Author’s response to reviews

Title: Weight loss as a predictor of cancer and serious disease in primary care: an ISAC approved CPRD protocol for a retrospective cohort study using routinely collected primary care data.

Authors:

Brian Nicholson (brian.nicholson@phc.ox.ac.uk)
Paul Aveyard (Paul.Aveyard@phc.ox.ac.uk)
Richard Hobbs (Richard.Hobbs@phc.ox.ac.uk)
Margaret Smith (Margaret.Smith@phc.ox.ac.uk)
Alice Fuller (Alice.Fuller@phc.ox.ac.uk)
Rafael Perera (Rafael.Perera@phc.ox.ac.uk)
William Hamilton (W.hamilton@exeter.ac.uk)
Clare Bankhead (Clare.Bankhead@phc.ox.ac.uk)

Version: 1 Date: 05 Jun 2017

Author’s response to reviews:

Dear Editors,

Thank you very much for considering our manuscript for publication in the journal. We apologise as we did not make it clear in the submission that this paper is a protocol for a research project, which has already undergone external peer-review by the Independent Scientific Advisory Committee to the MHRA, in order to obtain approval for its conduct.

We are therefore not in a position to respond to your reviewers’ comments with substantial changes to this approved protocol. We hope we have clarified this by the addition of a sub-clause to the title. To suit journal requirements, we have added an abstract and a short discussion to the protocol and removed statements about dissemination and patient and public involvement.
For your information and the reviewers’ we attach (below) a point by point reply to the reviewers: in many places we look forward to the opportunity to address their queries during the data management and analysis phases of the study.

I’m the meantime, we invite the editors to consider whether they can accept this for publication as a protocol that has already been peer reviewed. We hope you will share our belief in the value of publishing protocols for Prognostic studies.

Yours Sincerely,
Dr Brian Nicholson (on behalf of all authors).

Response to reviewers comments.

Reviewer #1: Weight loss as a predictor of cancer and serious disease in primary care: a retrospective cohort study using routinely collected primary care data.

Thank you for asking me to review this manuscript. It is an interesting topic, and could make an interesting and worthwhile study. There were some issues that needed clarification which are detailed below.

**Author Response - Thank you.

Reviewer #1. Concerns about the proposed study:

Sample size: should be based on potential number of events (ie number of patients with the outcome: cancer/serious disease) not just weight loss recorded

Stratification by cancer, stage and other covariates as planned may be impossible for many cancers due to small numbers in strata / cells

**Author Response - Subsequent work has established that there are 86,651 eligible patients >40yrs with an unexpected weight loss code (preliminary pilot work had suggested there was at least 30,000). This will therefore be the largest primary care cohort study using all eligible codes for unexpected weight loss in the CPRD. We originally calculated that only 2184 patients with
weight loss are required to detect a hazard ratio of 2 at 99% power (0.05% alpha) using an enrolment ratio of 1:5. That is, a change in a cancer risk from a PPV of 1.5% in patients without weight loss to 3% in patients with weight loss. It is anticipated that the study will therefore have sufficient power for stratification by cancer type. We agree that we may not be able to stratify for cancer stage and all other covariates with sufficient numbers remaining in each strata. However, this study aims to understand the association between weight loss and cancer, in as much detail as the data permits.

Reviewer #1. Amount of missing data: on weight loss, on stage of cancer, on other covariates (e.g. smoking/alcohol consumption). Not enough justification given of why these will not scupper the aims by biasing the results completely - those without these records may well be very different, and therefore dealt with very differently by GPs from those who do

**Author Response -** In the protocol, we have chosen to focus on weight loss coding as we appreciate that records of weight measurement are a source of missing data in NHS primary care records, and so weight measurement cannot be relied on as a means of defining weight loss. Cancer stage information is entirely unsatisfactory in CPRD, which is why we have linked to the cancer registry (which will also be incomplete, but less so). Lifestyle covariates are non-essential for our main aim (to determine the predictive value of weight loss for cancer) but we will explore multiple imputation to complete these (and all other relevant missing) variables. For symptoms, we have to assume that the absence of a code equates to the absence of a symptom. This is necessary for two main reasons: 1) GPs will normally only code the presence of a symptom; 2) although GPs may enter information in free text uncoded, CPRD do not allow access to these data. These factors are sources of recording bias, a limitation of primary care electronic database research.

Reviewer #1. Time needed to develop a codelist for serious disease - what is the time frame for the project? Is this part feasible within it?

