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)	An investigation of routes to cancer diagnosis in ten international jurisdictions, as part of the International Cancer Benchmarking Partnership; survey development and implementation
AUTHORS	Weller, David; Vedsted, Peter; Anandan, Chantelle; Zalounina, Alina; Fourkala, Evangelia-Ourania; Desai, Rakshit; Liston, William; Jensen, Henry; Barisic, Andriana; Gavin, Anna; Grunfeld, Eva; Lambe, Mats; Law, Rebecca-Jane; Malmberg, Martin; Neal, Richard; Kalsi, Jatinderpal; Turner, Donna; White, Victoria; Bomb, Martine; Menon, Usha; Working Group, ICBP

VERSION 1 - REVIEW

REVIEWER	Mary L McBride British Columbia Cancer Agency 675 West 10th Avenue Vancouver, British Columbia CANADA V5Z 1L3 Member (co-investigator) of a Canadian Institute of Health Research-funded Team grant on gaps between oncology and primary care for breast cancer patients with Drs Eva Grunfeld (PI) and Donna Turner (CI); also Dr Peter Vedsted is advisor on this Team grant.
REVIEW RETURNED	30-Oct-2015

GENERAL COMMENTS	This is an important paper, necessary to provide the appropriate methodologic background with which to analyse results from Module 4 of this international study. This module addresses a critical component of the cancer trajectory that has a significant effect on treatment intensity, survival, healthcare costs, and patient burden. The effort to compare international experience on this topic using a survey is novel, and should provide important information for all jurisdictions on gaps in care. The term “survey” is used in two contexts in this paper. As examples, the title and text refer to the study methods; however, the abstract objective, and design and setting, as well as the title of supplementary file 1, appears to refer to the data collection instrument. The paper would be more clear if the uses of this term were made distinct. Other aspects of the abstract could be made more clear and consistent with the text. For example, “Main Outcomes” includes one outcome and a statement of preliminary results; rather than results, additional outcomes discussed in the paper, relating to evaluation of the survey tool and recruitment development process, eg the kappa reliability scores, or development of “data rules” for reconciliation of responses, and progress, would be informative for the reader to

	include here (space permitting) . In the “main outcomes” section, it is not obvious that high response rates are indicative of validity of the tool; I suggest removing this comment. The comments on response in Results state a current recruitment rate of 10,000. I believe the word “rate” is misplaced, since this is a number, and, since this number is hard to reconcile with the target numbers quoted in the text, a more useful number might be the number still required to achieve complete ascertainment by type of cancer and jurisdiction. As this is an international paper and not all readers may be familiar with the abbreviation GP, the authors may want to consider spelling this out. Methods include a clear statement of design, information on governance and funding, and a framework for the study aims, that informs development of the survey tool. The reporting source(s) (?patient and/or provider) for each of the critical dates in the diagnostic process would be useful to add, either when these dates are defined in the section “Measuring time-points and intervals”, or in the section on Questionnaire development, rather than later on. In describing the various types of patient diagnostic journeys, semi-colons after each option would clarify the sentence where combinations of earlier options are considered. Pilot study: Received questionnaire response rate was quoted with all questionnaires (120) in the denominator (36%; 43/120). Since questionnaires could not be received back if they were not forwarded, the more appropriate denominator would appear to be 72 (59.7%; 43/72). Also, the exact percent (rather than “many”) of patients who provided consent to have their providers contacted would be useful, if it was available, as it would provide the denominator for the provider response rates. Referencing the IARC publication Cancer Incidence in Five Continents, which has data quality indicators by registry, would strengthen the statement of registry data quality. I believe a word is missing from the sentence fragment “…where questionnaire-based data collection from primary and secondary (ADD WORD HERE) was not feasible”. A query in the statistical calculations: In the calculation of a sample size of 200 cases per jurisdiction, is the 10% difference between countries (as stated) or between jurisdictions? Analysis plan: The first mention of health-related visits and investigations that I detected is in the analysis. A description in the Methods section of how these data are collected, and from which of the three identified data sources (or another data source?), would be helpful. In the section on “Recruitment progress/response rates”, nformation on mean time from diagnosis + 3-6 months to registry referral for study recruitment would be useful. Results and Discussion: The sentence in the first para would read better if the authors replaced one instance of the word “challenging” in the sentence fragment “It is a challenging exercise with a broad range of methodological challenges;…” A comment on the previous experience, in the different jurisdictions, of non-response or delays in response using registry ascertainment would add to the discussion of reasons for non-response.

REVIEWER	Gary Abel University of Cambridge, UK
REVIEW RETURNED	05-Nov-2015

GENERAL COMMENTS

This paper falls between a traditional paper reporting a questionnaire development process and a study protocol for an ongoing study. Whilst I see no problem with these two aspects sitting side by side I feel that the paper, to some degree, fails to deliver on either front. This is by and large due to the fact that the paper is very good at outlining the principles behind various aspects of the work it fails to deliver many specifics. I assume that the main purpose of this paper is to act as reference material for future publications, which is fine, but without supplying the specific outcomes of the questionnaire development and the final processes used in the ongoing study it will be of limited value. Notable absences include the fact that the final questionnaires are not included as supplementary material, that operational definitions of the various time points have not been given (presumably they would be obvious with a questionnaire), that the comprehensive set of data rules for combining data from the three sources are not included as supplementary material nor even any indication of how they might work and to which source in which circumstance preference is given. The paper as written is largely a set of assurances that sensible process will be applied. If the paper remains to be only this I do question its future worth as a reference source.

