Reviewer’s report

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.

Version: 0 Date: 07 Apr 2017

Reviewer: Melanie Morris

Reviewer's report:

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.

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

* 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

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

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)

Abstract

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

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

* 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.

* 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

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?)

* 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.

* 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)
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

Exposures, Outcomes and Covariates

* Quantitative weight measurements (line 20): presumably need measurements taken at at least two time points?
* Does the codelist from Hamilton and colleagues (line 46) have a reference or was it personal communication?

Covariates

* Many covariates are recorded poorly e.g. smoking/alcohol/ethnicity: give an indication of completeness and discuss in Limitations section
* 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

Data/Statistical analysis

* I assume the 200 patients will all have been determined to have a weight loss record?
* How well recorded is the "clinical purpose"? (Address in Limitations section)
* Cumulative incidence plots/ Cox regression: issue of small number in stratified analyses? Issue of missing data? How will it be dealt with?
  - The section on addressing missing data has little detail in it about how this will actually be dealt with
  - 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

* 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.

Level of interest
Please indicate how interesting you found the manuscript:

An article whose findings are important to those with closely related research interests

Quality of written English
Please indicate the quality of language in the manuscript:

Acceptable

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?
If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal