Peer Review File

Article information: https://dx.doi.org/10.21037/ace-22-4

Reviewer A

This paper describes essential work carried out to validate CPRD Aurum cancer diagnoses. Although many researchers have switched from CPRD Gold to Aurum recently, much of it hasn't been validated yet. My concerns are therefore not regarding the purpose of the study, but regarding the methods used.

Comment 1:
My main concern is the use of HES admitted patient care data as the comparison. Generally, for cancer diagnosis information, cancer registry is seen as the gold standard. However, even if for whatever reason (which doesn't seem to be explained in the paper) HES was considered more suitable, it would have been better to include all parts of the HES dataset. Not all cancer patients will be admitted, and the first date of diagnosis may be somewhat off if they are admitted at some point during their treatment (which would explain the 33% which were recorded more than 90 days later (line 174)). Please explain why HES admitted patient care was chosen as the best source to compare CPRD Aurum with, and how this compares to cancer registry and potentially a full HES dataset.

Reply 1:
The objective of this study was not to provide unbiased estimates of the validity of specific cancer diagnoses in CPRD Aurum; rather, the objective of this study was an initial assessment of the quality and completeness of data contained within CPRD Aurum. The patient population was selected to have both CPRD Aurum and HES APC records and was used to evaluate multiple disease outcomes (pulmonary embolism, myocardial infarction), not just cancer. HES APC was selected as an appropriate external data source because all outcomes of interest, including most cancers, would be expected to be reliably recorded in in-patient and out-patient hospital settings and hospitals are expected to send discharge letters to the GP summarizing the care received in-hospital. We recognize that not all cancer patients would receive in-hospital patient care. We addressed this in two ways. First, we list this as a limitation to our study (Discussion, paragraph 5, page 14-15, lines 280-290). We also described the presence of supporting clinical codes that in CPRD Aurum and provide evidence that there was a cancer diagnosis, and that the patient was receiving diagnostic tests, procedures, or treatments to care for their cancer (Figure 1).

We acknowledge that recording of cancer diagnoses is expected to be better in the Cancer Registry and we cite this as a limitation in the discussion (Discussion, paragraph 5, page 14, lines 290-294). Despite the benefits of using the cancer registry, however, availability and coverage of Cancer Registry data can present challenges. For example, at the time the data for this study was acquired, the

Cancer Registry data was available through 2016, which would limit the ability to evaluate current data recording practices. In addition, the wait time to access Cancer Registry data can be long (made longer by the UK's shift of focus to Covid-19 data, currently estimated to be 12-18 months to receive Cancer Registry data). We determined that these practical concerns created more challenges than benefits, and, as the focus of our study was a broad look at data recording of multiple outcomes in CPRD Aurum compared to HES APC, we chose to not pursue comparison to the Cancer Registry for this project. In our discussion, we state that additional validation work comparing Aurum to the Cancer Registry is warranted.

Since the completion of this study, the authors at Boston Collaborative Drug Surveillance Program have begun work to evaluate the presence of breast cancer in CPRD Aurum compared to HES APC, HES outpatient, and the Cancer Registry, which will provide insight into the added value of including each data source.

Changes in text: Edited the discussion of limitations (Discussion, paragraph 5, page 13-14, lines 277-294). to further clarify the limitations based on the data sources used: "In this study, we required all CPRD Aurum patients selected for this random sample to have at least one admission for any reason in HES APC. This was necessary to have two data sources to compare. HES, in general, is not a perfect reference standard because coders may be non-clinical staff and there may be non-specific coding of some hospital events. In addition, some cancer events may be treated in outpatient hospital settings or non-NHS facility where some patients with private insurance may have opted for care elsewhere. We did not evaluate HES outpatient data in this study; therefore, correctness may be underestimated, particularly for cancers treated solely in outpatient hospital or consultant settings. However, it is important to note that, unlike HES APC, it is not mandatory for diagnostic information to be recorded using ICD-10 codes in HES outpatient data and diagnostic information is captured in less than 5% of all attendances (18); therefore, the additional diagnostic information that could be provided by including HES outpatient data is likely to be small. It is also important to keep in mind that the goal of this study was to assess the quality of diagnosis recordings present in the CPRD Aurum data source in terms of completeness and correctness as compared to another easily accessible external data source (data source agreement assessment), not to undertake a validation of cancer diagnoses in CPRD Aurum per se. Cancer stage information is not available in CPRD Aurum or HES APC; therefore, we cannot assess differences in recording practices by cancer stage. Formal validation studies are still needed to assess the validity of cancer outcomes, including studies comparing CPRD Aurum to the Cancer Registry."