**Author Response -** This is part of a PhD: there is enough time to generate these lists.

Reviewer #1. Issues in the article: Title: it would be helpful to have an indication in the title that this is a protocol for a planned study, rather than the results of a completed study.

**Author Response -** This is an approved ISAC protocol. We have added “: an ISAC approved CPRD protocol for a retrospective cohort study using routinely collected primary care data.”
Reviewer #1. General. More definitions/descriptions/expansions of abbreviations are needed:

- "unexpected" on first use (Background, line 6): explain how this is defined in this specific context
- NICE (Background, line 26)
- ISAC (Background, line 49): expand abbreviation and reword "this ISAC application" - makes this sound like it was cut and pasted from that application (which I guess it partly was?)
- "acceptable records" (Background, line 50): explain what is meant by this in this context
- Read code (Background, line 56)
- "algorithm" (Aims and rationale, line 33): say more about what form this takes
- Open matched cohort study (Aims and rationale, 2nd page, line 1): worth a bit more description of this study design for anyone unfamiliar
- NCDR (Data Linkage Required, line 56)
- Medcode (Exposures, Outcomes and Covariates, line 27)

**Author Response - We have included an abbreviation list.**

Reviewer #1. Abstract. It would be useful to summarise some of the strengths/limitations in the abstract's Discussion section, not your aim for the findings

**Author Response - The abstract is unique to the D&PR submission and so we have edited the abstract to suit these comments.**

Reviewer #1. Background

It is not clear in this section (beginning line 49) why you did a preliminary search in only patients over 40. It becomes clear later, but worth explaining so the reader doesn't question it here. In fact, this whole paragraph could move to a data section (before Sample size)
Aims and rationale. I would usually expect an overall aim, with specific objectives that would be undertaken to meet that aim (not the other way round) but willing to be overruled on the semantics of this!

Sample size. I would restate the years of diagnosis etc at this stage, or as suggested, move the description of the data to before this. My understanding is that for your sample size calculation you would need to have an estimate of the number of events (ie cancer cases, or serious disease cases) you could expect among those with a weight loss record

**Author Response - This is the wording used in the approved ISAC protocol.

Reviewer #1. Say more about how you know that 60% of cancer cases will be able to be linked? Who are those who are not likely to be linked? Discuss the impact this will have in the Limitations section.


Reviewer #1. Stratifying by the variables you mention will leave you with very small numbers in certain cells: do you have enough data? Where will you get your cancer stage information from? Even from the cancer registry it is likely to be missing in very high proportions for the years before 2012, to varying degrees for different cancers. Same would probably be true for grade, tumour size, histology etc

**Author Response - Please see comments above.

Reviewer #1. Data Linkage Required. I would move this section to before Sample Size as it answers some of the questions I had while reading that section (maybe rename to Data Sources or similar?)

**Author Response -. This is wording used in the approved ISAC protocol.

Reviewer #1. What does the 38-68% refer to (2nd page, line 2): is it the anticipated range of missing data by cancer ie some will be 38% missing, others 68% missing? It should be clear
without having to look up the reference. In fact, I'm not sure how you came to these numbers as that reference is to a paper only about colorectal cancer.

**Author Response - We have removed this reference as it is misleading, and will amend the ISAC protocol.

Reviewer #1. Make it clearer that the "patient level" IMD that you mention as a proxy for SES is also an area level (ecological) measure (not measured for an individual patient, but rather based on their postcode of residence)

**Author Response – This is wording used in the approved ISAC protocol.

Reviewer #1. Selection of comparison group(s) or controls. Could match on more than GP practice? e.g. age/sex of patient, year of diagnosis, age/sex of GP - explain why these not used for matching

**Author Response – We wanted to adjust for age and sex in multivariate analysis without having used these variables to match. We don’t know that the association between weight loss and cancer is the same in all men aged 50 or in all women aged 60. We do know that coding behaviours clusters in GP practices.

Reviewer #1. Exposures, Outcomes and Covariates. Quantitative weight measurements (line 20): presumably need measurements taken at least two time points

**Author Response – We are only describing weight measurement, not using it to quantify weight loss, as we know that contemporaneous weight measurement data is largely missing in NHS primary care records.

Reviewer #1. Does the codelist from Hamilton and colleagues (line 46) have a reference or was it personal communication?

**Author Response – Personal communication.

Reviewer #1. Covariates. Many covariates are recorded poorly e.g. smoking/alcohol/ethnicity: give an indication of completeness and discuss in Limitations section
**Author Response – We will have to describe the completeness of the data in both the descriptive and cohort analyses.

