PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Which factors may determine the necessary and feasible type of effectiveness evidence?: A mixed methods approach to develop an instrument to help coverage decision makers
AUTHORS	de Groot, Saskia; Rijnsburger, Rian; Versteegh, Matthijs; Heymans, Juanita; Kleijnen, Sarah; Redekop, Ken; Verstijnen, Ilse

VERSION 1 - REVIEW

REVIEWER	C Craig Blackmore Director, Center for Health Services Research Virginia Mason Medical Center Seattle, WA USA and Chair, Washington State Health Technology Clinical Committee
REVIEW RETURNED	31-Dec-2014

GENERAL COMMENTS	The authors are to be congratulated for taking on the challenging issue of sufficient evidence for reimbursement decisions. The FIT toll has potential to help decision makers through this process. However, ultimately, to be useful, the tool has to improve the decision making process in some capacity. The authors interview potential users who express appreciation for the tool, but no evidence is presented that the tool improves the appropriateness, efficiency, consistency, or other aspect of the decision making process. I also fear that this tool will not address the fundamental question for policy makers, which is not “Is an RCT necessary?” but rather “How do we decide based on the existing, rather than desired strength of evidence?” This tool might be useful in the abstract sense in determining what evidence to attempt to acquire going forward, but doesn’t affect the reality of having to make decisions today. Seldom are funding decision makers satisfied with the strength of the evidence for or against, but still must make decisions based on the evidence that exists. I am also concerned that in some respects the tool sets the bar too low for required strength of evidence. The “intervention is common practice” and “good quality low level evidence” (from Table 3) are particularly troubling. There are many interventions which have diffused rapidly based on low level evidence, but it is only later that the RCTs are performed that demonstrate the lack of effectiveness. There are innumerable interventions that have been adopted with insufficient evidence, only to be abandoned when the appropriate RCT was done to evaluate their effectiveness. The FIT document would seem to imply that decision makers should accept lower quality evidence for such technologies rather than encouraging RCTs.

	In contrast, this approach could also encourage over-reliance on RCTs. There is no question that an RCT is the strongest experimental design for determining causality. However, because they occur in highly selected samples, with rigid protocols, and generally limited follow-up, RCTs may not provide the evidence necessary for coverage decisions. For example, RCT evidence for MoM hip prostheses, even if available, might not have shown the delayed effects of systemic metal resorption. Registries or observational studies with long follow-up often provide more accurate assessment of the safety of a procedure. A second challenge with many RCTs is that they are designed to answer the question of causality, not clinical effectiveness. Accordingly, RCTs are often performed on the subpopulation most likely to benefit under controlled circumstances most favorable to the intervention. The information from such a trial, though very strong under principles of evidence based medicine, may not be relevant to the policy decision makers. Large pragmatic trials can address this concern, but are not common. Finally, RCTs vary greatly in quality. The FIT captures some of this variability explicitly through discussion of double blinding, but does not also include quality factors in RCTs that may be equally or more important, including sample size, choice of endpoints, choice of comparators, clinical setting, and source of funding. A major challenge we have faced in coverage decisions is whether to extrapolate evidence from RCTs to different populations and or conditions than was covered under the original RCTs. This is explicit in table 3, but I think underemphasized by the tool. For example, the Washington State Health Technology Clinical Committee (a group in the United States which appears to be similar to the ZIN) provided coverage for MRI supplemental breast screening in women at high risk of cancer based on cohort data on cancer detection, with RCT mortality data from mammography. There was no direct RCT evidence of mortality benefit for MRI screening. On the other hand, the HTCC was willing to provide coverage for proton beam therapy for intraocular and pediatric CNS tumors (like the ZIN) based on case-series and treatment mechanism, but was not willing to extend coverage to prostate cancer local therapy without RCT evidence of effectiveness. Would this constitute “extension of the indication area of a procedure that is already…reimbursed?” There is judgment implicit in each of these determinations that cannot easily be summarized by a simple tool. In the non-coverage decision for proton beam therapy for prostate cancer, the HTCC decided not simply based on the absence of RCT evidence, but rather on a combination of the available evidence, the burden of disease to the individual, prevalence of disease in the population, potential complications, potential differential effectiveness in mortality and quality of life, and cost. None of these are binary responses. This level of detail is critical for decision making, but might be lost with a simple determination of RCT or no RCT. Specific comments: There are many additional questions that if covered in this paper would add to its value. How was the validation study performed? Was there a formal way of eliciting information from the decision makers in the validation study? How were suggestions for improvement captured and incorporated? How did the tool change the decision making process? Is there evidence that the decision making process was improved with the tool? Do you have any data on how the tool performs? What is the

	inter/intra-observer reliability? Do scores on the FIT correspond in any way with actual reimbursement decisions?

