific topics and research proposals. The incidence of patients who can benefit from the additional evidence may also change over time, although not necessarily decline as other types of effective health care change competing risks. However, in some areas recent innovations might suggest a predictable decline, e.g., the decline in the incidence of cervical cancer following the development of HPV vaccination.

The CRASH trial was proposed to the MRC in 2000. A time horizon of 15 years may have been a reasonable but conservative judgment at the time, given that there were no other trials underway and previously few major innovations which had transformed the treatment or understanding of THI.[20] The implications for an assessment of the overall expected benefits of CRASH are illustrated in Figure 4.9.[21]

CRASH was not expected to report before 2004 when the additional evidence it would provide could inform treatment choice for patients (between 2004 and 2015). Therefore, the overall (undiscounted) expected health benefits were an additional 10,266 years lived in full health in the UK. In the UK context both health benefits and NHS costs are discounted at the same rate of 3.5%

[20] UK incidence of severe THI in 2000 is used in this analysis, although predictions of changes in THI between 2000 and 2015 could have been used if they had been available in 2000.

[21] Some assessment of the chance that proposed research will be successfully conducted (e.g., able to recruit), be of sufficient quality to be useful and ultimately report is also needed (see Appendix B). Partly this is an assessment of quality of the proposal and the capacity of the research team, but it will also depend on other future events outside their control, e.g., a new and effective innovation may make it very difficult to complete recruitment into a trial.

specified by the UK Treasury,[22] so it is the discounted value of 8,946 years of full health that is most relevant. The question remains, however, whether these expected benefits were sufficient to justify the expected costs of this trial (£2.2m) and whether it represented a particular priority compared to the other research that could have been commission by the MRC using the same resources?

One way to start to address this question in the UK is to ask whether the NHS could have generated similar expected heath gains more effectively elsewhere, or equivalently whether the costs of the CRASH trial would have generated more health benefits if these resources had been made available to the NHS.

Very recent work in the UK, also commissioned by the MRC, has estimated the relationship between changes in NHS expenditure and health outcomes. This work suggests that the NHS spends approximately £75,000 to avoid one premature death, £25,000 to gain one life year and £20,000 to gain one quality adjusted life year. Using these estimates the costs of CRASH could have been used to avoid 29 deaths and generate 110 quality adjusted life years elsewhere in the NHS – substantially less than the expected health benefits of the CRASH trial. Alternatively the NHS would have to spend an additional £179m between 2004 and 2015 to generate expected health benefits similar to those offered by CRASH. This strongly suggests that the CRASH trial was indeed worthwhile at the time it was commissioned.

However, since the MRC itself has limited resources and cannot draw directly on the NHS budget, it is possible that other research proposed in 2000 may have been even more valuable than CRASH. Without a similar reanalysis of rejected research proposals at the time it is not possible to confirm that CRASH offered the greatest value. If similar analysis was conducted for all topics competing for limited research resources it does become possible to identify a short list of those which are likely to be worthwhile and then periodically select from this shortlist those that are likely to offer the greatest value.[23]

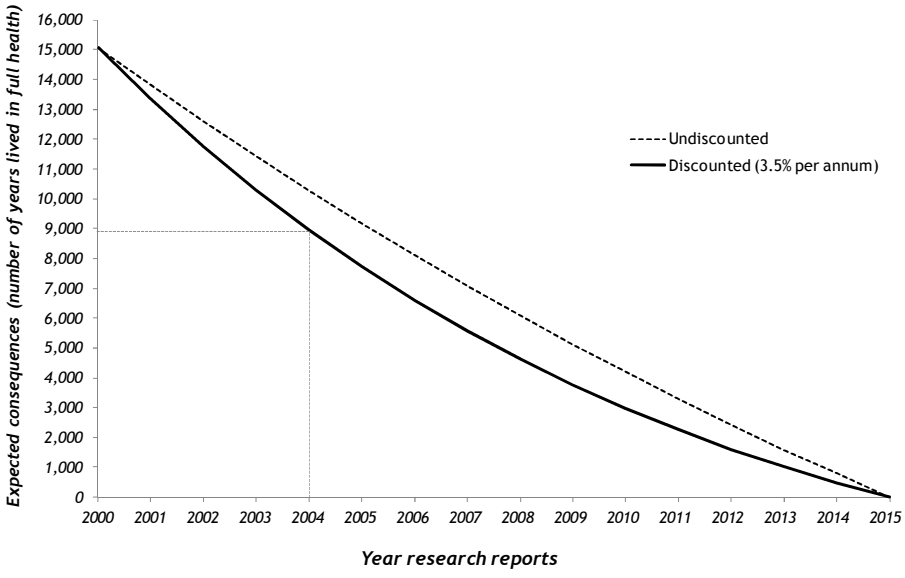


Figure 4.9 Time horizon and the value of proposed research

[22] Discounting is used to reflect the fact that NHS resources committed today could have been invested at a real rate of return to provide more resource in the future that could be used to generate more health.

[23] Unless research funding is especially generous it is unlikely that all proposed research that is likely to be worthwhile can be funded with available resources. Therefore, there is a danger that funding all proposals that appear worthwhile when they are submitted will mean that resources are insufficient to fund more valuable research that might be proposed later in the budget cycle.

