PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	A protocol for data extraction: how real-world data have been used in the National Institute for Health and Care Excellence appraisals of cancer therapy
AUTHORS	Kang, Jiyeon; Cairns, John

VERSION 1 – REVIEW

REVIEWER	Seamus Kent University of Oxford, Nuffield Department of Population Health
REVIEW RETURNED	12-Aug-2021

GENERAL COMMENTS	The planned activities described in this protocol will be of substantial interest to NICE and the wider research community. I very much look forward to seeing the results. I have provided a few minor comments below for consideration by the authors. 1. Definition of RWE a. The authors emphasise the importance of the frequency of data collection ("routinely collected") as being integral to some definitions of RWE (e.g. that of the FDA). I think a more common interpretation of "routinely collected" is "collected in routine care" rather than frequency. This different interpretation shouldn't have major effects on the data extraction but is perhaps worth considering. 2. Scope of studies covered a. While RWD generates substantial attention, often the interest lies in observational data more generally or, in the context of comparative effectiveness, the use of non-randomised studies, including those with interventional designs (non-randomised controlled studies or single arm trials with external control). 3. Providing context around the cancer Drugs Fund a. I think the general information extracted should clarify whether the submission was pre or post CDF (and +/- 2016 with reformed CDF & mandated data collection). b. The protocol could also explicitly state that the pre-and post-CDF evaluations are considered separately (where relevant). 4. Extraction of data about use of RWD in modelling a. Is data extraction based on the preferred model from the ERG or the company submission? b. Are uses of RWD for sensitivity analyses excluded? 5. Data analysis on factors associated with greater use of RWE a. Should the data analysis be restricted to those issues where RCTs would be the preferred source of evidence only (i.e., treatment effects)? For many other parameters (e.g., resource use) or event

	rates, RWD, assuming of sufficient relevance and quality, might be the preferred source of data and therefore less dependent on the limitations of the trial(s). 6. Questions of interest not captured by the review a. What are the characteristics of RWE submissions that make committees more willing to accept their use? b. What is the quality of RWE submissions across use cases? c. Where was the use of RWE proposed (either by companies, ERG, or committees) but rejected or otherwise not possible?

REVIEWER	Ash Bullement The University of Sheffield, ScHARR
REVIEW RETURNED	25-Aug-2021

GENERAL COMMENTS	Comments to the authors The authors have developed a protocol to aid with searching NICE assessments of cancer treatments for how RWD have been used to inform decision making. By following this protocol, the authors intend to extract information in a "reproducible, systematic and transparent" way. I believe substantial remedial work is required for the protocol to be deemed suitable for publication, based on clarity of reporting and justification of decisions made. Most of my comments are related to describing the planned approaches as opposed to major concerns with the planned analysis itself. However, there are also some concerns highlighted concerning potential for information loss, handling of missing data, and risk of inconsistent extraction/ reporting of information. General In general, some paragraphs are very large – suggest splitting some up where suitable. In addition, the paper would benefit from an English language review for flow. There are also a few instances of 'Error! Reference source not found' that need addressing. Introduction "Health Technology Assessment (HTA) requires valid and reliable information for the systematic evaluation of health technology". For this audience, can an explanation of HTA be included here? Health technology is also not an immediately understood term, especially when used to define HTA – could this be expanded upon? "Furthermore, the traditional design of RCTs is possibly less appropriate for new technologies such as those targeting rare genetic mutations or where there may be ethical issues with control arms". Here, suggest clarifying you are talking about the principle of equipoise, and why single-arm or uncontrolled trials may be undertaken. Ethical issues could be misconstrued as other types of problems which do not seem relevant here. "Moreover, RCTs tend to include strictly controlled populations". This is included at the end of an otherwise unrelated paragraph – suggest moving this point earlier in the text when describing other general issues with RCTs. "As a leading HTA agency, NICE has…". It would be my personal view that the start of this sentence should be removed as I think it's

	an unsubstantiated claim (even though in my opinion it's true!) “The evidence is structurally well-documented enough to find the key information and available on the NICE website.” Here, I think it would be clearer to say “… and is available on the NICE website” as this is a separate comment. “Although this study follows a more systematic approach to review the use of RWD, it does not fully explain how the data were extracted and what criteria were used to judge the use of data”. I agree with the authors that these aspects highlighted by the authors were not described fully in the study by Bullement et al. However, I think the authors could be clearer here about what is meant by "how the data were extracted" and "what criteria were used to judge the use of data". For the extraction, perhaps the authors could say: "a data extraction table was not provided" and for the criteria, perhaps it would be better to say "the authors focused only on how RWE influenced the cost-effectiveness analysis, and not RWE used to support the interpretation and/or perception of the results"? “As the process of reviewing appraisals is not clear enough, it is unclear whether the information presented provides a full picture of the use of RWD.” This sentence is difficult to understand because of the use of ‘not clear’ to describe why something else is ‘unclear’. Perhaps this would be easier to understand if the authors instead said: “Due to limited information presented concerning the review process in this study, it is unclear whether…”? The final paragraph in the Introduction is a little confusing – it is stated that the main purpose of the protocol is to extract data, but later it is noted that data will be analysed. Overall, I think this section could be better described, including specifically highlighting what the research questions are (as it is unclear exactly what these are) and signposting to the planned regression analysis which is only described very later in the paper. Also, I think it would be better to describe an analysis of data potentially providing a biased view of how RWE has been used, rather than the data themselves being 'biased'. Methods and analysis Figure 1: In my opinion, this diagram is unnecessary, and to an extent unhelpful - it raises extra questions (e.g., how are the data going to be validated? what is the analysis planned? etc.) without answering them at this point in the paper. I would suggest as a minimum signposting in the diagram to where these are discussed within the paper, and if not discussed consider removing this diagram altogether. “The information is extracted from identified appraisals in accordance with extraction rules”. What are the extraction rules? These should be explained here. “The extraction tool includes general appraisal information and appraisal-specific information such as characteristics of the main clinical evidence and the economic evaluation model”. An example of appraisal-specific information is noted, but what do the authors mean about general appraisal information, and how is this different to appraisal-specific information? From my understanding, the authors might be referring to things like TA number, data of