Reviewer #1. Comorbidities gathered from different sources (main GP record/prescribing data): need a careful algorithm for how these will be used/combined as they could give different information

**Author Response – If data from separate CPRD files is combined to create a variable, there will be a clear audit trail for this: as there will be for all the methods developed further during this study.

Reviewer #1. Data/Statistical analysis. I assume the 200 patients will all have been determined to have a weight loss record?

**Author Response – A weight loss code or a weight measurement.

Reviewer #1. How well recorded is the "clinical purpose"? (Address in Limitations section)

**Author Response – This is a developmental section of the protocol. We will generate codelists for clinical purpose during the descriptive section of the study, including codes for: medication review, chronic disease review, health check, antenatal, symptom groups, and diagnostic groups.

Reviewer #1. Cumulative incidence plots/ Cox regression: issue of small number in stratified analyses? Issue of missing data? How will it be dealt with?

**Author Response – We will only perform analysis on sub-strata when numbers permit. We will explore whether multiple imputation can overcome some of the limitation of missing data for variable where this is appropriate, as discussed above.

Reviewer #1. The section on addressing missing data has little detail in it about how this will actually be dealt with

**Author Response – This is wording used in the approved ISAC protocol.
Reviewer #1. - You only mention the missing weight data, but do not deal with the big issue of all the missing covariate data and how that will be addressed.

Limitations. More needs to said about why missing data, potential small number in strata etc will not be insurmountable. I'm not convinced about clustering of similarity of coding practices by GP practice. Perhaps they are trained together for the coding itself, but are GPs in a practice more similar than e.g. GPs of the same gender/age in different practices in their propensity to ask about/measure/code weight loss? If you know this to be the case (from the literature) it would be worth specifying

**Author Response – Please see comments above about missing data.

Reviewer #1. Not sure about the very last paragraph - I understood repeated weight measurements to be part of what you would be using to identify weight loss, but you mention it here as if it is a completely different study.

**Author Response – We are describing weight measurement and weight loss coding, then we are only using weight loss coding as our exposure variable in the cohort analysis.

Reviewer #2: The research questions the authors plan to address are interesting and worthwhile, and I am in principle I strongly supportive of the publication of protocol papers for electronic health record studies. However the manuscript in its current form doesn't provide enough information about certain aspects of the planned work and does not follow the journal Protocol format. Specific comments are listed below:

**Author Response – Thank you.

Reviewer #2. Most of this text appears to have been copied directly from the original ISAC application (see line 49, page 3 for concrete example) and does not follow the journal-specified format for a Protocol.

**Author Response – This is an approved ISAC protocol, reformatted to suit the journals formatting requirements as much as possible.
Reviewer #2. Related to this, under the heading "Data Linkage Required" the authors list requested linkages, but it is unclear whether these have been approved and will form part of the data set. There are many more instances, but I do not list them all here.

**Author Response – This is an approved ISAC protocol.

Reviewer #2. I note that the authors are using a validated set of readcodes for their cancer endpoint (please insert the appropriate reference for this as it is currently missing).

**Author Response – There is no reference for this list. The codelists have been developed through collaboration with other researchers working in this field, validated by local review of the codes chosen by searching the CPRD medcode library.

Reviewer #2. However, it appears that they propose the development of an EHR algorithm for weight loss - what steps do the authors propose to take in order to validate the algorithm developed?

**Author Response – We will develop this algorithm using a small subset of clinical records, and then cross-validate is using the remaining dataset.

Reviewer #2. If I have understood correctly, the authors also appear to be proposed the development of another algorithm for what they have termed "Serious Disease", but it was not clear what conditions might be included in the endpoint, or how this algorithm might be validated.

**Author Response – This serious disease codelists are being developed by literature review to identify likely candidate conditions that are known to cause weight loss.

Reviewer #2. It was also not clear to me which conditions might be included under "comorbidity" covariates.

**Author Response – Comorbidity codelists are also being developed through literature review to identify the most prevalent 20 comorbidites in English primary care patients.

Reviewer #2. The cohort definition was somewhat unclear given that different time windows are proposed for different variables. I appreciate that extracting relevant information from EHR is
complex and there will be rationales for the authors decisions, but these are currently hard to follow. A clear statement of inclusion/exclusion, cohort entry, outcome/exit definitions is needed, as well as the time windows used for each set of variables - a diagram may be helpful here?

**Author Response – This is wording used in the approved ISAC protocol.**