Reviewer’s report

Title:A novel method for studying the temporal relationship between type 2 diabetes mellitus and cancer using the electronic medical record

Version:2 Date:2 January 2014

Reviewer:Emma Eggleston

Reviewer’s report:

Comments for authors:
Strengths of this manuscript, which describes an electronic medical record (EMR) based algorithm for the detection of type 2 diabetes and correlation with cancer diagnosis, include large sample size, approach to defining diabetes and cancer diagnoses in the same cohort, and attention to timing of diabetes onset in algorithm development. There appears to be a lot of potential for meaningful findings with the ongoing work in this cohort. However there is a lack of clarity in both research question and conceptual framework for the work that makes the purpose, and the intended future application, somewhat difficult to understand. In addition, the algorithm is not validated, does not clearly distinguish type 1 from type 2 diabetes, or address onset or duration of pre-diabetes, the phase argued in the manuscript to be of critical importance. Recommended revisions are detailed below.

Discretionary Revisions

1) Abstract, results sentences: If this is a methods paper about the development of the diabetes/cancer detection algorithm, results should cover what algorithm found: for example, proportion of type 1 vs. type 2 diabetes, average time between onset of diabetes by laboratory and diagnosis of diabetes, or proportion of patients with diabetes who also had a cancer diagnosis etc.

2) Background: There is a body of recent and exciting literature on the pathophysiologic links, and clinical links, between malignancy and diabetes – recommend including more of these to strengthen the background and argument for importance of the algorithm.

3) Replace “diabetics” and “diabetic subjects” throughout the document with “subjects with diabetes” or “cohort with diabetes”.

4) A table with data on the pre-diabetes, e.g. potentially hyperinsulinemic, stage it is argued the algorithm captures would be very useful and informative. Specifically proportion with IGT, IFG, or IGT/IFG and timing in relation to a) diabetes onset and b) cancer onset.

Minor Essential Revisions

1) Abstract: Conclusion sentence: “and is unique in that it allows for examination of the
hyperinsulinemia-dominant pre-diabetes phase in addition to the diabetes phase, which is
characterized by progressive hyperglycemia” – This overstates study ramifications

- There is no measurement of hyperinsulinemia in the study or correlation between timing of diabetes onset and endogenous insulin status. This statement is fine as a speculation in the Discussion section, but should not be included in the abstract conclusions as it overstates study scope.
- Similarly there is no identification or reporting of pre-diabetes (proportion of pts with IFG, IGT, or both by glycemic testing in the time prior to onset of DM, rates of progression in whom and over what time frame) in the study – again an overreach of study scope.

2) Background – 3rd and 4’th sentence: “Electronic medical records (EMRs) have served as an important data source in initial studies of the relationship between cancer and diabetes, but such studies are limited by imprecision in capture of diabetes diagnosis date, inaccuracies in electronic data, difficulty in distinguishing between type 1 and type 2 diabetes, and biases inherent to retrospective and observational studies. Due in part to these limitations, the temporal and causal relationship between diabetes and cancer, if any, remains almost entirely unexplored. To address this issue…”

These sentences both under and over state the potential of the study to address the limitations specified and need to be amended to more accurately reflect what gap in knowledge the study will address: There is precision in the existing literature re: diagnosis date -- the current algorithm aims to distinguish between date of diabetes onset and date of diabetes diagnosis (a strength that could be set up by the limitations of current data sentence), but it does not measure or address inaccuracies in EMR data, does not distinguish well between type 1 and type 2 diabetes (or if it does, we are not given the validation data to establish this) and is itself a retrospective observational study.

The temporal and causal relationship between diabetes and neoplasia is an active area of research and far from unexplored, but even were it not, this study is not addressing causality, it is describing an algorithm that can be used to make inferences about the temporal association between diabetes onset and cancer.

3) Background, paragraph 2: – more citations are needed, several declarative sentences are without them: sentences 3,4, 6 and 7.

4) Background, paragraph 4: “Despite the powerful effects of insulin as a growth factor and the potential for hyperinsulinemia to impact cancer development, little attention has been paid to the pre-diabetes phase and the temporal relationship between the two diseases remains largely unexplored.”

See, and consider citing, the several studies from the Nurses Health Study (breast cancer and insulin levels, diabetes) and other prospective cohorts (for example Wolpin 2013, JNCI – pancreatic cancer and prediagnostic insulin
levels), prostate cancer and insulinemia, or the Paris Prospective cohort which looks at hepatic cancers.

The present algorithm and findings in the manuscript do not measure or report on pre-diabetes.

Methods:
1) Are glucose values in the EHR system labeled as “fasting” or “random”? The CPT code for fasting and random glucose is the same, so unless it is entered as such by the lab one cannot determine from CPT code alone, and this has significant ramifications for diagnosis of “diabetes”. Please specify if known fastings/random.