	publication, etc., but it is unclear to me why this is not considered information specific to the appraisal? A whole page of text to describe the definition of RWE seems excessive to me - I think this could be condensed substantially. Could the authors simplify this into a shorter paragraph noting that there is no standard definition, citing a range of alternative studies which give their own definition, and then provide their own definition? The reason(s) for the use of two different definitions of RWD are unclear to me when this is described within the methods and analysis section. Can this be succinctly summarized by the authors and added into the paper? “Relevant appraisal documents including the final scope, the manufacturer’s submission, the evidence review group (ERG) report, and the final appraisal determination are available for each appraisal. The appraisal documents are reviewed to establish whether RWD is used to determine any components of the economic evaluation”. Here, the authors state that appraisal documentation includes these, but I assume this is not an exhaustive list (e.g., clarification responses were also considered where applicable?). If so, I would omit the examples, and limit this to say “appraisal documentation”, as the wording otherwise may imply some documents were missed. “This research exclusively includes single-technology appraisals (STA) of oncology medicines.”. It would be helpful to clarify what the difference between STAs and MTAs is (acknowledging that the audience may not be aware of the differences in NICE TA processes). I think it’s perfectly reasonable to restrict to STAs, but clarification would be helpful here. Some comments on Figure 4/ Supplement 1 are presented in the list below. Here, these all highlight areas where a different researcher may record information in a different way based on interpretation of the protocol: • I’m unclear why outcome variables need to be separated by parametric versus non-parametric and disagree with the use of these terms here – it may be that a non-parametric estimate of OS informs the model, but this would be categorized here as ‘parametric’. Instead, would it be clearer for the authors to use ‘model input’ versus ‘not a model input’? • Recommendation: Is it not possible that there could be another recommendation (e.g., recommended in research?). Might not apply to any of the STAs included in your review, but I would check for possible recommendation types (including, for example, terminated appraisal). Suggest looking at the Excel file available via this link: https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-technology-appraisal-guidance/data/cancer-appraisal-recommendations • Incidence is often reported as a range – is this planned to be reflected? Guidance perhaps could be added for data such as incidence which might be reported as a range • Does the extraction account for studies with more than 2 arms?

	How should this be recorded? • Does number of participants matter by arm, or are you just collecting for the full population? Might be a situation where the comparator is essentially discarded if it's irrelevant for decision making • How is the risk of bias planned to be ranked/graded? This isn't clear from the extraction template – numbers are provided, but unsure where they are from/ how a reviewer would decide which rank to assign • I'm not sure grading maturity on % OS events is a universally appropriate mechanism to describe maturity. There may be some appraisals which simply don't report much in terms of OS, because they're for earlier stage disease. However, I'm not sure I have a better suggestion – is there some way of accounting for studies that are understandably 'immature' versus those that are actually just reflecting short follow up? Maybe there's a way of capturing median/minimum follow up, or expressing events as a proportion of the planned events per the primary end point of the study? "In case of the variables coarsely divided, the outcome of the extraction is so blunt that it cannot fully capture how RWD is used. Likewise, the variables overscrupulously divided are less likely to provide valid outcome to show the pattern of the use of RWD in the analysis". The phrasing of this statement is concerning, as it suggests that data will be coded such that it maximizes the chance of showing the results the authors are hoping to see. I would suggest the authors revisit this statement to explain that coding data effectively has advantages which include avoiding information loss, and also grouping 'similar' information across appraisals to establish patterns of RWE use. "Under the parametric use, the clinical effectiveness, health utility, cost side were thoroughly reviewed." Here and elsewhere, I suggest the authors refrain from using the phrase 'cost side' as this is not a standard term. Instead, the authors may wish to instead describe features of the submitted economic analysis, including inputs related to clinical effectiveness, health utility, and costs. "The extracted data will be analysed quantitatively in two different ways. First, a descriptive analysis will summarise where and how RWD has been used in appraisals. This will be supplemented by an analysis of the intensity of use of RWD in order to explore changes in the pattern of use of RWD over time and differences with respect to cancer type. Secondly, a regression analysis will be performed to investigate which factors are associated with the greater use of RWD in a company's submission". From my perspective, this instead sounds as though a qualitative summary of the appraisals will be undertaken, and then a quantitative regression analysis will be performed. I suggest the authors edit this description to clarify the analysis methods proposed. "A literature review and a pilot study were conducted to identify factors potentially associated with the use of RWD". Where are the review and pilot study findings – could these be referenced or included as supplementary material? Methodological issues