REVIEWER	Katharina Fischer Hamburg Center for Health Economics, University of Hamburg
REVIEW RETURNED	07-Jan-2015

GENERAL COMMENTS	The study develops a tool to assess the minimum level of evidence on effectiveness that is required for an intervention to make reimbursement decisions. The analysis was based on past decisions made in the Netherlands, literature review and expert consultations. This tool and thus, the study, is meaningful to guide discussions in committees dedicated to make reimbursement decisions at international level. Whilst I enjoyed reading the study, some changes are needed. Especially, clarifications with regards to the methods used and the modifications in the presentation of the results are required. Major comments Introduction Literature: The authors should refer to the existing literature that deals with the problem of lacking RCT evidence in coverage decisions. A collection of studies that have included variables that analyse the level of effectiveness evidence in coverage decisions is provided in the systematic review by Fischer (2012), Health Policy. Concepts of necessity and feasibility Across the article, the authors do not provide a definition of what they mean that a RCT is necessary and feasible. Does necessity mean that for different reasons other evidence exists that no additional RCT is needed. Feasibility suggests that there are ethical or other issues that discern the performance of an RCT. Whilst these two terms may seem self-explanatory, they should be made explicit as they are two key concepts in this study, also relating to other literature. Type of effectiveness In both marketing authorization and reimbursement decisions, evidence on effectiveness is used for assessing health technologies. Whilst in the marketing authorization stage, emphasis is put on efficacy related outcomes, reimbursement decisions focus on effectiveness in terms of mortality and morbidity oriented outcomes. It is unclear to which type of effectiveness evidence the authors relate to. Type of health care interventions Whilst it is described in the strengths and limitations of the study section, the focus on non-pharmaceutical medical specialist care does not become evident from the text. A rationale for this specification should be provided. Methods In the methods section, several methodological steps need elaboration as they seem rather arbitrary. Selection of reimbursement decisions: Please provide a rationale for the selected timeframe of Jan 2007 to Dec 2010. How has the data on decisions been accessed? Where there documents or minutes of each decision? Also, a rationale for the classification of complexity of the reimbursement reports is needed. It is unclear what complexity

	means in this context. Data form: If present, how were multiple patient groups for one intervention accounted for? What do the decisions of a positive and negative reimbursement imply? Does this relate to cost coverage also? Literature review: Has the literature been screened systematically? If yes, please provide more information on the number of hits, inclusion / exclusion criteria and why only one database was searched. Expert interviews: No information about participant consent has been provided. Use multiple methodological steps are applied. Authors should clarify the rationales for performing each step and the major results concluding in the final instrument. A figure displaying the research process could help here. Results The authors describe the reimbursement decisions included. A description of the sample of decisions is missing. The content provided in the two tables seems somewhat arbitrary. Authors should focus on presentation of those elements that contributed to specification of the FIT instrument. The results of the literature search and how they contributed to development of the FIT instrument need to be described besides citing these references in table 3. Especially, the references of the articles included should be added to the text. Discussion In the first paragraph, the authors describe that the assessment of necessity and feasibility follows a cascade of decisions. This two-step procedure is essential, but does not become evident from the methods / results section. For international application, authors should discuss the steps needed to transfer the instrument into other settings. In particular, what are evidence requirements stated by decision-making authorities? How do they relate to the FIT instrument? Relating to my comment on the introduction section, authors should discuss the peculiarities of effectiveness evaluation in contrast to marketing authorization if reimbursement is the stage they focus on. Minor comments Strengths and limitations of the study: Without having read the article, the definition of the FIT acronym is not clear in the bullet point listing. Table 3 and subheading on page 8: “When is randomization […] (un)necessary?” Please avoid double meaning using both positive and negative expressions. Please revise the citation style for this journal.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
The authors are to be congratulated for taking on the challenging issue of sufficient evidence for reimbursement decisions. The FIT tool has potential to help decision makers through this process. However, ultimately, to be useful, the tool has to improve the decision making process in some capacity. The authors interview potential users who express appreciation for the tool, but no evidence is presented that the tool improves the appropriateness, efficiency, consistency, or other aspect of the decision making process.

We agree with the reviewer that the current impact of the tool on the decision making process is not fully known. However, the aim of our study was the development of this tool. The instrument's face validity was discussed in a joint meeting with ZIN decision makers and the project team, and conclusions were made whether the factors included appeared to be relevant, reasonable, unambiguous and clear. Unfortunately, we were not able to study the instrument's reliability and additional forms of validity, since only three new interventions were assessed in the validation phase. Therefore, we do suggest further testing of this instrument in future reimbursement decisions, and we have added the following sentences to the discussion section:

"Not only does the instrument's completeness requires further testing, but also its impact on the decision making process in terms of efficiency, reliability (such as inter/intra-observer reliability) and additional forms of validity. The current validation phase was too short to properly address these issues. However, regardless of its impact on decisions, the instrument provides a degree of transparency in defining the appropriate evidence for an intervention, and this is in itself an important improvement over the status quo."