4.2.4 *Informing research design*

Once a single primary endpoint in the existing clinical trials is linked to other secondary endpoints and external evidence, the same type of analysis can be extended to identify which sources of uncertainty are most important and what type of evidence is likely to be most valuable. This is useful in two respects. It can help to identify the type of research design that is likely to be required (e.g., an RCT may be needed to avoid the risk of selection bias if additional evidence about the relative effect of an intervention is required) and identify the most important endpoints to include in any particular research design. It can also be used to consider whether there are other types of research that could be conducted relatively quickly (and cheaply). The results of which might confirm whether or not more lengthy and expensive research (e.g., a large RCT) is really needed, i.e., it can inform the sequence of research that might be most effective.

This type of analysis has been conducted for the CS case study and is reported in Appendix B. The different sources of uncertainty include: i) the effect of CS on the risk of death; ii) the effect of CS on the risk of disability and severe GOS states; and iii) the uncertainty in the health-related quality of life associated with GOS outcomes. The analysis suggests that it is additional evidence about the combined effect of CS on the risk of death and severe disability (vegetative and severe disability in the GOS) that is of value. It indicates that future research that only included mortality as an endpoint would be of limited value but that including the less severe GOS states as end points would offer little additional value. There is very little value in additional evidence about the quality of life associated with GOS states. Therefore, delaying a future trial, until a relatively rapid study of the health related quality of life associated with GOS outcomes is conducted, does not seem appropriate in this case.

It is also possible to extend this analysis even further and consider the optimal sample size of a future trial, which included mortality and the severe GOS states as endpoints, based on the benefits of additional sample information and the costs of acquiring it - including the additional delay to reporting findings and the additional resources required. This analysis is not reported in Appendix B, but is available from the authors. It suggests that the original sample size for CRASH may have been too large and most of the uncertainty could have been resolved more quickly and cheaply with a smaller trial. However, if the implementation of research finding depends on the trial reporting a statistically significant result for the effect size specified in the original research proposal, the sample size for CRASH does indeed appear appropriate.

4.2.5 *Impact of commissioned research*

Research prioritization decisions require an assessment of the *expected* potential value of future research, i.e., before the actual results that will be reported in the future are known. Therefore, using hindsight to inform research prioritization decisions is inappropriate for two reasons: i) such an (ex-post) assessment cannot directly address the (ex-ante) question posed in research prioritization decisions; and ii) assessing the (ex-post) value of research with hindsight is potentially misleading if used to judge whether or not the original (ex-ante) decision to prioritize and commission it was appropriate. This is because the findings of research are only one realization of the uncertainty about potential results that could have been found when the decision to prioritize and commission research must be taken.

For example, it would be inappropriate to judge whether or not the MRC made the right decision to commission CRASH in 2000 based on the results that were reported in 2004 and 2005. It may be tempting to calculate the number of iatrogenic deaths avoided because the results of CRASH found CS to be harmful. However, if at the time CS was not used at all in THI, then the realised value of CRASH, with hindsight, would have been zero, because it would not have changed clinical practice at all (it's real value would have been in confirming that it was correct). In these circumstances the

MRC would not have been mistaken in its research prioritization decision, because at the time it was also possible that CRASH might have reported a reduction in mortality and disability due to the use of CS. Since at the time only 12% of UK patients were treated with CS this result would, again with hindsight, have had a greater impact on mortality. Therefore, the appropriate assessment of the expected or potential value of research or a judgment about the quality of research decisions cannot be based on hindsight but on prior or ex-ante assessments which can be informed by the type of analysis set out above.

It is useful and instructive, however, to reconsider the analysis set out above once the results of CRASH became available by updating the meta-analysis and the estimates of the expected potential benefits of further research (see Appendix B).[24] When the results of the CRASH trial are included in the meta-analysis the chance that CS improves mortality is effectively zero (<0.0001) and the chance that CS improved survival and quality of life was also almost zero (probability of 0.005). Therefore, when the analysis of the potential value of additional evidence in Sections 4.2.1 and 4.2.2 is updated, there are no expected benefits of acquiring additional evidence.[25] In this sense CRASH was a definitive trial; appropriately prioritized and commissioned at the time.[26]

4.3 Different weights to reflect the relevance of evidence

The evidence about the effectiveness of probiotics (PB) for the prevention of infectious complications in severe acute pancreatitis (SAP) is mixed and limited. Two earlier and small trials in 2002 and 2007 reported (non-significant) improvements in outcomes associated with the use of PB. However, a much larger trial reported a significantly increased risk of mortality associated with PB use in 2008. These conflicting findings prompted a debate about the benefits of PB and the relative merits of these trials. A meta-analysis performed in 2009 concluded that "Future large-scale, high-quality, placebo-controlled, double-blind trials are still required to clarify the issues of the effect of probiotics in severe acute pancreatitis". The conflicting findings of these three trials can be represented as a standard forest plot reported in Figure 4.10.