	For Issue 1, it is unclear how the authors intend to mitigate the issue. The sub-section ends with the statement “This leaves room for discretion how to record the information.” – Could this be more clearly described? This issue highlights the potential for information to be extracted non-systematically, which raises concerned with the protocol itself. “First, the study will record the unclear information as ‘no RWD.’ The separation of ‘not clear’ is an intuitive way to extract the data, however, it is not useful for the analysis. The code ‘not clear’ cannot be independently analysed. It will be combined into ‘no RWD’ when analysing the data. In addition, having a ‘not clear’ category is unlikely to improve data quality”. The ordering of this paragraph is confusing – are the authors saying that unclear data will be coded as ‘no RWD’? If so, I would avoid implying that there is a separate code ‘not clear’, as this is confusing. In addition, and more importantly, I am concerned that grouping essentially missing data within the ‘no RWD’ category could lead to misinterpretations of evidence if there are a substantial proportion of ‘not clear’ records. Have the authors planned for a possible alternative analysis where records afflicted by substantially missing data are simply omitted from the analysis? Issue 3 states that the research has only used the definition of RWD by FDA, yet earlier in the paper it was stated that (for each instance concerning definition) that two different definitions would be covered. In addition, the paper highlights much earlier that a definition of RWD has been established by NICE, yet this has not been used to inform this research which is considering NICE appraisals. The authors later comment that there are many definitions which do not differ greatly, and so it is “unlikely there will be a marked divergence in the data extracted when using the different definitions”. Considering all this information together, it is extremely confusing which definition(s) of RWD is/are being used, and why. It is also unclear precisely how they differ, which would help understanding for selecting a given definition over an alternative. I would recommend that the authors reconsider the presentation of the definition of RWD used for clarity and describe why the NICE definition was rejected in favour of the FDA/ Makady et al. definition. This section fails to describe how the authors will deal with missing information due to redaction – this is a critically important feature of any study which aims to look into NICE assessments. How have the authors planned to deal with this type of 'missing' information?

VERSION 1 – AUTHOR RESPONSE

Reviewer 1	
1. Definition of RWE a. The authors emphasise the importance of the frequency of data collection ("routinely collected") as being integral to	We are aware of this point from the interview with ERGs and committee members. We agree that “routineness” can be interpreted in

some definitions of RWE (e.g. that of the FDA). I think a more common interpretation of "routinely collected" is "collected in routine care" rather than frequency. This different interpretation shouldn't have major effects on the data extraction but is perhaps worth considering.	different ways. We found several cases where companies presented data from a phase 1 clinical study as RWD because it is routinely collected outside randomised clinical trials. As we are wary of leaving room for discretion, our paper adopts a definition of RWD combining definitions from FDA and Makady et al. We focus on data that is “routinely collected” from a “non-experimental setting.” However, our approach cannot solve all issues around definitions. One such issue concerns the different interpretations of “routinely collected”. We discuss this example in “Issue 3”.
2. Scope of studies covered a. While RWD generates substantial attention, often the interest lies in observational data more generally or, in the context of comparative effectiveness, the use of non-randomised studies, including those with interventional designs (non-randomised controlled studies or single arm trials with external control).	We agree. Our protocol focuses on RWD, which are collected from a non-experimental setting. Our study will not directly cover the interest in single-arm trials or other types of non-randomised trial. However, some aspects of the absence of head-to-head comparisons, and the use of unanchored treatment comparisons will be included when examining their association with the use of RWD
3. Providing context around the cancer Drugs Fund a. I think the general information extracted should clarify whether the submission was pre or post CDF (and +/- 2016 with reformed CDF & mandated data collection). b. The protocol could also explicitly state that the pre-and post-CDF evaluations are considered separately (where relevant).	Thank you for giving the opportunity to clarify the extraction tool. In the tool, there is a “replace.” It aims to capture whether the appraisal replaces previous appraisal or not. Although the reason for being replaced can be guessed with the information of recommendation in the CDF, clear information about an appraisal of the CDF review has strong benefits in data completion as well as the analysis. Hence, we have created two new variables, pre-2016 CDF reconsideration and 2016 CDF review (supplementary p.4, row 8,9). We believe that these variables allow us to distinguish appraisals which explicitly had managed access agreements specifying data collection. This will reduce potential bias with respect to the use of RWD by mandated data collection, and aid