The manuscript ends with a statement about the potential of the instrument: "Although the instrument needs further refinement, and although a critical appraisal of the factors influencing the necessity and feasibility of blinded RCTs and its constituent study characteristics remains essential, this instrument has the potential to support transparent, reproducible and well-founded decisions on appropriate evidence of effectiveness in medical specialist care".

I also fear that this tool will not address the fundamental question for policy makers, which is not "Is an RCT necessary?" but rather "How do we decide based on the existing, rather than desired strength of evidence?" This tool might be useful in the abstract sense in determining what evidence to attempt to acquire going forward, but doesn't affect the reality of having to make decisions today. Seldom are funding decision makers satisfied with the strength of the evidence for or against, but still must make decisions based on the evidence that exists.

We totally agree with the reviewer that the fundamental question is how to decide on reimbursement based on existing evidence. The FIT instrument addresses a vital question that needs to be answered beforehand, which study characteristics of a blinded RCT are necessary and feasible. We have added the following section to the discussion to address this important comment.

"In the context of decision making, the fundamental question is "what to decide" given the particular context of an intervention including the existing evidence. The FIT instrument does not address this question, but addresses a vital question that needs to be answered beforehand, i.e. the FIT instrument examines which study characteristics of a blinded RCT are necessary, and then, whether these characteristics are feasible. Thereafter, the appropriate evidence of effectiveness can be compared with the actual available evidence. If the available evidence lacks one or more of the study characteristics that were deemed necessary, the decision maker may advise conditional reimbursement, the condition being that further evidence has to be assembled after reimbursement. However, the decision maker may decide not to advise conditional reimbursement if the explanations for the existing lack of necessary study characteristics provided by the FIT instrument are considered legitimate. In this latter situation, the decision maker will assume that no better-fitting evidence will probably appear in the future, and that a decision based on the available evidence will have to be made. Next to reasons for a suboptimal evidence base, additional factors are considered important in reimbursement decisions, such as disease severity and budget impact. These latter criteria concern equity in a given society as well as overall capacity for reimbursement, and fell beyond the scope of our research."

I am also concerned that in some respects the tool sets the bar too low for required strength of evidence. The "intervention is common practice" and "good quality low level evidence" (from Table 3) are particularly troubling. There are many interventions which have diffused rapidly based on low level evidence, but it is only later that the RCTs are performed that demonstrate the lack of effectiveness. There are innumerable interventions that have been adopted with insufficient evidence,

only to be abandoned when the appropriate RCT was done to evaluate their effectiveness. The FIT document would seem to imply that decision makers should accept lower quality evidence for such technologies rather than encouraging RCTs.

The factors "intervention is common practice" and "availability of good quality low level evidence of effectiveness" were cited in various reimbursement reports and therefore included in Table 3. However, as may have become clearer now from the above altered text; the identification of possible reasons for suboptimal evidence, do not automatically mean that these reasons are accepted as legitimate grounds.

Furthermore, the factors included in the FIT instrument were categorised in three groups: 1) those rendering randomisation, a control group or blinding (un)necessary, 2) those that render randomisation, a control group or blinding hard to achieve and 3) those that hinder the feasibility of randomisation, a control group or blinding. Group two and three include factors that influence the feasibility of the constituent study characteristics of blinded RCTs. The difference between group two and three is that the latter group includes factors that may by themselves be insufficiently strong, but may jointly argue for deviating from randomisation, a control group or blinding. The factors "intervention is common practice" and "availability of good quality low level evidence of effectiveness" are categorised in the latter group.

In contrast, this approach could also encourage over-reliance on RCTs. There is no question that an RCT is the strongest experimental design for determining causality. However, because they occur in highly selected samples, with rigid protocols, and generally limited follow-up, RCTs may not provide the evidence necessary for coverage decisions. For example, RCT evidence for MoM hip prostheses, even if available, might not have shown the delayed effects of systemic metal resorption. Registries or observational studies with long follow-up often provide more accurate assessment of the safety of a procedure. A second challenge with many RCTs is that they are designed to answer the question of causality, not clinical effectiveness. Accordingly, RCTs are often performed on the subpopulation most likely to benefit under controlled circumstances most favorable to the intervention. The information from such a trial, though very strong under principles of evidence based medicine, may not be relevant to the policy decision makers. Large pragmatic trials can address this concern, but are not common. Finally, RCTs vary greatly in quality. The FIT captures some of this variability explicitly through discussion of double blinding, but does not also include quality factors in RCTs that may be equally or more important, including sample size, choice of endpoints, choice of comparators, clinical setting, and source of funding.