Comment 2:
My second question is regarding the patient selection. If I understand correctly, 50000 patients were randomly selected from any available patients from April 1, 1997, through to December 31, 2017. From this selection patients were included

who were present in both HES and CPRD. It is unclear why the selection was done this way. Why was the initial sample 50000 patients? Why not include all patients with a cancer diagnosis in either CPRD or HES instead of creating the 50000 sample first, or if the resulting number is too high, restrict it to a number of years? The current included number of cancer diagnoses could have been much higher, without actually increasing data costs or data preparation time by much. I assume there are good reasons for selecting patients in this manner. It would be helpful if these were explained in the paper. It would also be helpful if patient selection could be explained a little bit clearer in the abstract.

Reply 2:
The objective of this study was not to provide unbiased estimates of the validity of specific cancer diagnoses in CPRD Aurum; rather, the objective of this study was an initial assessment of the quality and completeness of data contained within CPRD Aurum. The CPRD restricted the random sample to 50,000 patients, which was used for multiple validation assessments of other outcomes (references 9-11). We conducted a data source agreement assessment as described by Weiskopf and Weng comparing CPRD Aurum to HES APC. As described in the Methods (Page 3, Paragraph 3, lines 97-103), we required that patients in the 50,000-patient sample had at least one HES admission for any reason to ensure that there were available HES records for comparison with CPRD Aurum. For a hypothesis testing or disease epidemiology study, such a restriction could introduce bias; however, this restriction was necessary for assessing the correctness and completeness of the data and to fulfill our stated objective.

Changes in text:

- Edited Abstract (Abstract/methods, page 2, lines 39-41): “Methods: The source population was a 50,000 random sample of CPRD Aurum patients with HES linkage (1997–2017). Patients with cancer diagnoses recorded in either data source were selected for these analyses. Correctness was the proportion of patients with cancer recorded in CPRD Aurum with a concordant cancer diagnosis recorded in HES. Completeness was the proportion of patients with cancer recorded in HES with a concordant diagnosis in CPRD Aurum.”
- Edited Introduction (paragraph 2, page 4, lines 66-70): “This study uses the same patient population as prior assessments to describe data source agreement on the presence of malignant cancer diagnoses recorded in CPRD Aurum primary care data compared with Hospital Episode Statistics (HES) Admitted Patient Care (APC) data, which captures cancer diagnoses in hospital settings”

Comment 3:
Although I appreciate the additional information of supporting clinical codes for completeness, this would further complicate data preparation and would make

the use of other linked data more attractive. It would be very useful to have information on cancer stage though. Is this something that could be added, and if not, why was decided not to include this?

Reply 3:
The purpose of including supporting clinical codes in this evaluation was to assess the presence of information on patient cancer care as recorded by the GP. The presence of supporting clinical codes provides evidence that the GP had recorded diagnostic information, referrals or visits to specialists, cancer treatments (e.g. drug prescriptions like tamoxifen), cancer care, or palliative care that suggest that the patient was a "true" case and provides rich information on the broad spectrum of care received beyond that which is available in HES APC.

Data on cancer stage is not available in CPRD Aurum or HES APC data. Therefore, we were unable to evaluate quality and completeness of cancer stage information in this study. We have updated the discussion with this information.