	analysis of how RWD is collected and used in CDF reviews.
4. Extraction of data about use of RWD in modelling a. Is data extraction based on the preferred model from the ERG or the company submission? b. Are uses of RWD for sensitivity analyses excluded?	a. The information will be extracted from the company submission and from the final appraisal document (the model preferred by the committee). b. Please see the change on page 9. We plan to extract the data for both the base-case and for sensitivity analyses. Previously this was described near the end of the appendix. Now we made it clear in the body of the paper
5. Data analysis on factors associated with greater use of RWE a. Should the data analysis be restricted to those issues where RCTs would be the preferred source of evidence only (i.e., treatment effects)? For many other parameters (e.g., resource use) or event rates, RWD, assuming of sufficient relevance and quality, might be the preferred source of data and therefore less dependent on the limitations of the trial(s).	a. We are interested in the use of RWD more generally in our study. We believe this comment is helpful to interpret the result. However, the hypotheses include not only issues of trials but also the rareness of the disease, previously recommendation status of technology by NICE. Hence, we plan to test the hypotheses for all the outcome variables and aggregated groups given our research interest.
6. Questions of interest not captured by the review a. What are the characteristics of RWE submissions that make committees more willing to accept their use? b. What is the quality of RWE submissions across use cases? c. Where was the use of RWE proposed (either by companies, ERG, or committees) but rejected or otherwise not possible?	a. As we describe in “Step 2: Data extraction”, the use of RWD is extracted from both the company submission and the final appraisal document. We will review the acceptance of RWD by the committee as part of hypothesis testing. b. The quality of RWD is an important question. However, it is challenging to evaluate the quality of RWD in this study. During the data extraction, we collect benefits and challenges of the use of RWD if they are addressed by company/ERG/ committee in the documents. This is now described in supplementary table 1: “*** Benefits/challenges of the use of RWD are collected in outcome variables” (supplementary document p.13, bottom of the table). Although it won’t be directly used for

	the analysis, this information can aid interpretation of the results and understanding of the issue of quality of RWD indirectly. c. We plan to investigate the differences in the use of RWD between the company and committee.
Reviewer 2	
The authors have developed a protocol to aid with searching NICE assessments of cancer treatments for how RWD have been used to inform decision making. By following this protocol, the authors intend to extract information in a "reproducible, systematic and transparent" way. I believe substantial remedial work is required for the protocol to be deemed suitable for publication, based on clarity of reporting and justification of decisions made. Most of my comments are related to describing the planned approaches as opposed to major concerns with the planned analysis itself. However, there are also some concerns highlighted concerning potential for information loss, handling of missing data, and risk of inconsistent extraction/ reporting of information.	We address the reviewer's concerns in turn below.
General	
In general, some paragraphs are very large – suggest splitting some up where suitable. In addition, the paper would benefit from an English language review for flow. There are also a few instances of 'Error! Reference source not found' that need addressing.	We have trimmed paragraphs, particularly in "definition of RWD" following the suggestion. Errors with respect to reference sources are addressed.
Introduction	
"Health Technology Assessment (HTA) requires valid and reliable information for the systematic evaluation of health technology". For this audience, can an explanation of HTA be included here? Health technology is also not an immediately understood term, especially when used to define HTA – could this be expanded upon?	We briefly explained what HTA is: "Health Technology Assessment (HTA) refers to the systematic evaluation of clinical- and cost-effectiveness of health technology." (p.3, line 9-10)
"Furthermore, the traditional design of RCTs is possibly less appropriate for new technologies such as those targeting rare genetic mutations or where there may be ethical issues with control arms". Here, suggest clarifying you are talking about the principle of equipoise, and why single-arm or uncontrolled trials may be undertaken. Ethical issues could be misconstrued as other types of problems which do not seem relevant here.	To respond to this comment, we have focused more on why the traditional design of RCT is less appropriate and deleted the ethical issue. (p.3, line 16-18)
"Moreover, RCTs tend to include strictly controlled	We have changed the order in

populations". This is included at the end of an otherwise unrelated paragraph – suggest moving this point earlier in the text when describing other general issues with RCTs.	response to this comment. (p.3, line 17)
"As a leading HTA agency, NICE has…". It would be my personal view that the start of this sentence should be removed as I think it's an unsubstantiated claim (even though in my opinion it's true!)	We deleted these words in response to this comment. (p.4, line 6)
"The evidence is structurally well-documented enough to find the key information and available on the NICE website." Here, I think it would be clearer to say "… and is available on the NICE website" as this is a separate comment.	We revised in response to this comment. (p.4, line 11)
"Although this study follows a more systematic approach to review the use of RWD, it does not fully explain how the data were extracted and what criteria were used to judge the use of data". I agree with the authors that these aspects highlighted by the authors were not described fully in the study by Bullement et al. However, I think the authors could be clearer here about what is meant by "how the data were extracted" and "what criteria were used to judge the use of data". For the extraction, perhaps the authors could say: "a data extraction table was not provided" and for the criteria, perhaps it would be better to say "the authors focused only on how RWE influenced the cost-effectiveness analysis, and not RWE used to support the interpretation and/or perception of the results"?	To respond to the comment, we revised the text following the suggestion: "Although this study follows a more systematic approach to the review of the use of RWD, a data extraction table was not provided and the authors focused only on how RWE influenced the cost-effectiveness analysis, and not on how RWE was used to support or establish the appraisal." (p.5, line 11 -13). We note that Bullement et al. reviewed the use of RWD as inputs to the model, but not the use of RWD for justifying the model itself.
"As the process of reviewing appraisals is not clear enough, it is unclear whether the information presented provides a full picture of the use of RWD." This sentence is difficult to understand because of the use of 'not clear' to describe why something else is 'unclear'. Perhaps this would be easier to understand if the authors instead said: "Due to limited information presented concerning the review process in this study, it is unclear whether…"?	We have revised this in response to the reviewer's suggestion. (p.5, line 13-14)
The final paragraph in the Introduction is a little confusing – it is stated that the main purpose of the protocol is to extract data, but later it is noted that data will be analysed. Overall, I think this section could be better described, including specifically highlighting what the research questions are (as it is unclear exactly what these are) and signposting to the planned regression analysis which is only described very later in the paper. Also, I think it would be better to describe an	We have revised by describing the research questions: "The data can be analysed to answer the research questions including "how has RWD been used in NICE appraisals" and "which factors are associated with increased likelihood of the use of

analysis of data potentially providing a biased view of how RWE has been used, rather than the data themselves being 'biased'.	RWD." (p.6, line 8-10) Also, the brief explanation of the analysis is addressed at the beginning of the methods and analysis part.
Methods and analysis	
Figure 1: In my opinion, this diagram is unnecessary, and to an extent unhelpful – it raises extra questions (e.g., how are the data going to be validated? What is the analysis planned? Etc.) without answering them at this point in the paper. I would suggest as a minimum signposting in the diagram to where these are discussed within the paper, and if not discussed consider removing this diagram altogether.	We have removed Figure 1 in response to this comment.
"The information is extracted from identified appraisals in accordance with extraction rules". What are the extraction rules? These should be explained here.	Regarding the comment on "extraction rule", it is difficult to explain the extraction rules for each variable in a short sentence. Given the word limit, we prefer to provide all extraction rules in the supplement rather than in the body of the paper. In order to direct the reader to the detailed extraction rules in supplement 1, we added "The detailed extraction rules can be found in supplement 1." (p.6, line 14-15)
"The extraction tool includes general appraisal information and appraisal-specific information such as characteristics of the main clinical evidence and the economic evaluation model". An example of appraisal-specific information is noted, but what do the authors mean about general appraisal information, and how is this different to appraisal-specific information? From my understanding, the authors might be referring to things like TA number, data of publication, etc., but it is unclear to me why this is not considered information specific to the appraisal?	We have changed the description of the information categories. Evidence-related information refers to information used in the evidence submission. Other information refers to other general information such as TA number, date of publication which are not part of the evidence submission. (p.6, line 15-17)
A whole page of text to describe the definition of RWE seems excessive to me – I think this could be condensed substantially. Could the authors simplify this into a shorter paragraph noting that there is no standard definition, citing a range of alternative studies which give their own definition, and then provide their own definition?	We reduced discussion of different definitions of RWD and have focused on the definition which we use in our manuscript (p.7).
The reason(s) for the use of two different definitions of RWD are unclear to me when this is described within the methods and analysis section. Can this be succinctly summarized by the authors and added into the paper?	Thank you for the opportunity to clarify the definition of RWD. After careful review, we conclude using a single definition by focusing on "routinely collected" and "non-

	experimental setting." (p. 7, line 6-11)
"Relevant appraisal documents including the final scope, the manufacturer's submission, the evidence review group (ERG) report, and the final appraisal determination are available for each appraisal. The appraisal documents are reviewed to establish whether RWD is used to determine any components of the economic evaluation". Here, the authors state that appraisal documentation includes these, but I assume this is not an exhaustive list (e.g., clarification responses were also considered where applicable?). If so, I would omit the examples, and limit this to say "appraisal documentation", as the wording otherwise may imply some documents were missed.	Clarifications can be important in terms of understanding the analysis presented by a manufacturer or the ERG but with respect to the use of RWD they are not an additional source of information. Hence, we exclude the clarification responses or other documents for appeals in our study. We believe that addressing four type of documents is necessary for clear understanding.: "Among the documents, this study only reviews four type of appraisal documents, the final scope, the manufacturer's submission, the evidence review group (ERG) report, and the final appraisal determination." (p.7, line 20-22)
"This research exclusively includes single-technology appraisals (STA) of oncology medicines.". It would be helpful to clarify what the difference between STAs and MTAs is (acknowledging that the audience may not be aware of the differences in NICE TA processes). I think it's perfectly reasonable to restrict to STAs, but clarification would be helpful here.	To respond to this comment, we have provided the reason to exclude MTAs in our review (p.8, line 7 – 13).
Some comments on Figure 4/ Supplement 1 are presented in the list below. Here, these all highlight areas where a different researcher may record information in a different way based on interpretation of the protocol:	
• I'm unclear why outcome variables need to be separated by parametric versus non-parametric and disagree with the use of these terms here – it may be that a non-parametric estimate of OS informs the model, but this would be categorized here as 'parametric'. Instead, would it be clearer for the authors to use 'model input' versus 'not a model input'?	We have carefully considered your suggestion to use "model input" and "not a model input" rather than "parametric" and "non-parametric". In the section "Parametric and non-parametric use" we clearly state how we are using the terms and thus do not believe our usage is unclear.
• Recommendation: Is it not possible that there could be another recommendation (e.g., recommended in research?). Might not apply to any of the STAs included in your review, but I would check for possible recommendation types (including, for example, terminated appraisal). Suggest looking at the Excel file available via this link: https://www.nice.org.uk/about/what-we-do/our-	Our preliminary work suggests that "recommended in research" has not been used in cancer appraisals but we agree we should allow for the possibility. Hence, the code, "recommended in research" is included (supplementary document

programmes/nice-guidance/nice-technology-appraisal-guidance/data/cancer-appraisal-recommendations	p.5, row 1). Our inclusion/exclusion criteria states that terminated appraisals are excluded. Therefore, this code is not necessary. We exclude them because there is no information whatsoever about any of the data considered by the manufacturer or anyone else, prior to the decision to terminate the appraisal.
• Incidence is often reported as a range – is this planned to be reflected? Guidance perhaps could be added for data such as incidence which might be reported as a range	To respond the comment, details are added: “Most appraisals present the annual estimate of the number of patients who are eligible for the treatment in the “Budget Impact” section of company submission. This number is mainly used for the incidence. If this information is not available in the appraisal, the number in a previous appraisal for the same indication is used instead. (supplementary document p.5, row 11)”
• Does the extraction account for studies with more than 2 arms? How should this be recorded?	There appear to be few cases where trials have more than two arms. In such cases, only the arms considered as relevant for decision problem in evidence submission are included. If there are two intervention arms and these arms are separately used for different indications in appraisals, the data extraction is carried out separately. This is explained in “operational separation.” When two arms are relevant as comparators for same indication, the data are recorded without distinguishing these arms. (supplementary document p.13, row 5)
• Does number of participants matter by arm, or are you just collecting for the full population? Might be a situation where the comparator is essentially discarded if it’s irrelevant for decision making	The number comprises all participants in the trial. This information is ancillary, collected in order to see how many people are included in trials generally. It will not be used directly in the analysis, but

	can give some background, such as whether rareness of disease has an impact on the size of trials. We believe it is not necessary to separately record the number of participants in each arm for our purpose.
• How is the risk of bias planned to be ranked/graded? This isn't clear from the extraction template – numbers are provided, but unsure where they are from/ how a reviewer would decide which rank to assign	Further detail is added in response to the comment: "The ERG assesses the risk of bias of the included study using quality assessment tools. The ERG statement is directly quoted. The ERG often addresses the issue of quality of study narratively. Moreover, the ERG uses different terminology, whereas the domain of assessment is consistent. Therefore, the risk of bias would be narratively recorded. Prior to analysis, it will be scored by looking at the number of factors about which the ERG has expressed concern. (supplementary document p.6, row 10)"
• I'm not sure grading maturity on % OS events is a universally appropriate mechanism to describe maturity. There may be some appraisals which simply don't report much in terms of OS, because they're for earlier stage disease. However, I'm not sure I have a better suggestion – is there some way of accounting for studies that are understandably 'immature' versus those that are actually just reflecting short follow up? Maybe there's a way of capturing median/minimum follow up, or expressing events as a proportion of the planned events per the primary end point of the study?	It is true that there is no universal way to measure maturity. However, Tai & Latimer (2021) investigate data maturity in STAs by looking at the proportion of deaths in pivotal trials. This protocol adapts this criterion for measuring maturity. This additional explanation is included in the supplementary document. "This protocol adapts the criterion for measuring maturity of survival data in Tai et al. which investigates data maturity in STAs by looking at the proportion of deaths in pivotal trials. In their study, 20, 50 and 70 % of total deaths are used to discuss the maturity of survival data (1). This protocol only uses 20% and 50% to assess maturity, without the category "unclear." (supplementary document p.7, row 7,8)"
"In case of the variables coarsely divided, the outcome of the extraction is so blunt that it cannot fully capture how RWD is used. Likewise, the variables overscrupulously divided are less likely to provide valid outcome to show the pattern of the use of RWD in the analysis". The phrasing of this statement is	This is an important point. We thank the reviewer for the suggestion. We have revised this part based on his suggestion: "This coding system has advantages which include avoiding

concerning, as it suggests that data will be coded such that it maximizes the chance of showing the results the authors are hoping to see. I would suggest the authors revisit this statement to explain that coding data effectively has advantages which include avoiding information loss, and also grouping 'similar' information across appraisals to establish patterns of RWE use.	information loss, and also grouping together 'similar' information used during appraisals to establish patterns of the use of RWD." (p.11, line 10-12)
"Under the parametric use, the clinical effectiveness, health utility, cost side were thoroughly reviewed." Here and elsewhere, I suggest the authors refrain from using the phrase 'cost side' as this is not a standard term. Instead, the authors may wish to instead describe features of the submitted economic analysis, including inputs related to clinical effectiveness, health utility, and costs.	To respond to this comment, we re-worded it to make the meaning clearer: "Under parametric use, clinical effectiveness, health utility and cost and healthcare resource use were thoroughly reviewed. (p.11, line 19)"
"The extracted data will be analysed quantitatively in two different ways. First, a descriptive analysis will summarise where and how RWD has been used in appraisals. This will be supplemented by an analysis of the intensity of use of RWD in order to explore changes in the pattern of use of RWD over time and differences with respect to cancer type. Secondly, a regression analysis will be performed to investigate which factors are associated with the greater use of RWD in a company's submission". From my perspective, this instead sounds as though a qualitative summary of the appraisals will be undertaken, and then a quantitative regression analysis will be performed. I suggest the authors edit this description to clarify the analysis methods proposed.	As the reviewer suggests, we added the sentence "In addition to descriptive statistics, the association between years and the intensity of use of RWD will be examined. (p.12, line 19-21)"
"A literature review and a pilot study were conducted to identify factors potentially associated with the use of RWD". Where are the review and pilot study findings – could these be referenced or included as supplementary material?	We re-worded the literature review and a pilot study as these were not independent processes, but parts of the protocol development to identify the issues over time (p. 13, line 1).
Methodological issues	
For Issue 1, it is unclear how the authors intend to mitigate the issue. The sub-section ends with the statement "This leaves room for discretion how to record the information." – Could this be more clearly described? This issue highlights the potential for information to be extracted non-systematically, which raises concerned with the protocol itself.	NICE appraisal documents are well-structured, which facilitates the identification of what kind of evidence is used. It appears to be rare for there not be an explicit statement regarding the evidence used (mostly with respect to resource use). We record this as no use of RWD as we have explained in the paper. This approach can involve a loss of information. However, this