There are important differences between the two early trials and the Besselink study, other than their sample size, including: quality; follow-up; and relevance to clinical practice and the patient population in the UK and the Netherlands. For this reason there appears to be important sources of heterogeneity between these three studies. However, applying a random effect meta-analysis to try and reflect these sources of heterogeneity simply down weights the evidence of the larger Besselink trial and perversely suggests that the balance of this evidence favors the use of PB albeit with considerable uncertainty. On the other hand, applying a fixed effects meta-analysis would suggest that PB is harmful but would make the unlikely assumption (given the characteristics of these trials – see Appendix C) that each of these trials are estimating the same underlying effect. This apparent paradox illustrates that the use of a random effect meta-analysis is not an appropriate way to account for reasonably held scientific value judgments about differences in the potential for bias and relevance across studies (see also Section 4.2.4). Indeed it will tend to generate perverse results if

[24] Although the trials prior to CRASH were synthesised in a random effect meta-analysis, the results of CRASH (which was designed and commissioned to be of high quality and directly relevant to clinical practice and the target patient population) enter as a fixed effect (prior based on random effects is updated using a fixed effect). It would be inappropriate to enter CRASH as random effect which would down weight its results by imagining that CRASH, like the previous trials, was randomly drawn from the same population of previous studies

[25] If CS is no longer used in THI following CRASH there are no expected mortality benefits of additional evidence and the very small benefits in terms of quality of life (3.2 years of equivalent full health) – see Appendix B.

[26] This confirms the findings discussed at the end of Section 4.2.4 - that more evidence than was strictly necessary was generated by CRASH, because an optimal amount of information would leave some value of additional evidence but not sufficient to justify further research. However, if the implementation of research findings depends on delivering a statistically significant result, and other ways to implement its finding were costly or unavailable, then the undoubtedly definitive result of the CRASH trial may well have been worthwhile.

the larger studies also tend to be regarded as more relevant and of higher quality. Recent work has demonstrated how explicit weights to reflect such judgments can be elicited from relevant experts and integrated into meta-analysis. Here we do not conduct elicitation; instead we conduct the analysis for a range of weights that might reflect a particular view or the outcome of a deliberative process.

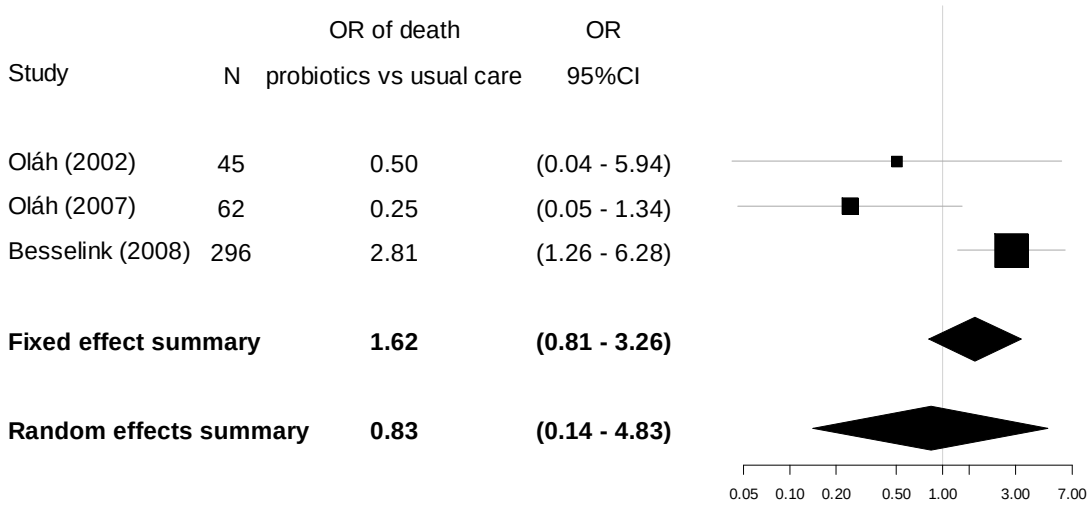


Figure 4.10 The trials of probiotics (PB) for severe acute pancreatitis (SAP)

4.3.1 *Reflecting the relevance of evidence in meta-analysis*

The meta-analysis can be conducted for each set of weights that might be attached to the two earlier trials (in additional to the precision of their summary estimates). The impact of weights on the central estimate of the odds ratio for PB and its confidence interval are reported in Figure 4.11 for the fixed effect meta-analysis.[27]

For example, a judgment that the previous smaller trials were of lower quality and less relevant could be reflected in a weight less than 1. In which case there would be greater confidence that the use of PB in SAP is not effective and potentially harmful. Indeed, if these two early trials are down weighted by 0.5 or more then existing evidence provides a statistically significant result against the use of PB. Figure 4.11 also exposes what implicit weight must be placed on the earlier evidence to believe that the balance of this evidence favors the use of PB (a weight greater than 2.5 on the earlier trials would be required). It shows that scientific value judgments about the quality and relevance of evidence not only changes the estimate of effectiveness but also the uncertainty associated with it – and therefore the value of additional evidence and need for further research.

[27] The results for a weighted random effects meta-analysis are reported in Appendix C. However, if the weight selected accounts for the apparent heterogeneity between the trials then including a random effect in the synthesis will overestimate uncertainty and the need for additional evidence. For this reason we focus here on the results of the weighted fixed effect analysis.

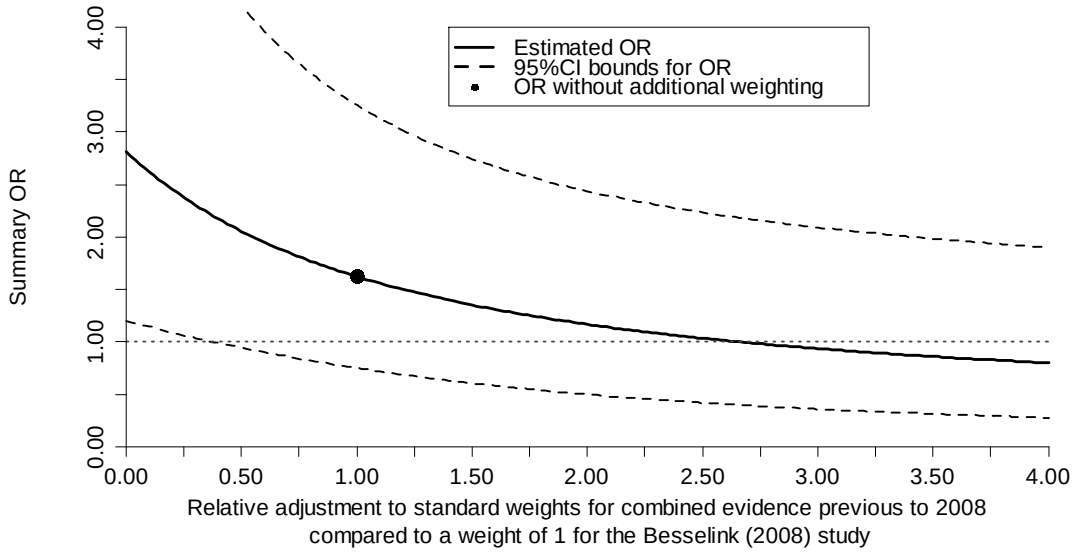


Figure 4.11 Reflecting the relevance of evidence in meta-analysis (probiotics for SAP)

4.3.2 *Implications for the value of additional evidence*

The implications for the assessment of the expected health benefits of additional evidence about the effectiveness of PB for SAP are illustrated in Figure 4.12 (expressed as deaths averted per year in the UK). These potential benefits of further research are calculated for three different weights (0.5, 1 and 1.5) for the small trials and for a range of minimum clinical differences which might be required before PB would be implemented for SAP (expressed as a percentage reduction in the absolute mortality risk). The fixed effect analysis in Figure 4.12(a) suggests that even if very modest differences are required to implement PB use, the value of additional evidence is very limited and unlikely to justify further research in this topic (3 deaths avoided per year with a weight of 1 and 1 death per year with a weight of 0.5 on the earlier trials). Even if a weight of 1.5 is assigned to the earlier trials, because they are regarded as of higher quality and of greater relevance than Besselink, the upper bound on the potential value of future research is 8 deaths averted per year. The conclusions of the 2009 meta-analysis, quoted above appears misplaced, focusing as it does on the precision of estimation and the 'rules' of statistical significance rather than the consequences of uncertainty and the potential health benefits of being able to resolve it.

Importantly, however, these results contrast sharply with those results reported in Figure 4.12(b) when a random effect is applied to this weighted evidence. As discussed above if the weights assigned to different studies accounts for the apparent heterogeneity between the trials then including a random effect in the synthesis will overestimate uncertainty and the need for additional evidence. For this reason there is a real danger that inappropriate use of random effect analysis, to account for what are really judgments about the quality and relevance of different studies, will wrongly suggest that further research is needed. For example, in Figure 4.12(b), conducting a standard random effect analysis (with a weight of 1 on the earlier studies) rather than explicitly weighting the earlier trials (a weight of 0.5 in (a) above) would seriously overestimate the value of additional research. In fact, it would suggest that additional research in this area would have higher value than the CRASH trial in Section 4.2, even if a minimum clinical difference of 4% reduction in mortality was required. The implications are that the standard application of random effect meta-analysis should not be used as a substitute for considering explicit scientific value judgments about the sources of heterogeneity between different studies and their relevance to clinical practice and the target population. Even if explicit elicitation and/or analysis based on a range weights is not

conducted, careful examination of the impact of a random effect on estimates of uncertainty and the expected benefits of additional evidence is important.

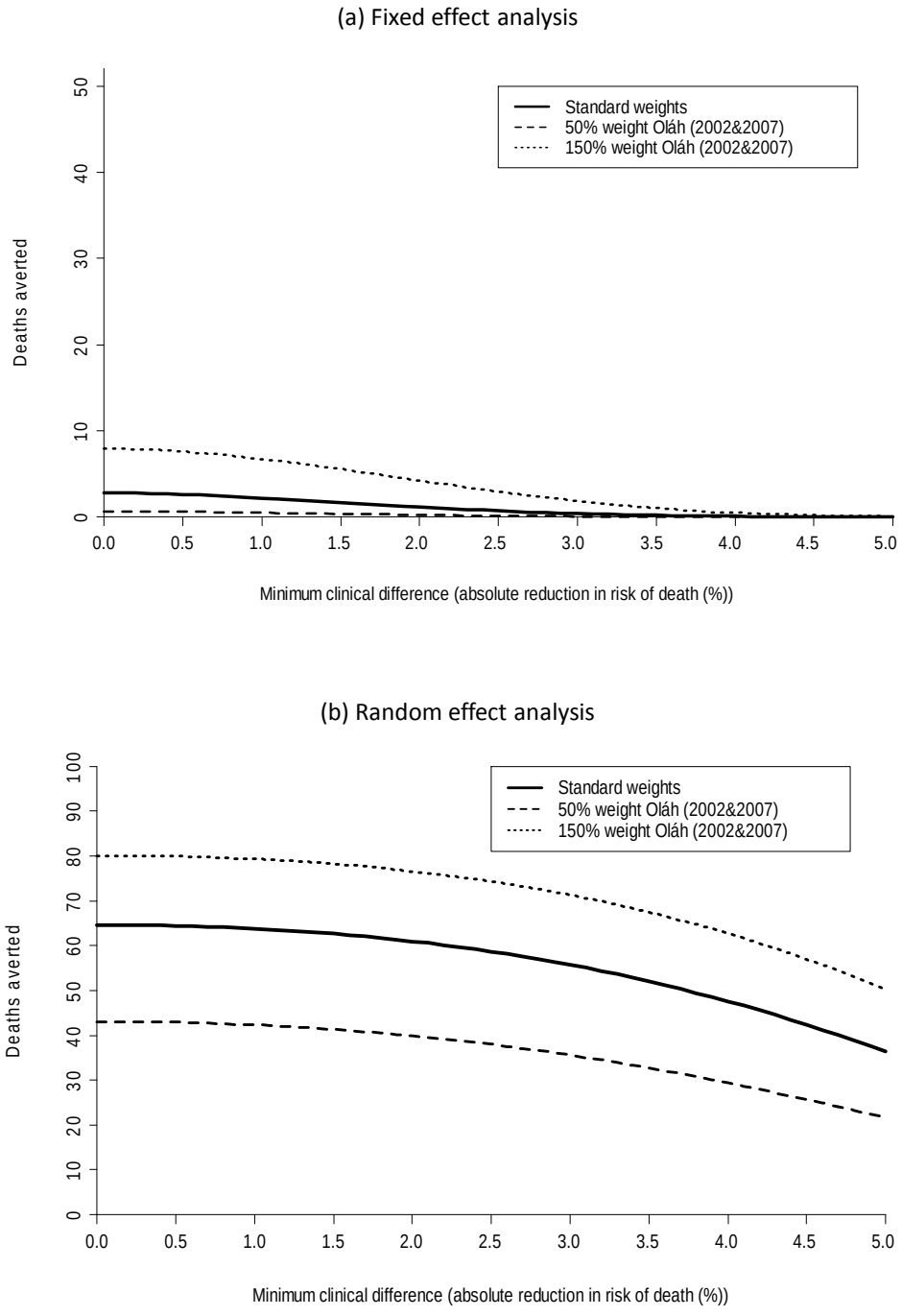


Figure 4.12 Expected health benefits of additional evidence (probiotics for SAP)

4.4 More than two alternative interventions need to be compared

In previous sections the case studies offer examples where the choice of which intervention to select are restricted to two alternatives (or have been simplified to be so for ease of exposition – see Section 4.1). Commonly, however, there are a number of alternative interventions available, only

one of which can be selected as the appropriate treatment for a patient (i.e., they are mutually exclusive). For this reason an intervention is better thought of as a strategy of treatment rather than a particular technology or single procedure. For example interventions might include different combinations or sequences of treatments and/or the same treatment but with different dosage or starting and stopping rules. Here we consider how the methods of analysis illustrated in previous Sections can be applied and interpreted when there are multiple alternative interventions which could be used to improve patient outcomes in a specific indication.

For example, when Topotecan and pegylated liposomal doxorubicin hydrochloride (PLDH) became available as second line treatments of advanced ovarian cancer in 2005, it posed the question of whether they should be used rather than paclitaxel which had been available for some time. The key question was which of the three alternative interventions is likely to be most effective in promoting survival and whether the evidence for improvements in survival is sufficient to justify the use of Topotecan (Top) or PLDH for this indication.

At the time, the evidence of effectiveness for these three alternative interventions was limited to three trials which made different comparisons, all of which are relevant to the question of which intervention was most effective. This 'network' of evidence is described in Table 4.1.

Table 4.1 Network of evidence for second-line treatment of advanced ovarian cancer

Trial ID	Alternative interventions for advanced ovarian cancer		
	Paclitaxel (Pac)	Topotecan (Top)	PLDH
039	53.0 (n=114)	63.0 (n=112)	-
30-49	-	59.7 (n=235)	62.7 (n=239)
30-57	56.3 (n=108)	-	46.6 (n=108)

Median weeks survival (number of patients analyzed)

The question was whether:

i. The use of these new technologies (Top or PLDH) should be encouraged, if on the balance of evidence they appear more effective than paclitaxel (Pac), but accept the chance that the other interventions maybe more effective; or
ii. Wait until additional evidence can confirm which intervention is most effective but accept that the intervention, which on the balance of evidence expected to be most effective, may not be widely used until further research reports.

4.4.1 *Synthesis of evidence with multiple comparisons*

The network of evidence in Table 4.1 offers a confusing and inconsistent picture of the relative effectiveness of Top and PLDH compared to Pac. The observed trial results suggest that either:

i. Top is more effective than Pac (trial 039) *and* PLDH is more effective than Top (trial 30-49), which means that PLDH must logically be more effective than Pac. However, evidence from trial 30-57 contradicts this and would need to be set aside.
ii. Alternatively, Pac is more effective than PLDH (trial 30-57) *and* Top is more effective than Pac (Trial 039), which means that Top must logically be more effective than PLDH. However, the evidence from the largest of the three trial (30-49) contradicts this.