Changes in text: We edited the limitations to state that stage information is not available in CPRD Aurum or HES APC data (Discussion, paragraph 5, page 15, lines 290-292): "Cancer stage information is not available in CPRD Aurum or HES APC; therefore, we cannot assess differences in recording practices by cancer stage."

Comment 4:
Line 55 "which captures cancer diagnoses in hospital settings" is probably a bit of an exaggeration seeing as it will capture that only if patients are admitted to hospital, and cancer is one of the main diagnoses the treatment is focussed on.

Reply 4: HES APC captures data that represent the most complete set of diagnosis and procedures provided in the hospital setting. The relevant references were in the methods section, but we edited this statement and renumbered the references to include them in the introduction.

Changes in text: We edited the Introduction (paragraph 2, page 4, lines 66-70) to read "This study uses the same patient population as prior assessments to describe data source agreement on the presence of malignant cancer diagnoses recorded in CPRD Aurum primary care data compared with Hospital Episode Statistics (HES) Admitted Patient Care (APC) data, which captures the most complete set of diagnosis and procedures provided in the hospital setting (13, 14)."

Comment 5:
Line 56 " because cancer often requires treatment in-hospital", do you have a reference to support that statement?

Reply 5: HES APC data represent the most complete set of diagnosis and procedures provided in the hospital setting (see references 13, 14). We edited the sentence to reflect the broad range of information contained in HES APC.

Changes in text: We edited Introduction (paragraph 2, page 4, lines 66-70) "This study uses the same patient population as prior assessments to describe data source agreement on the presence of malignant cancer diagnoses recorded in CPRD Aurum primary care data compared with Hospital Episode Statistics (HES) Admitted Patient Care (APC) data, which captures the most complete set of diagnosis and procedures provided in the hospital setting (13, 14)."

Comment 6:
Line 92 " the patient's 20th birthday", Why only include from 20 years old instead of 18?

Reply 6: The objective of this study was to assess the quality and completeness of cancer diagnoses recorded in CPRD Aurum compared to an external data comparator, not to conduct a study on cancer. There is unlikely to be a material difference in how data are recorded for patients who are 18 versus 20 years of age.

Changes in text: none

Comment 7:
Line 98 "The start and end of each patient's active CPRD Aurum electronic record were estimated using available registration, prescription, and clinical data.", How did you estimate this?

Reply 7: CPRD Aurum does not have a universal date of first registration for all patients and may include diagnoses of patient history captured up to decades before the start of electronic records as well as data migrated from a prior GP or software platform. Thus, for many analyses researchers using CPRD Aurum must define a "start" and "end" date that can be universally applied across all patients. The algorithm we applied to estimate these dates cannot be briefly described beyond what we have already written in the manuscript, however we are willing to share our methods with interested researchers (see supplement 1), but the inclusion of these details in the text would not add meaningful information to the current manuscript. We have included supplement 1 in the resubmission documents, which could be published if desired by the Editorial board.

Changes in text: Methods/study population, paragraph 2, page 6, lines 105-7) has been edited to read "The start and end of each patient's active CPRD Aurum electronic record were estimated using available registration, prescription, and clinical data (Supplement 1: Start/End)."

Comment 8:
Line 100 " Patient's cohort entry date was defined as April 1, 1997, (start of HES data) or their estimated CPRD Aurum record start date, whichever came later", Does that mean that not everyone was included for the full period between April 1, 1997 and December 31, 2017? If that is the case, you may want to clarify line 97 "From the source population, we selected patients who were present in both the CPRD 97 Aurum and HES APC data from April 1, 1997, through December 31, 2017 (time frame when data from both sources was present)."