	problem appears to arise in very few appraisals. Also, the information which is not clearly recorded in the appraisal documents is usually not major information with respect to the evidence synthesis. Therefore, we don't think this issue challenges our systematic approach to data extraction.
"First, the study will record the unclear information as 'no RWD.' The separation of 'not clear' is an intuitive way to extract the data, however, it is not useful for the analysis. The code 'not clear' cannot be independently analysed. It will be combined into 'no RWD' when analysing the data. In addition, having a 'not clear' category is unlikely to improve data quality". The ordering of this paragraph is confusing – are the authors saying that unclear data will be coded as 'no RWD'? If so, I would avoid implying that there is a separate code 'not clear', as this is confusing. In addition, and more importantly, I am concerned that grouping essentially missing data within the 'no RWD' category could lead to misinterpretations of evidence if there are a substantial proportion of 'not clear' records. Have the authors planned for a possible alternative analysis where records afflicted by substantially missing data are simply omitted from the analysis?	'Unclear' is recorded separately in order to provide a more accurate description of the use of RWD. However, 'unclear' is not commonly found. For purposes of data analysis we anticipate treating these instance as "no RWD". As you can see in the appendix, the variables are finely divided. Most information required for this extraction is available in the appraisal documents. Some information such as incidence rate, or maturity is likely to be unavailable. How to mitigate these issues is described individually in supplement 1. The protocol has been developed, in part by reviewing appraisals to identify the main issues which arise and need to be addressed. As part of this process, we find that missing data is not a major problem. If it is unclear whether the evidence is RWD or not, the study design in the original paper can be checked. The issue of code 'not clear' is raised mostly with respect to resource use. It is unlikely to have a major impact on our analysis.
Issue 3 states that the research has only used the definition of RWD by FDA, yet earlier in the paper it was stated that (for each instance concerning definition) that two different definitions would be covered. In addition, the paper highlights much earlier that a definition of RWD has been established by NICE, yet this has not been used to inform this research which is considering NICE appraisals. The authors later comment that there are many definitions which do not differ greatly, and so it is "unlikely there will be a marked divergence in the data extracted when using the different	The issue of definition is important and we appreciate the opportunity to clarify our choice. NICE have not published a definition of RWD. The definition provided earlier is presented in a NICE-associated document. As there is no clear definition of RWD which NICE prefer, we used the most widely used definition. We use a definition

definitions". Considering all this information together, it is extremely confusing which definition(s) of RWD is/are being used, and why. It is also unclear precisely how they differ, which would help understanding for selecting a given definition over an alternative. I would recommend that the authors reconsider the presentation of the definition of RWD used for clarity and describe why the NICE definition was rejected in favour of the FDA/ Makady et al. definition.	merging the definition of the FDA and Makady et al. to minimise the operational flexibility. "This research uses the definition of RWD merging definitions by FDA and Makady et al. The distinctive part of the definition used in this research is 'routinely collected' data from 'non-experimental or non-interventional study'." (p.14 line 14-16) Although some findings might differ, any divergence is unlikely to be substantial as the definitions overlap considerably.
This section fails to describe how the authors will deal with missing information due to redaction – this is a critically important feature of any study which aims to look into NICE assessments. How have the authors planned to deal with this type of 'missing' information?	Please see the added information on page 7 in the supplementary document. We deal with missing information by clear rules. However, the reason why we don't emphasise missing information in this protocol is that our preliminary work suggests that most of the information to be extracted in this study is available. The information that is usually redacted is information concerning the clinical trial such as median follow-up and number of events. This information is not central to our analysis. However, when it comes to measuring maturity, it is a problem. We already recognise this issue and take several approaches to improve the completeness of the data. First, the reference is checked. When the clinical information is not available, the published paper reporting the clinical data is reviewed. If still this information is not available, we use a surrogate indicator.

VERSION 2 – REVIEW

REVIEWER	Seamus Kent University of Oxford, Nuffield Department of Population Health
REVIEW RETURNED	11-Nov-2021

GENERAL COMMENTS	I would like to thank the authors for addressing the comments. I look

	forward to seeing the results of the analysis.