The reviewer is right in claiming that an RCT has its disadvantages, such as a limited generalizability and follow-up. Before applying the FIT instrument, the PICO (i.e. population – intervention – comparator – outcome) should be specified. The FIT instrument should be used for all outcomes of interest. We have added the following sentences to the discussion section:

"(…) an RCT has its disadvantages, such as a limited generalizability and limited follow-up. Whereas efficacy might be more important for market authorization to address causality, effectiveness might be more important from a decision making perspective which is the focus of this study. The optimal study design depends on the outcome of interest. If the outcome of interest appears to be a safety outcome, a blinded RCT may no longer the optimal study design; the FIT instrument will show that randomisation is hard to achieve since 'outcomes occur in distant future'. Therefore, the FIT instrument should be used for all outcomes of interest."

Furthermore the reviewer is right in claiming that RCTs differ in quality. Therefore, we recommend using the GRADE system to assess the available evidence. As stated in the discussion section, the GRADE system and our FIT instrument complement each other; "While the GRADE approach focuses on assessing the quality of available evidence, we identified factors that outline which types of evidence can in principle be available, thereby providing arguments for reimbursement decisions when the available evidence does not match the evidence that is considered necessary and feasible."

A major challenge we have faced in coverage decisions is whether to extrapolate evidence from RCTs to different populations and or conditions than was covered under the original RCTs. This is

explicit in table 3, but I think underemphasized by the tool. For example, the Washington State Health Technology Clinical Committee (a group in the United States which appears to be similar to the ZIN) provided coverage for MRI supplemental breast screening in women at high risk of cancer based on cohort data on cancer detection, with RCT mortality data from mammography. There was no direct RCT evidence of mortality benefit for MRI screening. On the other hand, the HTCC was willing to provide coverage for proton beam therapy for intraocular and pediatric CNS tumors (like the ZIN) based on case-series and treatment mechanism, but was not willing to extend coverage to prostate cancer local therapy without RCT evidence of effectiveness. Would this constitute "extension of the indication area of a procedure that is already…reimbursed?" There is judgment implicit in each of these determinations that cannot easily be summarized by a simple tool. In the non-coverage decision for proton beam therapy for prostate cancer, the HTCC decided not simply based on the absence of RCT evidence, but rather on a combination of the available evidence, the burden of disease to the individual, prevalence of disease in the population, potential complications, potential differential effectiveness in mortality and quality of life, and cost. None of these are binary responses. This level of detail is critical for decision making, but might be lost with a simple determination of RCT or no RCT.

We agree with the reviewer that all factors mentioned above have to have an impact on the reimbursement decision. The extension argument (if it is legitimate in this case) just adds another argument to the decision about which study characteristics of a blinded RCT are necessary and feasible. The FIT instrument, by addressing the extension question, ensures that the argument is considered.

Still, we agree that every decision on the necessity and feasibility of blinded RCTs and its constituent study characteristics needs a critical appraisal; with this tool we have at least tried to make this judgement transparent, reproducible and well-founded. Therefore, we have added the following sentence to the end of the discussion section:

"Although the instrument needs further refinement, and although a critical appraisal of the factors influencing the necessity and feasibility of blinded RCTs and its constituent study characteristics remains essential, this instrument has the potential to support transparent, reproducible and well-founded decisions on appropriate evidence of effectiveness in medical specialist care."

Specific comments:
There are many additional questions that if covered in this paper would add to its value. How was the validation study performed? Was there a formal way of eliciting information from the decision makers in the validation study?
Information from the decision makers was elicited in a joint meeting. Instructions about the validation were sent to the decision makers beforehand. These instructions explained the aim of the validation, and pointed out particular questions to the decision makers, such as 'Do all questions apply?', 'Do you miss any questions?' and 'Are the questions clearly stated?' These sentences were added to the methods section.

How were suggestions for improvement captured and incorporated?
Suggestions for improvement of the instrument were applied if everyone (both the decision makers and the project team) agreed. A sentence indicating this was added to the methods section.

How did the tool change the decision making process? Is there evidence that the decision making process was improved with the tool? Do you have any data on how the tool performs?
Unfortunately, we do not have any information about how the tool performs (yet). We recommend further testing of the instrument. This recommendation was added to the discussion section.