In this example the observed evidence is inconsistent (i.e. not transitive) even when only considering the qualitative question of which intervention is expected to be most effective. Even if the results were transitive when considering this qualitative question, the difficulty of inconsistency is likely to remain when considering the quantitative questions of the magnitude of relative effects and the uncertainty about any overall assessment of relative effectiveness. There are three possible approaches that can be taken:

i. Attention could be restricted to the three pair-wise comparisons made in the existing trials. The major difficulty is that as soon as the effects on survival are compared across these three separate pair-wise analyzes randomization is implicitly broken, i.e., this comparison is equivalent to comparing the intervention and control arms of different trials. To be appropriate this would require an heroic assumption that patients in the control and intervention arms of different trials were on average comparable in both observed and unobserved respects, i.e., it assumes away the very reason for randomized design in the first place. The other difficulty is that such an assessment would be inconsistent. It would generate three different estimates of effectiveness each associated with ignoring one of the three trials. Any view about effectiveness and the uncertainty associated with it would require at least one of the trials to be ignored.

Over recent years methods to extend meta-analysis to a more generalized form of evidence synthesis has developed rapidly. There are two evidence synthesis approaches that could be taken in response to this network of evidence in Table 4.1:

ii. A meta-analysis which made indirect treatment comparisons (ITC) between all three alternatives based on a single common comparator could be undertaken. Unlike pair-wise comparisons this would respect randomization, making very similar assumptions as a standard pair-wise meta-analysis. In this case, there is no single comparator that is common to all three trials, therefore there are three possibilities:
   - Pac is the common comparator in trials 039 and 30-57 but applying an ITC to this network would be inappropriate as trial 30-49, would be excluded. This trial provides the most information, so its exclusion would bias the estimate of effectiveness and over estimate the uncertainty associated with it.
   - PLDH is the common comparator for trials 30-49 and 30-57, but applying an ITC to this network would exclude trial 039 which provide more information than trial 30-57.
   - Top is a common comparator for the two largest trials, 039 and 30-49, but an ITC based on Top would exclude the information in trial 30-57.

   Although all of these possibilities are inappropriate because they exclude potentially relevant evidence, we include an ITC based on using Top as a common comparator for purposes of illustration (it makes use of the two largest trials) and comparison with the more appropriate mixed treatment comparison.

iii. Mixed treatment comparison (MTC) can exploit the full network of evidence, making use of direct evidence, common comparators and the indirect comparisons from combinations of studies. For example, when comparing PLDH to Top the information from trial 30-49 (PLDH vs Top) provides direct evidence. However, the combination of 039 (Top vs Pac) and 30-57 (PLDH vs Pac) tells us something about PLDH vs Top, based on the common comparator of Pac. In addition, both 039 (Top vs Pac) in combination with 30-49 (PLDH vs Top) and 30-57 (PLDH vs Pac) in combination with 30-49 (PLDH vs Top) also tells us something about PLDH

compared to Top. Again this type of analysis respects randomization making very similar assumptions to standard pair-wise fixed or random effect meta-analysis.[28]

Inconsistency in evidence, just as in standard pair-wise meta-analysis, is dealt with statistically in ITC and MTC. There may, however, be good reasons why one or other trial is more likely to be inconsistent than by chance under fixed or random effects (as in all meta-analysis larger studies carry more weight so will tend to force others to appear inconsistent with the overall results of MTC).

This is illustrated graphically in Figure 4.13 where the width and shading of each arrow is proportional to sample size and the direction indicates which intervention is favored by each pair-wise comparison (which in this example is a single trial). It shows that among the pair-wise comparisons the strength of evidence was greatest for the comparison of Top and PLDH and the weakest for PLDH and Pac. Therefore, the weight of evidence will tend to force consistency through the weakest link (the PLDH vs Pac comparison). MTC does this statistically (some weight on the PLDH vs Pac evidence) rather than by simply ignoring trials in the ITC (giving zero weight) or by ignoring trials *and* breaking randomization when pair-wise comparisons are used to assess comparative effectiveness.

There may be good reasons why certain trials are less relevant or more likely to offer a biased estimate of effects in the target population than others. In these circumstances, careful consideration of the quality and relevance of the particular trials is needed, rather than allowing statistical analysis to substitute for judgments about clinical epidemiology (see Section 4.3 and the discussion of random effects). Although explicitly weighting the quality and relevance of evidence is possible, it is also instructive to conduct simple sensitivity analysis by comparing the implications and inconsistencies associated with simple multiple pair-wise analyzes with MTC and ITC while re-estimating relative effects when dropping potentially inconsistent trials from the network in turn. Consideration of which analysis best represents a reasonable view of the potentially disputed scientific value judgments is for a deliberative process, albeit it one that is informed about the nature of the particular studies and the implications of coming to alternative views.