The other major shortcoming is the discussion of statistical aspects. As a Statistician I am particularly sensitive to this area, but I feel it requires considerable improvement in the clarity of the writing. In particular the sample size section is utterly confusing and I have no idea as to how or on what basis the sample size has been derived. The analysis plan is vague such that it is no more than a generic thought on how various types of data would be analysed. Nowhere are any outcomes (primary or secondary) defined or any particular scientific questions outlined.
These aspects could easily be addressed making the paper a much more useful future resource.

Detailed comments
1. An overarching comment – there is somewhat of a miss-match between the stated objectives and what is presented as results rather than methods. The objective is the development of a survey (see comment 2 below for more on this) but all results of the development (i.e. results of cognitive testing, test-retest reliability etc. are contained within the methods and the actual results (brief and bundled with the discussion) relate to preliminary findings regarding response rates in the on-going study. Personally I have no issue with the paper as currently structured, but I wonder if the continued use of the section labels “Methods” and “Results” is not advisable and that a different structure should be adopted with appropriate section titles. This potentially applies to the abstract.
2. The objectives section of the abstracts states that the objective is the development of a survey – I would suggest breaking this down into the development of a questionnaire/instrument and the design of a survey.
3. The participants, main outcomes and results section of the abstract relate to the main study, rather than development of the survey as is the stated objective.
4. The main outcomes section of the abstract would normally be a

specification of outcome measures. What is currently there is a conclusion.

5. The article summary should specify that these statements apply to cancer.

6. The first bullet point in the article summary is not a sentence.

7. The sentence in the background section “Survival differences between populations are most probably due to a range of factors including lifestyle, levels of comorbidity, availability of screening programmes, primary care system, and availability and quality of diagnostic and treatment services.” would benefit from references

8. On line 27 of page 7 the patient, primary care, diagnostic and treatment intervals are mentioned, but have not as yet been defined. This may be an issue for some readers so at the very least it may be useful to signpost the reader to later in the paper where these are defined.

9. Page 7 line 42 – the sentence “Routes to diagnosis have an important influence on cancer outcomes.” Makes one think of hard cancer outcomes such as survival or other medical issues. This is then followed by two patient experience examples before returning to survival. Whilst I think it is entirely right that patient experience is given the prominent position here this either needs flagging as being a “cancer outcome” or else the paragraph should be restructured to improve the flow.

10. Page 8 line 4. I think that reference 12 is the wrong one to use here. Also I would question your interpretation. Emergency presentations defined in the “Routes to Diagnosis” project include both those who present in an emergency setting (such as A&E) and those who present to a general practitioner who then instigates an emergency referral. As such the only contact they may have prior to entering the secondary care system (which is where the RTD project considers the emergency presentation to start) may have been a single consultation where the GP acted appropriately and is still a manifestation of an unavoidable emergency presentation. Currently, there is no evidence, to my knowledge to quantify how often, patients consult with a GP prior to either the instigation of emergency referral or attendance at A&E.

11. Page 9 – The final sentence of the background section is a far better summary of the objectives than that found in the abstract and the abstract would benefit from a similar sentence to this one.

12. Measuring time points and intervals. It was not clear to me, that precise operational definitions of all intervals have been given. The bulleted text on page 10 outlines potential challenges and table 1 outlines the conceptual definitions. Presumably if the questionnaire was present this might become clear. This is particularly a problem for date of diagnosis because even a conceptual definition is missing, just outlines that different conceptual definitions exist. The paper states that the respondents would outline their understanding which would be aligned with international standards, but I am then left wondering how comparable these dates will be when no standard definition is available.

13. Page 12, Routes to Diagnosis. It is noted that the questionnaire draws on previous RTD work, however, the resulting definitions are different from those used in previous work and this should be noted. (See comment 10 above for one such difference).

14. The bullet points at the top of page 13 outline the importance of using data from PCPs and STCs but not from patients. Perhaps this is self-evident to the authors, but I think specifying why patient reported data is important would be of benefit.

15. As I said above it would be beneficial to include the final questionnaires as supplementary material, but if this is not possible

(and possibly even if it is) the authors should at least summarise the number of items in each questionnaire and the topics covered along with the number of questions used for each topic. As it stands the reader is left in the dark about the contents.

16. Page 15 line 31 – what other pilots? Some details would be useful.

17. Reliability section – I find it worrying that the test-retest reliability is based on only 12 patients. This is a very small number. This should at least be acknowledged as a limitation. Whilst I accept that statistical significance if not often reported for kappa statistics, given the small sample size it would be interesting to know if the level of agreement was consistent with chance.