Reply 8: Thank you for pointing out an area where we can improve clarity of the study methods. We have edited the text as follows "The study period was April 1, 1997, through December 31, 2017 (time frame when data from both sources were present). The start and end of each patient's active CPRD Aurum electronic record were estimated using available registration, prescription, and clinical data (Supplement 1). Patient's cohort entry date was defined as April 1, 1997, (start of HES data) or their estimated CPRD Aurum record start date, whichever came later. The end of follow-up was defined as first of the patient's estimated CPRD Aurum end date, death date, or December 31, 2017 (end of HES data). We excluded patients whose CPRD Aurum and HES APC record did not overlap or who did not have a valid birth date. We also excluded patients with a record of a prior cancer diagnosis in either data source before cohort entry because recording of cancer may vary based on prior cancer history in either data source".

Changes in text: Methods/study population, page 6, lines 104-114) was edited as follows: "The study period was April 1, 1997, through December 31, 2017 (time frame when data from both sources was present). The start and end of each patient's active CPRD Aurum electronic record were estimated using available registration, prescription, and clinical data. Patient's cohort entry date was defined as April 1, 1997, (start of HES data) or their estimated CPRD Aurum record start date, whichever came later. The end of follow-up was defined as first of the patient's estimated CPRD Aurum end date, death date, or December 31, 2017 (end of HES data). We excluded patients whose CPRD Aurum and HES APC record did not overlap or who did not have a valid birth date. We also excluded patients with a record of a prior cancer diagnosis in either data source before cohort entry because recording of cancer may vary based on prior cancer history in either data source."

Comment 9:
Line 110, I was wondering how you verified the cancer diagnosis code lists for Aurum? I believe the usual procedure is to have it verified by clinicians after an extensive search through the medical dictionary. Was this the case?

Reply 9: As a reminder, the objective of this study was to evaluate information recorded in CPRD Aurum against an external data source, not to validate cancer diagnoses for a specific study. We used a combination of key word searches, code mappings from existing cancer code lists, and review of patient records to check the completeness of code lists. To align coding systems between the two data sources, we organized CPRD Aurum codes to match ICD-10 neoplasm groupings at specified cancer sites (Supplement 2).

Changes in text: This supplement had been renumbered as Supplement 2: codes (methods/cancer diagnosis and assessments, page 6, line 116).

Comment 10:
Line 181, where correctness wasn't as high, were there perhaps other similar cancers recorded instead or was there no recording of cancer at all?

Reply 10: Figure 1 summarizes the reasons explaining why a patient had a cancer diagnosis recorded in CPRD Aurum but did not have a concordant diagnosis recorded in HES APC. These reasons included presence of a different cancer code recorded in HES (e.g. secondary, benign, in situ, neoplasm of unspecified or uncertain behavior), codes for diagnostic procedures or treatments recorded in HES in the absence of a cancer diagnosis, or other coding discrepancies.

Changes in text: none

Comment 11:
Line 200-214, the most likely reason for extensive recording of a cancer in CPRD but not in HES admitted patient care surely must be that this was a cancer that didn't result in admission to hospital. There is no reason to believe it wasn't treated in hospital though. Outpatient data would have been very useful here. For now, the lines suggesting that it was managed by just the GP seems somewhat presumptuous.

Reply 11:
In the discussion, we acknowledged as a limitation that not using HES outpatient data may underestimate the correctness estimate reported in this study. However, it is important to note that per the HES Outpatient data documentation (unlike HES APC data), it is not mandatory for diagnostic information to be recorded in HES OP, where diagnostic information is captured in less than 5% of all attendances (18). Therefore, the additional information that could be obtained by including HES outpatient data is likely to be small and should not materially impact our reported correctness estimate. We have edited the limitation section to provide further context on this for the readers of our paper. We are currently working on a study evaluating the recording of breast cancer coding in CPRD Aurum, HES APC, HES OP, and the Cancer Registry which will provide insight into how much, if any, additional diagnostic information is present in HES OP data compared with that recorded in HES APC.