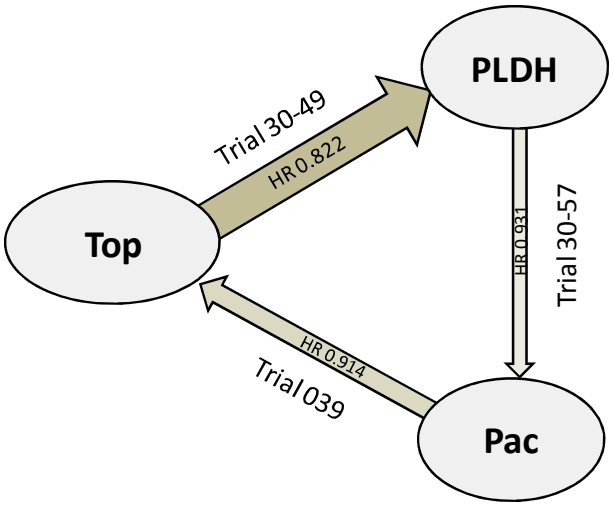


Figure 4.13 Strength of evidence for separate pair-wise comparisons

[28] Commonly, pair-wise, ITC and MTC meta-analysis assume that relative effects on the chosen scale are invariant with respect to the baseline. If this assumption is thought to be unreasonable then it poses the same difficulties for pair-wise as ITC and MTC. Generalization through meta-regression is available which in principle could be used to estimate the relationship between baseline and relative effect.

4.4.2 *Implications for the value of additional evidence*

The details of the three forms of meta-analysis (pair-wise, ITC and MTC) are detailed in Appendix D. Both the ITC and MTC suggest that PLDH is expected to be the most effective at extending survival in advanced ovarian cancer. However, they differ in the estimated magnitude of the effect, the uncertainty associated with it and whether or not Top is more effective than Pac. The implications of these different types of meta-analysis for an estimate of the expected health benefits of additional evidence are illustrated in Figure 4.14.

An analysis restricted to the pair-wise comparisons of Top and PLDH with Pac (i.e., that forced consistency by ignoring trial 30-49) provides an estimate of the expected benefits of evidence about Top compared to Pac and of PLDH compared to Pac in Figure 4.14(a). At the time, with the limited network of evidence, there was substantial uncertainty about the most effective intervention and this is reflected in the estimates of the potential benefits of further research, e.g., the expected health benefits of resolving the uncertainty about Top compared to Pac are 305 deaths averted per year. This remains high even if a substantial reduction in mortality risk would be required to change clinical practice from Pac to Top. Unsurprisingly, the value of additional evidence about PLDH is lower (it is much less likely to be the most effective intervention based on trial 30-57 alone). It is tempting to sum these values to get an indication of the overall value of additional evidence and the potential value of further research in the topic. However, the results reported in Figure 4.14(b) based on the MTC analysis show that this is likely to overestimate the potential value of further research (potential benefits of 227 death averted per year based on MTC if the minimum clinical difference required to change practice is close to zero). This should not be surprising as MTC includes all the evidence available in the network and borrows strength in estimating effects from all the direct and indirect comparisons. Including the expected benefits of additional evidence based on the pair-wise PLDH/Top comparison, which was excluded from Figure 4.14(a), would only increase this overestimation. The results in Figure 4.14(b) demonstrate the implications of using an ITC, which inappropriately excludes trials that cannot be linked to a common comparator. Again, as expected the uncertainty is overestimated and the potential value of further research is also overestimated.

Finally it should be noted that a comparison of the expected benefits of additional evidence across these topics examined in previous sections is possible because at some point the expected benefits can be expressed in terms of deaths averted. Such comparison should, however, be interpreted with caution because the overall health effects depend, among other things, on how long a death is likely to be averted for and what the health experience of any extra years of survival are likely to be. Unfortunately, in advanced ovarian cancer averting a death does not necessarily extend life to any great extent compared to avoiding mortality in acute MI or surviving in a healthy or mildly disabled state following THI. Therefore, in this example the relatively high potential value of further research, expressed in deaths averted, does not necessarily mean that it should have been regarded as a higher research priority than SK for acute MI or CS for THI. A more comparable picture of the relative value of additional evidence across these topics requires the endpoints that only capture some aspects of health outcome to be linked to others to provide a more complete picture of health outcomes; whether this is done quantitatively in explicit analysis (see Section 4.2) or considered more implicitly, possibly expressed as different MCDs judged to be required through a deliberative process.

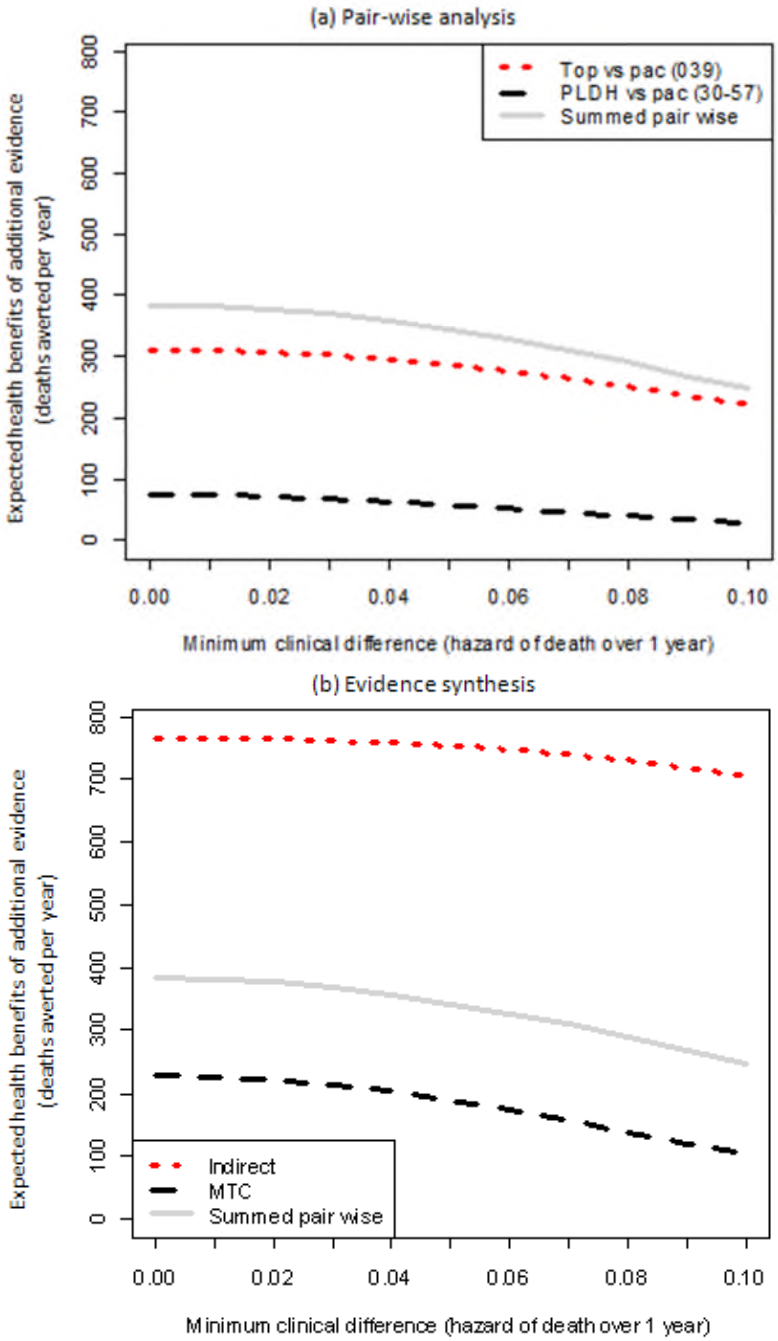


Figure 4.14 Expected health benefits of additional evidence (Top and PLDH for ovarian cancer)

5. Considerations

The previous sections have illustrated how explicit and quantitative analysis, based on systematic review and meta analysis, can be used to inform the assessments that need to be made when considering research prioritization and commissioning. Nonetheless, no quantitative analysis, no matter how assiduously conducted or sophisticated, can capture all aspects of scientific and social value relevant to making decisions about research priorities. Not least because both scientific and social value judgments are quite reasonably disputed. The more relevant question is whether they offer a practical and useful starting point for deliberation and add to the transparency and accountability of the decision making process. It is on this basis that the usefulness of more explicit and quantitative analysis ought to be judged.

We believe their potential usefulness has been demonstrated, as long as the results are appropriately interpreted and the limitations of the analysis are understood and explored. The more difficult question is their practicality and feasibility within the time and resource constraints of a deliberative process of research prioritization. We have taken it for granted that any topic suggestion or specific research proposal will include a systematic review of existing evidence and, where appropriate, a meta-analysis, since funding additional research without knowledge of existing evidence would seem inappropriate and potentially unethical for experimental research designs. Although extending meta-analysis in the way described above is not technically challenging, nor does it pose particular computational problems, there is an issue of familiarity with these principles and methods (the principle that it is the consequences of uncertainty that matters and mainly methods of simulation) amongst those who most commonly conduct systematic review. Some of the contexts examined also required more sophisticated forms of meta-analysis, e.g., Bayesian meta-analysis to link multiple endpoints in Section 4.2 and a MTC evidence synthesis in Section 4.4. These methods were required to estimate effectiveness even before the value of additional evidence was considered, so would be required in any event. Some questions related to practicality and implementation includes:

i. Who should be primarily responsible for conducting this type of analysis? Should it be funding agencies, such as PCORI, and their academic affiliates or stakeholders suggesting a topic and/or submitting specific research proposals?
ii. Should this type of analysis be required or recommended? Whether required or recommended should this be for all suggested topics and proposals? If not, how should efforts be focused, e.g., only required for specific proposals but not necessarily for a suggested topic?
iii. How might the capacity (if lacking) to conduct this type of analysis, and the type of meta-analysis that is likely to be required in many circumstances, be generated?
iv. Can access to information that might commonly be required be provided and specified by funding agencies, such as PCORI, so all analysis could be based on common sources of more easily accessed information? For example, to inform possible time horizons for the value of additional information from international trial registries (Phase II, III and IV), patents, and licensing applications are useful but are often propriety and partial.
v. Can similar simple methods to identify the potential value of understanding the reasons for variability in patient outcomes and the potential value of more individualized care be developed which can also be implemented routinely as part of systematic review?
vi. There are a number of areas were methods of evidence synthesis and the analysis of observational data are developing rapidly to more fully inform the type of assessments required in research prioritization. They include: combining individual patient level data with summary statistics from published trials, synthesizing observational data with trial evidence, reflecting potential bias, relevance and the likely difference between effectiveness and

efficacy, accounting for potential selection bias within observational studies and *into* clinical trials as well as revealing information about the joint distribution of effects in trials and observational data. The question that these developments pose is what process might funding agencies adopt to make best use of developing methods of analysis?