"*Not only does the instrument's completeness requires further testing, but also its impact on the decision making process in terms of efficiency, reliability (such as inter/intra-observer reliability) and additional forms of validity. The current validation phase was too short to properly address these issues. However, regardless of its impact on decisions, the instrument*

provides a degree of transparency in defining the appropriate evidence for an intervention, and this is in itself an important improvement over the status quo."

What is the inter/intra-observer reliability?
We have also added the recommendation to test the instrument's reliability (such as inter/intra-observer reliability) in future reimbursement decisions. (see previous paragraph)

Do scores on the FIT correspond in any way with actual reimbursement decisions?
Whether the results of the FIT instrument correspond with actual reimbursement decisions is not (yet) known. As stated before, the aim of our study was the development of this instrument and only the instruments' face validity was tested. Further testing is recommended (see previous paragraphs).

Reviewer: 2

The study develops a tool to assess the minimum level of evidence on effectiveness that is required for an intervention to make reimbursement decisions. The analysis was based on past decisions made in the Netherlands, literature review and expert consultations. This tool and thus, the study, is meaningful to guide discussions in committees dedicated to make reimbursement decisions at international level. Whilst I enjoyed reading the study, some changes are needed. Especially, clarifications with regards to the methods used and the modifications in the presentation of the results are required.

Major comments

Introduction
Literature: The authors should refer to the existing literature that deals with the problem of lacking RCT evidence in coverage decisions. A collection of studies that have included variables that analyse the level of effectiveness evidence in coverage decisions is provided in the systematic review by Fischer (2012), Health Policy.

Thank you for this suggestion. In the introduction we now refer to existing literature, and included a reference to Fischer (2012):

"(…) A systematic review by Fischer (2012) showed that, indeed, the presence of suboptimal evidence plays an important role in coverage decision making."

Concepts of necessity and feasibility
Across the article, the authors do not provide a definition of what they mean that a RCT is necessary and feasible. Does necessity mean that for different reasons other evidence exists that no additional RCT is needed. Feasibility suggests that there are ethical or other issues that discern the performance of an RCT. Whilst these two terms may seem self-explanatory, they should be made explicit as they are two key concepts in this study, also relating to other literature.

We agree with the reviewer, and therefore we made these two concepts explicit in the methods section:

"The factors included were categorised in three groups: 1) those rendering randomisation, a control group or blinding (un)necessary, 2) those that render randomisation, a control group or blinding hard to achieve and 3) those that hinder the feasibility of randomisation, a control group or blinding. The first group includes factors that highlight whether the constituent study characteristics of blinded RCTs are necessary to determine an unbiased effect of an intervention. Groups two and three include factors that influence the feasibility of the constituent study characteristics of blinded RCTs, in other words, are these characteristics possible given the study population, the intervention, its setting and the desired outcomes? The difference between groups two and three is that the latter group includes factors that may by themselves be insufficiently strong, but may jointly argue for deviating from randomisation, a control group or blinding."

Type of effectiveness

In both marketing authorization and reimbursement decisions, evidence on effectiveness is used for assessing health technologies. Whilst in the marketing authorization stage, emphasis is put on efficacy related outcomes, reimbursement decisions focus on effectiveness in terms of mortality and morbidity oriented outcomes. It is unclear to which type of effectiveness evidence the authors relate to.

We agree with the reviewer that this remains unclear from the manuscript. In fact, the FIT tool can be used to answer questions regarding efficacy and effectiveness. However, this publication focuses on reimbursement decisions where effectiveness is considered most important, including safety. We have added the following sentences to the discussion section:

"(…) Whereas efficacy might be more important for market authorization to address causality, effectiveness might be more important from a decision making perspective which is the focus of this study. The optimal study design depends on the outcome of interest. If the outcome of interest appears to be a safety outcome, a blinded RCT is no longer the optimal study design; the FIT instrument will show that randomisation is hard to achieve since 'outcomes occur in distant future'. Therefore, the FIT instrument should be used for all outcomes of interest."

Type of health care interventions
Whilst it is described in the strengths and limitations of the study section, the focus on non-pharmaceutical medical specialist care does not become evident from the text. A rationale for this specification should be provided.

The reviewer is right, a rationale for the focus on non-pharmaceutical medical specialist care is currently lacking. Non-pharmaceutical medical specialist care was chosen as a starting point. We have added the following sentences to the introduction of our manuscript:

"We focused on evidence of effectiveness for non-pharmaceutical, therapeutic medical specialist care as a starting point. Although the instrument is based on non-pharmaceutical therapeutic medical specialist care, it may be applicable to other types of interventions."

Methods
In the methods section, several methodological steps need elaboration as they seem rather arbitrary. Selection of reimbursement decisions: Please provide a rationale for the selected timeframe of Jan 2007 to Dec 2010. How has the data on decisions been accessed? Where there documents or minutes of each decision?

We understand the reviewer's comments, and elaborated the methods section:

"Reports were available on all reimbursement decisions. January 1st 2007 was chosen as a starting point, since from this date onwards ZIN officially applied the principles of evidence-based medicine to determine whether care is effective. As data extraction was performed in 2011, 2010 was the last year for which complete reports were available."

Also, a rationale for the classification of complexity of the reimbursement reports is needed. It is unclear what complexity means in this context.

The reviewer is right that a rationale is missing. Therefore, the following sentences are added to the methods section:

"A stratified sample was implemented in order to ensure that a variety in reimbursement reports was achieved rather than selecting reports where the necessity and feasibility of blinded RCTs was evident. The reimbursement reports were therefore classified into three groups based on their level of complexity regarding evidence of effectiveness, i.e. simple (few necessity or feasibility issues), intermediate (moderate necessity or feasibility issues) or complex (complex necessity or feasibility issues), by the three researchers from ZIN (J.H., S.K., I.V.)."

Data form: If present, how were multiple patient groups for one intervention accounted for? What do the decisions of a positive and negative reimbursement imply? Does this relate to cost coverage also?

A data form was available. The PICO (i.e. population – intervention – comparator – outcome) was documented on this form. So if reports dealt with multiple patient groups, this was reported here (besides arguments that advocated for (or against) the acceptance of the available evidence for any of the patient groups).

The decision of a positive or negative reimbursement decision implies that the intervention was considered or not considered suitable for reimbursement by adding it to the basic health insurance package in The Netherlands. We have added the following sentences to the discussion section:

"A negative or positive reimbursement decision implies whether the intervention was considered suitable for reimbursement by adding it to the basic health insurance package in The Netherlands. These decisions did not always rely on the assessment of effectiveness only, but may also take additional criteria into account, such as cost-effectiveness."

Literature review: Has the literature been screened systematically? If yes, please provide more information on the number of hits, inclusion / exclusion criteria and why only one database was searched.

Yes, the literature has been screened systematically using the following search strategy:

Database: Ovid MEDLINE(R) without Revisions <1996 to March Week 3 2011>
Search Strategy:
---
1 ((level? or degree? or criteri$ or hierarch$ or require$ or assess$ or standard$) adj4 (evidence or evidentiary)).ti,ab. (20004)
2 exp *Evidence-Based Practice/ec, lj, mt, og, st, sn, td [Economics, Legislation & Jurisprudence, Methods, Organization & Administration, Standards, Statistics & Numerical Data, Trends] (4386)
3 1 and 2 (401)
4 decision support techniques.mp. or Decision Support Techniques/ (6941)
5 decision making.mp. or Decision Making/ (65910)
6 decision making organizational.mp. or Decision Making, Organizational/ (7149)
7 policy making.mp. or Policy Making/ (8390)
8 health policy.mp. or exp Health Policy/ (49171)
9 health planning.mp. or exp Health Planning/ (154168)
10 insurance, health.mp. or exp Insurance, Health/ (59251)
11 delivery of health care.mp. or exp "Delivery of Health Care"/ (441478)
12 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 (580015)
13 3 and 12 (153)
14 exp Evidence-Based Practice/ec, lj, st [Economics, Legislation & Jurisprudence, Standards] (2370)
15 1 and 14 (251)
16 15 and 12 (98)

The search strategy identified 98 publications (this was also mentioned in the results section). Although this literature review was conducted in Medline only, this database has indexed over 5,000 journals and we therefore decided not to search additional databases. We agree with the reviewer that further details on the literature review are missing, therefore we have added the following sentences to the methods:

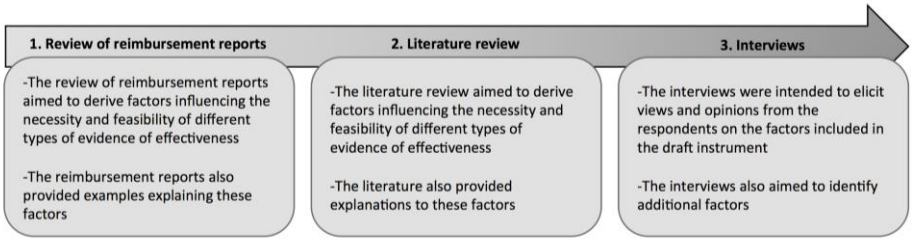
"The resulting publications were screened systematically. Selection on relevance was carried out by two researchers (S.G., R.R.). First, titles and abstracts were screened; publications that might discuss any factors influencing the necessity and feasibility of different types of evidence of effectiveness were included. The full texts of the selected publications were then examined; publications that did not discuss any factors influencing the necessity and feasibility of different types of evidence of effectiveness were excluded, just as publications not in English and publications not available as full text. The reference lists of relevant publications were screened by hand to identify any additional publications for inclusion."

Expert interviews: No information about participant consent has been provided.

"All participants were informed about the purpose and content of the interview and agreed to participate by e-mail." This sentence was added to the methods section.

Use multiple methodological steps are applied. Authors should clarify the rationales for performing each step and the major results concluding in the final instrument. A figure displaying the research process could help here.

As recommended by the reviewer we have added a figure to the methods section showing the research process.



Results
The authors describe the reimbursement decisions included. A description of the sample of decisions is missing.

The first paragraph of the results section and the text beneath Table 1 includes a description of the sample of decisions.

"Eleven of 22 reimbursement decisions were negative, i.e. the intervention was not considered suitable for reimbursement by adding it to the basic health insurance package in The Netherlands, and the remaining 11 decisions positive. In five of 11 positive reimbursement decisions, evidence from (blinded) RCTs was lacking."

"Of the 11 negative reimbursement decisions, evidence from (blinded) RCTs was lacking in three of them."

The content provided in the two tables seems somewhat arbitrary. Authors should focus on presentation of those elements that contributed to specification of the FIT instrument.

The arguments listed in the tables are the only arguments explicitly mentioned in the reimbursement reports. Hence, this is an exhaustive list rather than a selection of arguments selected by us. Therefore, we have added the following sentence to the results section:

"The arguments listed in the tables are the only arguments explicitly mentioned in the reimbursement reports."

Finally, we have improved the design of the tables.

The results of the literature search and how they contributed to development of the FIT instrument need to be described besides citing these references in table 3. Especially, the references of the articles included should be added to the text.

With adding to the manuscript the figure shown above, we have tried to be clearer about how the results of the literature review contributed to the development of the FIT instrument. Factors influencing the necessity and feasibility of randomization, a control group and blinding derived from the literature are included in the final FIT instrument (including reference). Furthermore, explanations to these factors derived from the literature are included in our manuscript (including reference).

Discussion
In the first paragraph, the authors describe that the assessment of necessity and feasibility follows a cascade of decisions. This two-step procedure is essential, but does not become evident from the methods / results section.

We agree with the reviewer that this was not clearly described in the methods section. Based on the second comment from this reviewer we have added text in the methods section which describes this cascade of decisions.

For international application, authors should discuss the steps needed to transfer the instrument into other settings. In particular, what are evidence requirements stated by decision-making authorities? How do they relate to the FIT instrument?

As stated in the discussion section, although reimbursement decisions vary across countries, the FIT instrument might be useful for reimbursement agencies in other countries since most items are based on general epidemiological principles. While we agree with the high value of the additional information as recommended by the reviewer, we feel that an adequate description of the transferability of the instrument requires additional research to the evidence requirements in other countries. We feel, therefore, that this falls outside of the scope of our research, but are eager to address this in follow-up research.

Relating to my comment on the introduction section, authors should discuss the peculiarities of effectiveness evaluation in contrast to marketing authorization if reimbursement is the stage they focus on.

We have added a full paragraph on this topic in the discussion section:

"(…) an RCT has its disadvantages, such as a limited generalizability and limited follow-up. Whereas efficacy might be more important for market authorization to address causality, effectiveness might be more important from a decision making perspective which is the focus of this study. The optimal study design depends on the outcome of interest. If the outcome of interest appears to be a safety outcome, a blinded RCT is no longer the optimal study design; the FIT instrument will show that randomisation is hard to achieve since 'outcomes occur in distant future'. Therefore, the FIT instrument should be used for all outcomes of interest"

Minor comments
Strengths and limitations of the study: Without having read the article, the definition of the FIT acronym is not clear in the bullet point listing.
We agree with the reviewer that this is unclear and adapted the bullet point listing:

Strengths and limitations of this study

- ***In this study multiple sources were used, including 20 reimbursement reports made by the Dutch National Health Care Institute, literature and expert opinion.***
- ***Since most items that are used to examine which study characteristics of a blinded RCT are necessary and feasible, are based on general epidemiological principles, results of this study might be useful for reimbursement agencies in other countries besides The Netherlands.***
- ***Not all possible study characteristics are taken into account; the study was limited to the necessity and feasibility of randomisation, a control group and blinding.***
- ***The instrument's completeness requires further testing, just as its impact on the decision making process in terms of efficiency, reliability (such as inter/intra-observer reliability) and additional forms of validity.***

Table 3 and subheading on page 8: "When is randomization […] (un)necessary?" Please avoid double meaning using both positive and negative expressions.
We agree with the reviewer that it is better to avoid double meaning. Therefore we have changed this expression:

"When is randomisation, a control group or blinding necessary, or on the contrary unnecessary?" We have changed this subheading as well in Table 3.