It is also important to note that in the results (page 10, line 192, and Figure 1) we provide reasons for the presence of a cancer diagnosis record in CPRD Aurum, but not in HES APC. Most patients (68.1%) had codes in their CPRD Aurum record that indicated that cancer was under evaluation or cared for by the GP or specialists ("supporting clinical codes", which includes codes for suspected cancer, cancer care, referrals or visits to cancer specialists, and palliative care). In addition, 2.3% had a cancer diagnosis in CPRD Aurum near the beginning or end of the patient's follow-up and therefore they may have been admitted to hospital at a time outside the study follow-up period which could explain why they were not captured in the results. Finally, 20.1% appeared to have coding discrepancies where the diagnostic code recorded in HES APC (e.g. secondary, benign, in situ, neoplasm of unspecified or uncertain behavior) differed from the cancer code recorded in CPRD Aurum, or where the diagnostic procedure or treatment was recorded in HES APC in the absence of a cancer diagnosis code. Our results suggest that there are other potential reasons for the discrepant coding beyond not using the HES outpatient data.

Changes in text: We have edited the limitation section of the discussion (Discussion, paragraph 5, page 14, lines 290-294) as follows: "In this study, we required all CPRD Aurum patients selected for this random sample to have at least one admission for any reason in HES APC. This was necessary to have two data sources to compare. HES, in general, is not a perfect reference standard because coders may be non-clinical staff and there may be non-specific coding of some hospital events. In addition, some cancer events may be treated in outpatient hospital settings or non-NHS facility where some patients with private insurance may have opted for care elsewhere. We did not evaluate HES outpatient data in this study; therefore, correctness may be underestimated, particularly for cancers treated solely in outpatient hospital or consultant settings. However, it is important to note that, unlike HES APC, it is not mandatory for diagnostic information to be recorded using ICD-10 codes in HES outpatient data and diagnostic information is captured in less than 5% of all attendances (18); therefore, the additional diagnostic information that could be provided by including HES outpatient data is likely to be small. It is also important to keep in mind that the goal of this study was to assess the quality of diagnosis recordings present in the CPRD Aurum data source, not to estimate unbiased measures of sensitivity and specificity. Cancer stage information is not available in CPRD Aurum or HES APC; therefore, we cannot assess differences in recording practices by cancer stage. Formal validation studies are still needed to assess the validity of cancer outcomes, including studies comparing CPRD Aurum to the Cancer Registry."

Comment 12:
Line 248, I may have missed it, but I believed correctness to be about 88%, so how does this match the over 95% correctness stated as necessary for most

observational research?

Reply 12: In the results, we stated "Overall, 96.1% of the 3,864 patients with a cancer diagnosis at a specified site recorded in the CPRD Aurum sample had a concordant cancer diagnosis coded in HES APC (87.7%) or had a cancer diagnosis plus presence of supporting clinical codes recorded in CPRD Aurum indicating the cancer was cared for by a GP or specialist outside a hospital setting (8.4%)." We have edited the discussion to clarify and match the results section.

Changes in text: We have edited the Discussion (paragraph 1, page 12, line 238-243) to state "The results of this study indicate that cancer diagnoses recorded in CPRD Aurum, where present, are of sufficient quality for most observational research. Throughout the study period (1997–2017), 87.7% of cancer diagnoses at a specified site recorded in CPRD Aurum were concordant with HES APC, while an additional 8.4% had a cancer diagnosis plus presence of supporting clinical codes recorded in CPRD Aurum indicating the cancer was cared for by a GP or specialist (correctness)."

Reviewer B

Comment 1:

The paper is clearly written, and the description of the study, the results and the discussion are comprehensive. I read it with interest and will find it useful when planning future studies. I am, however, rather disappointed that the authors used only HES APC data. In terms of cost, a CPRD licence covering HES APC data also includes ONS mortality data, which is commonly used with HES APC for identifying cases or outcomes. It could also be useful to know whether adding outpatient data for additional cost would be worth the outlay. Cancer Register data would be a 'gold standard' additional source, but I understand that exploration of this might be outside the budget of the project.