18. I am rather confused by the use of a weighted Kappa for the reliability of dates. Kappa statistics are used to quantify agreement in categorical variables, whilst dates are, in effect, continuous variables. Given in this particular case there are only 12 patients, there must be 24 or fewer unique dates. As such it would be possible to apply a kappa (weighted or unweighted) statistic, but I would suggest that this is inappropriate. I would suggest that a classical inter-rater reliability based on a one-way analysis of variance would be more appropriate.

19. Why has test-retest reliability not been considered for the other two questionnaires?

20. Page 21 line 40 – Errors can only be corrected for the 10% of questionnaires which have been checked. Presumably however, this is not the prime purpose of checking for errors or all data entry would be checked. Presumably the prime purpose is to check the error rate is low enough not to be of concern. When would this prompt a concern?

21. Analysis plan section. I would suggest reordering the section such that data manipulation/combination comes before the outline of analyses as it will do when analysis is performed.

22. As mentioned in my opening paragraph details of primary outcomes would be useful in this section and/or questions to be addressed.

23. Page 22 line 16 – why will prevalence rate ratios be used to look at diagnostic routes rather than logistic (or multinomial logistic) regression as you will be considering the proportion of your sample diagnosed through an emergency route, for example.

24. Page 22 line 23 – multi-level models are mentioned but with no justification nor with any detail of when they will be used and what the clustering variable will be.

25. Page 22 line 27 – the comprehensive set of rules could be provided in supplementary material along with a sense of in what circumstances each of the data sources takes precedence over the others.

26. Page 22 line 55 – It is stated that ecological analyses will be undertaken, presumably with each jurisdiction as the unit of analysis. This means that the sample size will be 10, which is very small and as such will have limited power. In fact such studies will have only 80% power to detect a correlation coefficient of 0.79. I would suggest that this is inadequate, particularly if multiple testing is factored in.

27. The authors' contribution is less than enlightening. Given there are so many aspects covered in this paper I would have thought some attribution would have been possible – who did the cognitive testing, who did the reliability analysis, who did the sample size calculation etc.

REVIEWER	Jessica Sheringham UCL, England
REVIEW RETURNED	11-Nov-2015

GENERAL COMMENTS	Thank you for inviting me to review this interesting paper. I note the review criteria employed relate to scientific credibility and research/publication ethics only. They do not require a judgement on the significance of the study. My comments therefore are mainly in relation to demonstrating its scientific credibility by enhancing clarity in a few key aspects: Abstract 1. Objective - this doesn't note the survey was to collect data on differences in diagnostic pathways IN CANCER. I think this is the case, rather than being developed as a generic tool? 2. Results – would be useful to reflect the response rates here (to give credibility to why they are encouraging). Methods 3. I see the survey developed is available for readers on request (p23, line 48), but without seeing the survey I didn't feel I could gauge whether the study has met its aims in designing a tool that can capture international differences in diagnostic pathways and intervals. Is there a reason for not making the survey available with the paper? 4. The analysis plan and the statistics used looked appropriate and the justification for using % over the median duration for diagnostic interval was clear. However, I am not a statistician so have suggested the paper needs statistical review. 5. A minor observation: from our qualitative research of emergency pathways, the diagnostic pathway may influence the degree to which credible diagnostic intervals are possible to determine, and also their relevance to patients' experiences of diagnosis. For patients diagnosed as emergencies, their pathways were characterised by circuitous, prolonged pathways, where consultations were not necessarily a conduit to diagnosis.1 Results 6. Overall, the paper structure of having results and discussion together with some results presented in methods was confusing to me. In particular: 7. In the abstract, response rates are presented in results. In the paper they are reported in methods and in supplementary data. As a consequence, I almost missed them in the paper. Given response rates are the main outcome metric for this paper, I think these should be reported in the body of the results. 8. The reporting of response rates (methods p23) makes just passing reference to the fact Denmark has better response than the UK countries and the results/discussion focuses on the potential for patient-level factors to influence response (p24 line 36-p25 line 2). However, the differences between UK countries and all the other jurisdictions seems quite stark and has implications for the final ICBP study's validity. Therefore, I think it would be worthy of more discussion to address such questions as: – is it administration method or other factors responsible for the differences in response rates? Were other administration methods explored in the UK? – why did primary care in the UK countries not send what looks to be a significant number of questionnaires? (it would be useful, if not for this paper but for the administration of the survey to record the numbers excluded by reason, i.e. that not actually have cancer, or

	were not aware of their diagnosis – the former group rightly should not be included in the study, but the latter group should and patients' lack of awareness of their diagnosis is another indication of problems in the system.) – how do the response rates compare with response rates from other surveys using similar administration strategies? (i.e. if primary care recruitment strategies normally lead to lower response rates, then the abstract's conclusion that responses were encouraging would be more justified). In summary my recommendations for revision are: 1. Make the requested changes to the abstract 2. Include the survey as an appendix to the paper or justify why this is not necessary or appropriate 3. Include the response rates in the results section 4. Discuss the differential response rates by jurisdiction in more detail in the discussion 1. Black G, Sheringham J, Spencer-Hughes V, et al. Patients' Experiences of Cancer Diagnosis as a Result of an Emergency Presentation: A Qualitative Study. PLoS One 2015;10(8):e0135027.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1
This reviewer notes that it is an important paper which will provide the methodological background for our international study.

• Survey versus data collection instrument - we now use these terms entirely distinctly

• Main outcomes in abstract - we have restructured have also removed the comment about high response rates, comments on response and results section. We have moved the figure of 10,000 and made a comment about our likely achievement of complete ascertainment of type of cancer and jurisdiction, and we've not commented on the validity of the instrument

• Use of abbreviation 'GP' – we have now removed all of these abbreviations, replaced with primary care provider ('PCP) and spelt the term out

• The reporting source for critical dates in the diagnostic process – we now spell this out much more clearly under the 'measuring time points and intervals' heading

• Various types of patient diagnostic journeys – we now use semi-colons after each option

• Pilot study response rates – the reviewer highlights the difficulty in our two stage questionnaire distribution process and we've addressed the issue of denominator (120 versus 72) and the exact percent of patients who gave consent to have their providers contacted

• IARC Cancer Incidence – we've now included the IARC publication Cancer Incidence in Five Continents reference

• Missing word in sentence beginning "where questionnaire-based data collection from primary and secondary…" - we have now added this

• Sample size – we now make it clear that we are looking for a difference of 10% between jurisdictions, not countries

- Health related visits and investigations – we now include a section under ‘Methods’ on how these data were collected

- Mean time from diagnosis + 3-6 months to registry referral: unfortunately we can’t produce a meaningful estimate of this, and it can change over time in our participating registries. We do, however, now provide an estimate of range across jurisdictions, based on discussion with jurisdiction leads

- ‘It’s a challenging exercise with a broad range of methodological challenges’ – we have removed one of the ‘challenges’

- Reasons for non-response - we have added in some material on this.

Reviewer 2

- Final questionnaires not included as supplementary material – we have now addressed this and made the questionnaires available via a web-link. We have also produced a new table (Table 2) which gives an overview of questionnaire content and questionnaire extracts

- Operational definitions of the various time points – readers who so wish can now access the questionnaires; we have also provided more material in the text on definitions of the time points.

- Sample size section – this has now been rewritten and we believe it is much simpler to follow (see below).

- Data rules – see our response below

- Analysis plan – this is also been rewritten and we have outlined our key scientific questions.

- It’s not straightforward to define ‘outcomes’ in an international descriptive survey but we are now clearer about the precise research questions, and expected data the survey will produce.

- Use of the terms ‘methods’ and ‘results’ – we have taken this reviewer’s advice and re-structured the paper under different headings. It is always quite difficult to fit a methods/protocol paper into a traditional aims, methods, results, discussion format and we think that the new structure is more fit for purpose. We are also clearer about the objectives of the paper.

- Participants main outcomes and results section of the abstract relate to the main study rather than the development of the survey as its stated objective – we have now addressed this.

- Currently the main outcomes section of the abstract is the conclusion – we have addressed this.

- Article summary should specify that these statements apply to cancer - now done

- The first bullet point in the article summary is not a sentence – fixed

- We’ve now added in references to support our sentence in the background section around ‘Survival differences between populations’

- Line 27, page 7 – primary care, diagnostic and treatment intervals are now defined before they are mentioned.

• Page 7, line 42, routes to diagnosis have an important influence on cancer outcomes – we have addressed this comment.

• Page 8, line 4, reference 12 – we have used a different reference

• Emergency presentations and routes to diagnosis – the reviewer rightly points out that data on emergency presentations doesn't typically make a distinction between patients who have been referred there by their GP and those who present initially to A & E – we have now addressed this point.

• Page 12, routes to diagnosis – we now make it clear that the definitions are different to those used in previous work – they draw strongly from the Aarhus Statement to which we refer.

• Importance of using data from PCPs and STCs and patients – we now emphasise that patient reported data is also important.

• We now include a summary of the items in each questionnaire and the topics covered as well as indicating where readers can access the questionnaires themselves

• Page 15, line 31 – we are now clearer about the sites for the pilot work

• Reliability section – this has now been rewritten and the comments about numbers of patients for test-retest reliability, Kappa statistics and inter-rater reliability have been addressed – as has the comment on why we didn't undertake test-retest reliability for the other two questionnaires:

• I find it worrying that the test-retest reliability is based on only 12 patients. This is a very small number. This should at least be acknowledged as a limitation.
We agree and now mention the number as a limitation in the Discussion section.

• Whilst I accept that statistical significance if not often reported for kappa statistics, given the small sample size it would be interesting to know if the level of agreement was consistent with chance.
That is a good idea and we now give the confidence intervals for the Kappas. Note changes undertaken in reliability section - the confidence intervals for Kappa have been calculated and bias corrected using bootstrapping.