2) Clarify purpose of measure of “glucose control”. Substantial work clearly went into an approach to predicting A1C based on glycemic values, however others have previously published similar approaches and it is unclear why this is done for this manuscript – which details an algorithm for detection of type 2 diabetes. Glucose control is not part of the algorithm – if meaning glucose at onset, would not term it “control” as if onset there is no control yet – more importantly one can have a normal A1c but abnormal FPG or RPG at onset with potentially relevant implications for neoplasia. If glucose control will be measured for the future EHR-based studies addressing relationship between diabetes and cancer the authors are undertaking, and this is an additional methods piece for that future work – state it upfront in the abstract and background section and make it more clear in the results and discussion, or remove it from the manuscript

3) The glycemic criteria for diagnosis of diabetes changed between 1995 and 2009 – how was this handled in the analysis? This could also potentially affect the non-diabetes group.

4) What proportion of patients in the diabetes group without a diagnosis code had an abnormal test prior to the normal test that set the 3 year time frame? Were they different than others?

5) Identification of type 2 vs. type 1 diabetes is poor via ICD-9 codes alone (Klompas 2013) – how else did the algorithm distinguish between the two other than requirement that type 1 code occur at least 1 yr after type 2 code ? If there were not further criteria, and algorithm validation is not presented, then language needs to be changed to “diabetes algorithm” not type 2 diabetes. As this has very important implications for insulinemic state and link with neoplasia, this limitation will need to be addressed in the discussion.

6) Methods, subject selection, par 2, sentence 2: “Prior to algorithm application extensive manual exam of the EMR was performed” Does this mean chart review – if so what was the sample size, was it a random sample? What did it find?

7) A1C was not recommended by the ADA for the diagnosis of diabetes until 2010, a year after the cohort ending date – this should be addressed re implications for results, analysis and potential bias. Similarly, the accuracy and
reproducibility of the A1C has changed dramatically since 1995.

8) Results, paragraph 1 sentence 2: “Of note, application of our algorithm including laboratory parameters to the pool of potential diabetic subjects resulted in exclusion of approximately 70% of patients…” – This is a surprisingly high proportion. Report the most common reasons for exclusions found on chart review (labs outside the system? Misapplication of DM code? Transient hyperglycemia – (TPN, steroids etc); diagnosis codes but no abn lab values; vice-versa?) and comment on implications of this – does it match the population level prevalence of diabetes that would be expected for the region or are there subjects with actual diabetes that are potentially being excluded?

9) Results, paragraph 2, sentence 2: “Mean observation in the EMR was similar for both groups with approximately 13 – 14 years before the reference date and 5 – 7 years after the reference date” But in the Discussion it states “with over 16 years of follow-up after diabetes onset and 6 – 7 years of observation before… “ Which is it?

10) Discussion Limitations section: Important limitations to be mentioned include
   a) Glycemic values or diagnosis codes made in other systems or by other providers, a very common EMR based data flaw, which could dramatically impact determination of diabetes onset and diagnosis.
   b) Confounding by indication – providers are more likely to draw glycemic or labs in patients in whom they have concern for risk of diabetes – pts who may also have risk for cancer. As the “non-diabetes” group was defined based on only 1 or more (not 2 or more) normal glycemic labs, this is a potential concern. This potential for confounding by indication is increased by the higher percentage of visits to clinicians pre – diabetes onset in the diabetes group.
   c) Potential impact of change in diabetes diagnostic thresholds and methods (A1C) on findings.
   d) “Potential for additional confounding as a result of selection bias will be minimized via statistical adjustment.” Specify how these will be adjusted – and that the reference is to future work – if applicable to current results in the methods section, adjust for them and put in the methods and results – not the discussion section.

Major Compulsory Revisions

1) Prospective validation of the diabetes algorithm is needed. If this is not possible, a second but less rigorous, option is validation of the algorithm in a random sample of subjects from 2010-2013 (years not included in the current cohort) with calculation of the positive predictive value for type 1 and type 2 diabetes, and predictive value for date of onset of diabetes.

2) “Potential non-diabetic subjects with no normal glucose or HbA1c test, as defined by the ADA, … were excluded” Per Figure 1 this did not exclude pre-diabetes level values. If this is the case, the analyses needs to be re-rerun
with the exclusion of the prediabetes level values in the non-diabetes group as the primary stated goal of the algorithm is identification of the hyperinsulinemic prediabetes phase.

3) Report algorithm findings – proportion with type 1 and type 2 diabetes, average time between onset of diabetes and diagnosis of diabetes, proportion of patients with abnormal FPG, vs., RPG, Vs A1C vs. combination, average time between onset of diabetes onset and cancer diagnosis – overall and by cancer type.

4) Report proportion of diabetes subjects with a pre-diabetes range value as their “normal” value and proportion with values suggestive of true normoglycemia (eg <6.5%, FPG<100, RPG<140). Did rates of cancer incidence differ between these groups?

5) Results, paragraph 3: Crude adjusted cancer rates - As the stated aim of the study is to allow assessment of cancer incidence pre and post diabetes onset, give RR for total cancer rates and rates by cancer in the diabetes and non-diabetes groups so readers don’t have to calculate on their own from absolute rates.

6) Results, paragraph 3: Crude adjusted cancer rates are minimally informative on their own in light of the potential confounders the authors outline in the background and collect data for on their subjects. Either do not report them or provide adjusted estimates – adjusting for BMI, clinic visits, reference date range, smoking status, age, MESA residency, and selected co-morbidities.

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:

I declare that I have no compelling interests.