Please revise the citation style for this journal.
We have changed the citation style.

VERSION 2 – REVIEW

REVIEWER	Katharina Fischer Hamburg Center for Health Economics, University of Hamburg, Germany
REVIEW RETURNED	10-Mar-2015

GENERAL COMMENTS	The authors have incorporated the majority of the comments such that the manuscript has improved in clarity. Especially, the methods are described appropriately. Some comments remain. Specific comments Methods, definition of the FIT instrument While I appreciate the authors' approach to define necessity and feasibility in the context of this study, the definitions provided need further refinement. The way the definitions are presented is somewhat confusing. I suggest restructuring the paragraph by providing the definitions after specification of the three categories. Moreover, how do the three categories relate to the literature identified in the systematic review? Also again, please avoid double meaning using the phrase "(un)necessary". Sampling of decisions The description how decisions were selected to identify the issues of necessity and feasibility raises further issues in the methodology applied. Have the authors selected the decisions based on the rating of complexity regarding necessity and feasibility? If not (I assume so), what were the selection criteria for decisions to be included (e.g. randomization)? What was the need for decisions to be rated by complexity? Please specify or consider deleting this point. Table 3: When is randomization, a control group or blinding necessary, or on the contrary unnecessary? The latter clause is not needed to completely avoid double-meaning. I recommend a general language revision to improve the clarity of the arguments stated. Many passages and sentences appear somewhat lengthy and difficult to understand.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

The authors have incorporated the majority of the comments such that the manuscript has improved in clarity. Especially, the methods are described appropriately. Some comments remain.

Specific comments

Methods, definition of the FIT instrument. While I appreciate the authors' approach to define necessity and feasibility in the context of this study, the definitions provided need further refinement. The way the definitions are presented is somewhat confusing. I suggest restructuring the paragraph by providing the definitions after specification of the three categories. Moreover, how do the three categories relate to the literature identified in the systematic review? Also again, please avoid double meaning using the phrase "(un)necessary".

We agree with the reviewer that the way the definitions of necessity and feasibility are presented now is somewhat confusing, and that the definitions need further refinement. Also, it was unclear that the three categories do not relate to the literature identified in the systematic review, but were created through agreement within the project team. Therefore, as recommended, we have restructured this paragraph, refined the definitions and added subheadings to the methods section. Furthermore, double meaning has now been avoided:

"The resulting instrument was called the FIT instrument (Feasible Information Trajectory). The factors included were categorised in two groups: The first group deals with the necessity of randomisation, a control group and blinding, relating to situations in which one or more of these three characteristic are not required. The second group deals with the feasibility of randomisation, a control group and blinding, and was subdivided into two groups (2A and 2B). Group 2A refers to factors that, stand alone, are sufficiently strong to deviate from randomisation, a control group and/or blinding. Group 2B refers to factors that in itself are insufficiently strong, but may jointly provide a compelling case to do so."

Sampling of decisions
The description how decisions were selected to identify the issues of necessity and feasibility raises further issues in the methodology applied. Have the authors selected the decisions based on the rating of complexity regarding necessity and feasibility? If not (I assume so), what were the selection criteria for decisions to be included (e.g. randomization)? What was the need for decisions to be rated by complexity? Please specify or consider deleting this point.

The reimbursement reports were, indeed, selected based on their level of complexity regarding the necessity and feasibility of randomisation, a control group and blinding. In this way, ending up with a sample just containing reports where the necessity and feasibility of blinded RCTs was evident, was prevented.

The reviewer is right that the selection procedure did not become clear from the manuscript. We have now specified this more clearly in the methods section:

"A stratified sample was implemented in order to prevent ending up with a sample just including reimbursement reports for which the necessity and feasibility of blinded RCTs was evident. The reimbursement reports were therefore classified into three groups based on their level of complexity, i.e. simple (few necessity or feasibility issues), intermediate (moderate necessity or feasibility issues) or complex (complex necessity or feasibility issues), by the three researchers from ZIN (J.H., S.K., I.V.). (…)"

Table 3: When is randomization, a control group or blinding necessary, or on the contrary unnecessary?
The latter clause is not needed to completely avoid double-meaning.
We agree with the reviewer, and removed the latter clause.

I recommend a general language revision to improve the clarity of the arguments stated. Many passages and sentences appear somewhat lengthy and difficult to understand.
As recommended by the reviewer, we revised some of the language in our manuscript and shortened some passages and sentences. Furthermore, we have added subheadings to the methods section in order to improve the clarity of our manuscript.