• I am rather confused by the use of a weighted Kappa for the reliability of dates. Kappa statistics are used to quantify agreement in categorical variables, whilst dates are, in effect, continuous variables. Given in this particular case there are only 12 patients, there must be 24 or fewer unique dates. As such it would be possible to apply a kappa (weighted or unweighted) statistic, but I would suggest that this is inappropriate. I would suggest that a classical inter-rater reliability based on a one-way analysis of variance would be more appropriate.
We agree that the dates would be rather unique for this group and Kappa should not be used.
Changes undertaken in Reliability section: the agreement for date of diagnosis has been measured by Lin's concordance correlation coefficient (CCC), which is known to be robust on as few as 10 pairs of data [ref1, ref2]. We have excluded other continuous variables from the test-retest analysis, as less than 10 patients responded to the corresponding questions.

• Page 21, line 20 –checking of questionnaires – we have now removed this. We don't have precise estimates of error rates, they differ between questions, questionnaires and jurisdictions

- Analysis plan section – we have now rewritten this: The analysis plan is vague such that it is no more than a generic thought on how various types of data would be analysed. Nowhere are any outcomes (primary or secondary) defined or any particular scientific questions outlined.

We agree that the analysis plan will benefit from more details, and trust the re-written section is suitable

- Page 22 line 16 – why will prevalence rate ratios be used to look at diagnostic routes rather than logistic (or multinomial logistic) regression as you will be considering the proportion of your sample diagnosed through an emergency route, for example.

The comment is very relevant. Since the output of multinomial logistic regression tends to be hard to interpret, it was decided to dichotomize Diagnostic route into 'screening' and 'non screening', and then for the non-screened (symptomatic) divide the routes into the relevant groups. Generalized linear models for the binomial family will be used to quantify the prevalence ratio for screening among jurisdictions. Prevalence ratio is chosen as an outcome measure over odds ratio, as odds ratio overestimates the association when the prevalence of the outcome measure is above 20% [Barros & Hirakata, 2003]. We have added this clarification to the Analysis plan section.

- Page 22 line 23 – multi-level models are mentioned but with no justification nor with any detail of when they will be used and what the clustering variable will be.

Thank you for the comment. Although it was originally planned, the multi-level analyses will not be performed in this study. We have deleted it from the text of the Analysis-plan section.

- Page 22 line 27 – the comprehensive set of rules could be provided in supplementary material along with a sense of in what circumstances each of the data sources takes precedence over the others.

These rules definitely belong to the public domain. However, we would prefer to publish them together with the specific papers as they would be a little different for each cancer type (e.g. on screening). Further, as the data analysis has not been finished yet, changes may occur. This could prevent different versions of the rules to be used.

- Page 22 line 55 – It is stated that ecological analyses will be undertaken, presumably with each jurisdiction as the unit of analysis. This means that the sample size will be 10, which is very small and as such will have limited power. In fact such studies will have only 80% power to detect a correlation coefficient of 0.79. I would suggest that this is inadequate, particularly if multiple testing is factored in.
- This is a good observation. Our expectation is that correlation coefficient is 0.85, that allows power of about 90% given 10 units of analysis (Machin et al, 2009). We have added to the Analysis-plan section, that we expect a high correlation between survival and time intervals. We also refer to the paper published from ICBP module 3 where similar analyses were conducted (Rose et al, 2015).

- Authors' contribution – we have now given a more detailed description of the contribution of authors to the work, including their roles in ICBPM4.

Reviewer 3

- Comments on abstract have now been addressed

- Availability of the survey instrument – see previous response

- Patients diagnosed as emergencies – we agree that these patients often have circuitous, prolonged pathways and have now included a comment on this

- The paper structure – see previous responses, we have now completely restructured the paper and believe that the reviewer will no longer find it confusing

- Differences in response rates between jurisdictions – we have now discussed this in greater depth

- Were other administration methods explored in the UK? – They were but the mechanism we used was specified by our ethical permissions

- Why did primary care in the UK countries not send what looks to be a significant number of questionnaires? – We have not followed this up and don't have the means to do this now but our experience with primary care research would suggest that in the majority of cases it was simply lack of time and workload

- How do response rates compare with the response rates of other surveys using similar administration strategies? – we've now included a reference to address this.

References
Barros AJ, Hirakata VN. Alternatives for logistic regression in crosssectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Med Res Methodol 2003;3:21.

Machin D, Campbell MJ, Tan SB, Tan SH. Sample size tables for clinical studies 3.ed. 2009

Rose PW, Rubin G, Perera-Salazar R, et al. Explaining variation in cancer survival between 11 jurisdictions in the International Cancer Benchmarking Partnership: a primary care vignette survey. BMJ Open. 2015 May 27;5(5):e007212