Reply 1: Please refer to our responses to Reviewer A, Comment 1 and Comment 11 for detailed explanations about the choice to use HES APC in this study.

Changes in text: We have edited the Discussion (paragraph 5, page 14, lines 290-294)as follows: "In this study, we required all CPRD Aurum patients selected for this random sample to have at least one admission for any reason in HES APC. This was necessary to have two data sources to compare. HES, in general, is not a perfect reference standard because coders may be non-clinical staff and there may be non-specific coding of some hospital events. In addition, some cancer events may be treated in outpatient hospital settings or non-NHS facility where some patients with private insurance may have opted for care elsewhere. We did not evaluate HES outpatient data in this study; therefore, correctness may be underestimated, particularly for cancers treated solely in outpatient hospital or consultant settings. However, it is important to note that, unlike HES APC, it is not mandatory for diagnostic information to be recorded using ICD-10 codes in HES outpatient data and diagnostic information is captured in less than 5% of all attendances (18); therefore, the additional diagnostic information that could be provided by including HES outpatient data is likely to be small. It is also important to keep in mind that the goal of this study was to assess the quality of diagnosis recordings present in the CPRD Aurum data source, not to estimate unbiased measures of sensitivity and specificity. Cancer stage information is not available in CPRD Aurum or HES APC; therefore, we cannot assess differences in recording practices by cancer stage. Formal validation studies are still needed to assess the validity of cancer outcomes, including studies comparing CPRD Aurum to the Cancer Registry."

Reviewer C

This is a highly relevant and well-written study that will be useful in guiding and providing context for future cancer research in CPRD. The openness in sharing the disease codes used is appreciated.

Comment 1:
Some minor comments:
It would be interesting to know what proportion of patients in the source population of AURUM had no admission in HES (despite linkage) so that researchers have an idea of what would be the implication for the completeness of AURUM data in a study where the restriction to patients with active HES records is not imposed. If the proportion was large enough, an analysis similar to the one presented of the AURUM population without this restriction would be of interest.

Reply 1: The reviewer's point is well taken. Given the study population selection criteria, we are unable to provide this information. CPRD can help researchers with these kinds of feasibility inquiries on a study-by-study basis.

Changes in text: none

Comment 2:
In the discussion it may be worth mentioning more explicitly that completeness in AURUM is very low (<30%) for cancers cared mostly in a hospital setting. AURUM is essentially useless for these cancers if HES data is not obtained (which comes at an additional cost, if I am not mistaken).

Reply 2: This is a good point. We have added a sentence to the discussion to address this.

Changes in text: We added a sentence to the Discussion (paragraph 1, page 13, line 251-253): "Researchers should consider use of HES data linkage in addition to CPRD Aurum data if studying cancer sites where completeness estimates are low."

Comment 3:
Also, some comparison on equivalent research performed in GOLD would be interesting since -if my understanding is correct- GOLD data are still eligible for research.

Reply 3:
The results for CPRD Aurum presented in this study are consistent with results of prior studies that compared CPRD GOLD to HES, cancer registry, and death registration data (references 18-19), as stated in the Discussion (paragraph 6, page 15, lines 295-296). In addition, the authors at Boston Collaborative Drug Surveillance Program have begun work to evaluate the presence of breast cancer in CPRD Aurum and CPRD GOLD compared to HES APC, HES outpatient, and the Cancer Registry. This study will provide insight into the additional clinical details that can be obtained from each data source.

Changes in text: none

Comment 4:
Some minor revision for typos is needed.

Reply 4: Thank you for pointing this out. We have reviewed the revised manuscript to correct typos, where appropriate.

Reviewer D
Wilcox Hagberg et al. conducted a validation study of the correctness and completeness of the documentation of malignant cancer diagnoses within CPRD Aurum compared to HES APC linkage between 1997-2017. This analysis is important to understand the strengths and limitations of using CPRD Aurum as a resource in cancer research. I have a few comments and suggestions for improvement.