REVIEWER	Ash Bullement The University of Sheffield, ScHARR
REVIEW RETURNED	17-Nov-2021

GENERAL COMMENTS	The authors have taken on board several of the comments raised both by myself and the other reviewer, which I believe has improved the paper substantially. However, I am still concerned with several aspects of the paper which should be resolved before this manuscript may be deemed suitable for publication. These are mostly focused on presentation and/or explanation, rather than some of the more methodological concerns raised in my previous review and presented separately below. The authors note within the protocol several sections of text which imply data extraction has already commenced, which include: “this problem appears to arise in very few appraisals” and “the data extraction is planned from January 2020 to October 2021”. Any reference to extraction having been started should be removed from this protocol. As highlighted in my previous set of comments, the paper would benefit from an English language review for flow. Please see below some specific sections of text that require attention, though this is by no means an exhaustive list and a more thorough editorial review would be of great benefit to the paper: “Restricted population makes replication of finding challenging” “… different actors have the principal responsibility for producing the main evidence in each process” “Even same definition can be interpreted in a different way” The authors have revised their definition of HTA, but I am still concerned that the definition of HTA is effectively defined on the basis of evaluating a health technology. I would suggest that the authors explain what is meant by a health technology or provide some examples (such as medicines and devices). Re-iterating a point raised in my original review, it is my view that the authors introduction conflates the purpose of the extraction template/ protocol, with the overall intention behind populating the template/ executing the protocol – to consider an analysis of RWD use in NICE assessments. I would suggest changing the following text: “With such data, the analysis can provide more robust answers to questions regarding how RWD has been used in NICE technology appraisals” to “By consolidating these data, subsequent analysis can provide more robust answers to questions regarding how RWD has been used in NICE technology appraisals”. While I believe relevant RWD may arise at the clarification or technical engagement stages, I understand that should such data be considered material to the final decision, then the relevant information would be cited in the final appraisal determination document (FAD). I would, however, suggest the authors comment on the possibility that some RWD may be provided outside of the four documents noted (acknowledging the FAD is by necessity short in length), and so this is perhaps a limitation of the planned study (even though the risk of missing relevant information is low). I am still unclear what the authors mean by a “descriptive analysis”

	which is described as “quantitative”, and so a comment I raised previously was to enquire whether this analysis was truly quantitative or if this could be better explained as a more qualitative analysis. Please can the authors provide further explanation on the analysis methods and adjust the explanation within the paper accordingly? Readers should be able to clearly understand what will be done with the data. The explanation of the regression analysis is clear, but this first analysis is not.

VERSION 2 – AUTHOR RESPONSE

Reviewer 2	
The authors note within the protocol several sections of text which imply data extraction has already commenced, which include: “this problem appears to arise in very few appraisals” and “the data extraction is planned from January 2020 to October 2021”. Any reference to extraction having been started should be removed from this protocol.	The extraction has been started in some appraisals as a part of developing and refining the protocol. However, the data extraction is still in progress. We have revised the text following this suggestion: “In addition, having a ‘not clear’ category in the analysis is unlikely to improve data quality since we anticipate that this problem will arise in very few appraisals.” (p.15, line 17) Also, the date for finishing extraction has been overtaken by time. The extraction timeline is no longer relevant information. Hence, we have removed it (p.6).
As highlighted in my previous set of comments, the paper would benefit from an English language review for flow. Please see below some specific sections of text that require attention, though this is by no means an exhaustive list and a more thorough editorial review would be of great benefit to the paper: “Restricted population makes replication of finding challenging” “… different actors have the principal responsibility for producing the main evidence in each process” “Even same definition can be interpreted in a different way”	We have revised the specific statements identified by the reviewer, and in the same spirit made revisions throughout the paper. “Moreover, RCTs often have strict inclusion criteria reducing generalizability” (p.3, line 21-22). “It is challenging to gather the same information in the MTA process as different actors are responsible for producing and reviewing the main pieces of evidence.” (p.8, line 10-11). “Although this definition provides a specific and clear definition for this research, there is no consensus on the best definition of RWD. Even the same definition can be interpreted in different ways.” (p.14, line 19-22).

The authors have revised their definition of HTA, but I am still concerned that the definition of HTA is effectively defined on the basis of evaluating a health technology. I would suggest that the authors explain what is meant by a health technology or provide some examples (such as medicines and devices).	To respond to this comment, we have clarified what health technologies are (p.3, line 12 -14).
Re-iterating a point raised in my original review, it is my view that the authors introduction conflates the purpose of the extraction template/ protocol, with the overall intention behind populating the template/ executing the protocol – to consider an analysis of RWD use in NICE assessments. I would suggest changing the following text: “With such data, the analysis can provide more robust answers to questions regarding how RWD has been used in NICE technology appraisals” to “By consolidating these data, subsequent analysis can provide more robust answers to questions regarding how RWD has been used in NICE technology appraisals”.	We have revised the text in response to this comment (p.6, line 7).
While I believe relevant RWD may arise at the clarification or technical engagement stages, I understand that should such data be considered material to the final decision, then the relevant information would be cited in the final appraisal determination document (FAD). I would, however, suggest the authors comment on the possibility that some RWD may be provided outside of the four documents noted (acknowledging the FAD is by necessity short in length), and so this is perhaps a limitation of the planned study (even though the risk of missing relevant information is low).	To respond to this comment, we have included this point under ‘strengths and limitations’ (p.3, line, 6-7 & p.17, line 15-20).
I am still unclear what the authors mean by a “descriptive analysis” which is described as “quantitative”, and so a comment I raised previously was to enquire whether this analysis was truly quantitative or if this could be better explained as a more qualitative analysis. Please can the authors provide further explanation on the analysis methods and adjust the explanation within the paper accordingly? Readers should be able to clearly understand what will be done with the data. The explanation of the regression analysis is clear, but this first analysis is not.	We now describe the first analysis as follows: “First, counts and proportions will summarise where and how RWD has been used in appraisals.” (p.12, line 20).