VERSION 2 – REVIEW

REVIEWER	Gary Abel University of Exeter, UK
REVIEW RETURNED	02-Feb-2016

GENERAL COMMENTS	The authors have addressed most of my comments on the previous version of the manuscript. Combined with the additions prompted by other reviewers this is a much improved paper. However I still have a few outstanding concerns. Page 12 line 2 – The description of the cancer registries in England is now a historic description. It is my understanding that even at the time of data collection there was officially only a single registry covering the whole of England. This should be corrected or made clear that it is a historical situation. Page 12 line 52 – Whilst I appreciate the fact that the authors have taken on board my suggestion to include a measure of precision of the Kappa estimates, the problem here is that for most measures considered agreement is perfect. With the bootstrapping method employs this means that confidence intervals cannot be created. Whilst it should still be possible to estimate the potential Kappa values which would potentially give rise to perfect agreement 5% of the time or more from first principles, a quick search does not suggest a tool for doing this is in the public domain (it should be noted that I only looked quickly). Assuming you are unable to find a tool and do not wish to calculate this from first principles I wonder if

	a short sentence in this section giving a narrative reason why the Cis could not be calculated, and acknowledging that the precision is still likely to be large (though unknown). Page 12 line 55 – You have applied Lin’s Concordance Correlation Coefficient to the date of diagnosis. Whilst I admit that this was following my suggestion that a more conventional reliability calculation is used, I now think this too is inappropriate – apologies. The reason being that such measures of reliability (such as the ICC or the CCC you have used) put the uncertainty in the context of the true variation between measures. Formally an ICC would be interpreted as the proportion of variance in dates which is explained by the true variation. On this basis the ICC or CCC is a measure of how well different individual dates can be distinguished. So simply by choosing individuals across a large period of time the reliability can be increased. However, in this situation, my impression is that the variation in dates between respondents is actually of little interest in this case, and likewise being confident in who was diagnosed before who is not what you really want to capture with reliability. I wonder if what is simply required is a description of the variation in the differences between the dates in the test and retest surveys. All that said I am surprised that you have interpreted a CCC of 0.829 as poor agreement. Translating the CCC to an ICC (to which it is sometimes equivalent) this would suggest that around 83% of the observed variation in dates is due to the true variation in dates. I would consider this reasonable agreement. Page 17 – Sample size – I am still at a lost as to what your sample size is actually trying to detect. Reference number 38 has a test for detecting proportions outside of that expected when the data are described by an overdispersed binomial distribution. I.e. you would be looking to detect outliers. However, to do this you would have to have an estimate of the amount of overdispersion that is real variation) between jurisdictions. You discuss a minimally important change of 10%, but how is this theorised to occur? Is it nine Jurisdictions at 75% and one at 85%? Or a spread of values? My calculations suggest that the sample size is too low to make pairwise comparisons (i.e. 2 jurisdiction with 200 patients each which would have 71% power to see the theorised difference). It may be that your power calculation is to detect an overall variation between jurisdictions. In which case, it should be noted that pairwise comparisons will have less power, even between the extremes. The bottom line is that I just don’t know what you are trying to detect and this should be made clear. Also I do not understand what is meant by a multinomial sample. Multinomial is a distribution and while data may have that distribution, I don’t see how a sample can. Page 17 – line 55 – The final sentence (which goes over the page) does not make sense. I do not know what the word “regress” means in this context and there is an “and” with nothing obvious following (apart from in brackets) and then an “is” which should be an “are” given the “and”. Also there is an implication that a length of time in days is a count variable (although that is not clear given the sentence as written). I would not consider a number of days to be a count, rather it is a continuous measure which has been rounded. Page 18 line 13 – An odds ratio does not overestimate the strength of association. It may overestimate a prevalence ratio when prevalence is high, but it is a legitimate measure of association in its own right (though different to a prevalence ratio). The sentence

	needs amending accordingly. Page 18 line 35 – Should the word “implies” be replaced with the word “includes”? Page 17 line 37 – You say “We develop rules for …” There should be a “will” or “have” in this statement to indicate whether they have been, are or yet to be, developed. Page 19 line 10 – See above comment on CCC and it’s application to dates. I suggest this may be a reasonable measure for intervals, though some consideration (such as a transform) may be needed to account for the skewed data. Page 19 line 20 – I am not sure why you a-priori expect a high correlation between survival and time intervals when so many other factors may be at play (e.g. treatment which, as you point out on page 6, may be impacting on survival). You suggest a figure of 0.85 in your reply to my comments. This seems hugely optimistic with no previous findings to back this up. It is worth noting that existing similar studies with small sample sizes are likely to suffer from publication bias, in effect overestimating the true association. With regard to response rates I wonder how the (potentially) different processes for dealing with deceased patients might impact the response rates. For example, a registry which is informed very early of deaths may exclude these patients from the sample and thus they would not obtain a non-response, whereas another registry may send more questionnaires to deceased patients who then, by definition, are non-responders. Having said all that being deceased is not listed as an exclusion criteria and so in theory this should not be an issue as all patients should be included, whether alive or not. Supplementary file 6 – It might be nice if it were possible to update these data (and any corresponding text) prior to publication.

REVIEWER	Jessica Sheringham UCL
REVIEW RETURNED	07-Feb-2016

GENERAL COMMENTS	I have reviewed this substantially revised and improved paper. I am content my major comments have been resolved: - The objectives now reflect the content. - The paper structure is clearer, although I would still have found it helpful for it to have followed a conventional journal structure with the development of the data collection instruments in the methods and preliminary response rates in the results. - I note the the survey tool is now included in the appendix. - The discussion of response rates is welcomed, particularly how the study team will investigate and address it.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 2

Page 12 line 2 – The description of the cancer registries in England is now a historic description. It is my understanding that even at the time of data collection there was officially only a single registry covering the whole of England. This should be corrected or made clear that it is a historical situation.