Comment 1:
- The second sentence in the abstract needs a subject ("Our objective") and a tense correction ("examine").

Reply 1: Thank you for this recommendation. We have corrected this.

Changes in text: Abstract/background, page 3, line 35-38 has been edited: "The objective was to examine agreement of cancer diagnoses recorded in CPRD Aurum compared with linked Hospital Episode Statistics (HES) data to provide information on CPRD Aurum data correctness (accuracy, validity) and completeness (presence, missingness)."

Comment 2:
- Were analyses performed after the practice up to standard (UTS) date? If so, please state this in the methods. If not, I suggest performing a sensitivity analysis where entry criteria are limited to later of UTS date, HES coverage start, or the start of the patient's electronic record.

Reply 2: Unlike CPRD GOLD, CPRD Aurum does not include practice up-to-standard (UTS) date; therefore, it is not possible to restrict analyses until after practice UTS date. The study period was restricted to HES coverage start (started April 1, 1997). We created supplement 1 to explain how we derived a patient level start and end date with CPRD Aurum electronic record.

Changes in text: none

Comment 3:
- I appreciated that the authors examined completeness by sex but think it might be important to also look at this by age. Are older people more or less likely to have their diagnoses recorded in Aurum?

Reply 3: We added correctness (Table 2) and completeness (Table 3) estimates stratified by age group. Yes, similar to CPRD GOLD, older patients (≥80 years) are less likely to have their cancer diagnoses recorded in CPRD Aurum compared to younger patients.

Changes in text: See Table 2 for correctness estimates and Table 3 for completeness estimates stratified by age group. We also added the following text to the results section:

Results/Correctness of cancer diagnoses recorded in CPRD Aurum (paragraph 1, Page 9, lines 164-167): ". Correctness was greater than 80% regardless of diagnosis year and age at first cancer diagnosis (Table 2)."

Results/ Completeness of cancer diagnoses recorded in CPRD Aurum (paragraph 1, Page 11, lines 210-212): "When stratified by age at first cancer diagnosis, completeness estimates were similar for those aged 20-49 (64.5%), 50-59 (69.0%), 60-69 (65.3%), and 70-79 (61.9%), but lower for those aged 80 years or older (50.2%) (Table 3)."

Comment 4:

- One of the reasons the authors give for a cancer diagnosis not being present in HES APC is that the cancer care may have occurred in an outpatient setting. It is also possible that the cancer care may have occurred at a non-NHS facility as some patients with private insurance may have opted for care elsewhere. Could you examine this by looking at completeness by the Index of Multiple Deprivation? If those with higher SES have a lower level of completeness it might suggest they went elsewhere for treatment. This measurement is available with HES linkages at the practice level.

Reply 4: We added a statement that non-NHS facilities would be missing from the data. We agree that an assessment of completeness by the Index of Multiple Deprivation would be interesting and informative, but we did not conduct such an analysis in this study. This would be interesting future work.

Changes in text: Discussion (paragraph 5, page 14, lines 280-282) has been edited as follows: "In addition, some cancer events may be treated in outpatient hospital settings or non-NHS facility where some patients with private insurance may have opted for care elsewhere."

Comment 5:

- It would be a nice addition to the analysis to show what the estimated prevalence of each cancer is in England and then compare CPRD Aurum alone, HES APC alone, and Aurum + APC together to the expected prevalence estimate. This would provide researchers with a better estimate of the completeness for specific cancers in CPRD Aurum + HES APC overall.

Reply 5. The source population for this study was a random sample of 50,000 CPRD Aurum patients from among practices with a recent HES APC update in October 2018. To enable comparison of data recordings, patients in the source population were required to have at least one admission for any reason recorded in HES APC. Due to these selection criteria (which were appropriate for our study objective) it would be inappropriate to estimate prevalence and to compare to estimates in England, as the underlying populations are different. As we stated in the discussion, formal validation studies are still needed to assess the validity of cancer outcomes.

Changes in text: none