We have done this

Page 12 line 52 – Whilst I appreciate the fact that the authors have taken on board my suggestion to include a measure of precision of the Kappa estimates, the problem here is that for most measures considered agreement is perfect. With the bootstrapping method employs this means that confidence intervals cannot be created. Whilst it should still be possible to estimate the potential Kappa values which would potentially give rise to perfect agreement 5% of the time or more from first principles, a quick search does not suggest a tool for doing this is in the public domain (it should be noted that I only looked quickly). Assuming you are unable to find a tool and do not wish to calculate this from first principles I wonder if a short sentence in this section giving a narrative reason why the Cis could not be calculated, and acknowledging that the precision is still likely to be large (though unknown).

We have taken this suggestion on board – we appreciate the difficulty with Kappa estimates, and agree this more narrative approach is appropriate. See revised 'Reliability testing' section

Page 12 line 55 – You have applied Lin's Concordance Correlation Coefficient to the date of diagnosis. Whilst I admit that this was following my suggestion that a more conventional reliability calculation is used, I now think this too is inappropriate – apologies. The reason being that such measures of reliability (such as the ICC or the CCC you have used) put the uncertainty in the context of the true variation between measures. Formally an ICC would be interpreted as the proportion of variance in dates which is explained by the true variation. On this basis the ICC or CCC is a measure of how well different individual dates can be distinguished. So simply by choosing individuals across a large period of time the reliability can be increased. However, in this situation, my impression is that the variation in dates between respondents is actually of little interest in this case, and likewise being confident in who was diagnosed before who is not what you really want to capture with reliability. I wonder if what is simply required is a description of the variation in the differences between the dates in the test and retest surveys. All that said I am surprised that you have interpreted a CCC of 0.829 as poor agreement. Translating the CCC to an ICC (to which it is sometimes equivalent) this would suggest that around 83% of the observed variation in dates is due to the true variation in dates. I would consider this reasonable agreement.

We no longer use Lin's CCC, and have reduced the reliability analysis for dates to a description of the variation in their differences (also in revised 'Reliability testing' section).

Page 17 – Sample size – I am still at a lost as to what your sample size is actually trying to detect. Reference number 38 has a test for detecting proportions outside of that expected when the data are described by an overdispersed binomial distribution. I.e. you would be looking to detect outliers. However, to do this you would have to have an estimate of the amount of overdispersion that is real variation) between jurisdictions. You discuss a minimally important change of 10%, but how is this theorised to occur? Is it nine Jurisdictions at 75% and one at 85%? Or a spread of values? My calculations suggest that the sample size is too low to make pairwise comparisons (i.e. 2 jurisdiction with 200 patients each which would have 71% power to see the theorised difference). It may be that your power calculation is to detect an overall variation between jurisdictions. In which case, it should be noted that pairwise comparisons will have less power, even between the extremes. The bottom line is that I just don't know what you are trying to detect and this should be made clear. Also I do not understand what is meant by a multinomial sample. Multinomial is a distribution and while data may have that distribution, I don't see how a sample can.

We appreciate we may not have expressed this section as clearly as we'd hoped, so have further modified it. We have emphasized that only 9 comparisons will be performed (between a reference jurisdiction and each of other jurisdictions). Reference 38 has been deleted; in fact it was left over from a previous version and we'd meant to delete it earlier. The changes are highlighted in the revised Sample size section.

Page 17 – line 55 – The final sentence (which goes over the page) does not make sense. I do not know what the word "regress" means in this context and there is an "and" with nothing obvious following (apart from in brackets) and then an "is" which should be an "are" given the "and". Also there is an implication that a length of time in days is a count variable (although that is not clear given the sentence as written). I would not consider a number of days to be a count, rather it is a continuous measure which has been rounded.

We've fixed that final sentence, and have addressed the 'count' vs 'continuous measure' issue. See revised Analysis plan section

Page 18 line 13 – An odds ratio does not overestimate the strength of association. It may overestimate a prevalence ratio when prevalence is high, but it is a legitimate measure of association in its own right (though different to a prevalence ratio). The sentence needs amending accordingly.

Agree – we have done this.

Page 18 line 35 – Should the word "implies" be replaced with the word "includes"?

Done

Page 17 line 37 – You say "We develop rules for …" There should be a "will" or "have" in this statement to indicate whether they have been, are or yet to be, developed.

It now reads 'we have developed rules'

Page 19 line 10 – See above comment on CCC and its application to dates. I suggest this may be a reasonable measure for intervals, though some consideration (such as a transform) may be needed to account for the skewed data.

This is an interesting and thoughtful comment on CCC and its application to dates, which prompted some discussion amongst the authors. While we accept that it is relevant for test-retest reliability of dates (which is under investigation in the Reliability testing section), we believe that CCC can be applied to assess inter-rater agreement, which is a focus of our triangulation analysis, with variation in dates between different responders as a point of interest. Nevertheless, we have added clarification regarding data transformation (all in revised Analysis plan section)

Page 19 line 20 – I am not sure why you a-priori expect a high correlation between survival and time intervals when so many other factors may be at play (e.g. treatment which, as you point out on page 6, may be impacting on survival). You suggest a figure of 0.85 in your reply to my comments. This seems hugely optimistic with no previous findings to back this up. It is worth noting that existing similar studies with small sample sizes are likely to suffer from publication bias, in effect overestimating the true association.

We accept that there are multiple factors at play, and the suggested correlation may be too optimistic. The most prudent approach would seem to withdraw the ecological analysis from the analysis plan – again, you'll find this in the revised Analysis plan.

With regard to response rates I wonder how the (potentially) different processes for dealing with deceased patients might impact the response rates. For example, a registry which is informed very early of deaths may exclude these patients from the sample and thus they would not obtain a non-response, whereas another registry may send more questionnaires to deceased patients who then, by definition, are non-responders. Having said all that being deceased is not listed as an exclusion criteria and so in theory this should not be an issue as all patients should be included, whether alive or not.

Yes – it is possible that differing speeds of death notification might lead to some registries excluding patients who have died, while others would have included them. We struggled with this, as we don't have complete information on this speed of death notification, and couldn't find a form of words to express this caveat. Of course questionnaires sent to patients who had died were ideally returned to us, but we didn't have the mechanisms to measure the completeness of this process. In the end we felt that the effects on response rates were likely to be small, so we didn't mention the issue in the paper.

Supplementary file 6 – It might be nice if it were possible to update these data (and any corresponding text) prior to publication.

Data in the table has been updated to reflect recruitment since June 2015, until end of February 2016.
As indicated, please do let us know if you would like us to respond to any further reviewer comments.
Can we again thank you and the reviewers for the very helpful and constructive feedback we've received on the paper.
Kind Regards,

David Weller (on behalf of the authors of 'An investigation of routes to cancer diagnosis in ten international jurisdictions, as part of the International Cancer Benchmarking Partnership; survey development and implementation')

VERSION 3 - REVIEW

REVIEWER	Gary Abel University of Exeter, UK
REVIEW RETURNED	05-Apr-2016

GENERAL COMMENTS	The authors have largely addressed my comments. I have found one typographical error in the new text plus two remaining comments which the authors may wish to consider. Page 12 line 58 - Rather than "the precision is still likely to be large", this should read "the imprecision is still likely to be large" Sample size - the authors have now clarified that 9 comparisons will be made - each of 9 jurisdictions with a reference jurisdiction. I would refer them to my previous review where I suggested that "2 jurisdiction with 200 patients each which would have 71% power to see the theorised difference". Thus I am confused as to where their numbers come from. Given my confusion I have now located the text book referenced and while I agree that using the referenced method on a 2 by 2 table does suggest 200 patients per jurisdiction I have a concern that this is an approximation better suited to tables with higher degrees of freedom. The method in section 6.5.1 gives a

	better approximation to what might be expected in my opinion. In the analysis section the authors are still suggesting they will analyse a number of days as count data. As I suggested before I do not believe this is appropriate. Count data should refer to independent events (such as the number of incident cancer cases in a region, or the number of attendances at an emergency room). In this sense number of days is not count data. For day 5 to occur days 1 to 4 must previously have occurred (i.e. these are not independent events). Really a number of days is no more count data than height rounded to the nearest metre.

REVIEWER	Jessica Sheringham UCL, England
REVIEW RETURNED	18-Apr-2016

GENERAL COMMENTS	Thank you for the opportunity to review the second revision of this paper. As stated in my comments on the first revision, the paper was substantially revised and improved and addressed my major comments. My comments therefore are very minor: Abstract, page 5 line 39 – In response to decision letter, authors stated they have removed Lin’s concordance correlation coefficient for assessing agreement in the test retest for dates – this is still in the abstract. I appreciate they will retain use of Lin’s CCC to compare agreement between sources of evidence but I think the abstract result still refers to the agreement in the test-retest (unless I’ve misunderstood – in which case, clarify!) line 46 –authors could consider updating “collection will be completed in early 2016” to reflect the fact the paper reports on response rates to Feb 2016 (particularly if data collection has now finished). page 12 line 47 – typo - patient’s should read patients’

VERSION 3 – AUTHOR RESPONSE

Reviewer 2
- Page 12 line 58: We have rewritten this sentence to address reviewer comments.
- Sample size: We are happy that the reviewer now agrees about the principles.
- Analysis section: We thank the reviewer for this information.

Reviewer 3
- Abstract line 39: We thank the reviewer for spotting this, this has now been deleted from the abstract.
- Abstract line 46: Abstract has been amended to reflect the fact that data for breast has been collected, but other cancer types continue and will be finalised by the end of 2016.
- Page 12, line 47: We thank the reviewer for spotting this typo, we have corrected this.