# The creativity of the study

Suggestion: make it clear the primary objective of the study

1. Although this is a large and frequently utilized US population cohort database, of such studies addressing outcomes in patients with COVID-19 and cancer only a few of which the authors have cited or otherwise addressed.

2. The authors reported finding of a predominant impact of COVID-19 over a diagnosis of cancer is of interest but also has been mentioned in other publications.

this study is not particularly novel since prior studies have addressed this question (e.g. OpenSAFELY; PMID: 32640463) using a similar methodology to the present study and did find a significant association between cancer diagnoses and poor outcomes (which conflicts with the results of the present study). Such studies are insufficiently cited and discussed in this paper

The following statement in the introduction is inaccurate, there are now multiple consortia (some of which are cited, but others not) that have presented data on COVID-19 and cancer. The authors should acknowledge this and amend this statement.

"While the time-sensitive nature of these studies is appreciated, the majority of such studies involves a small number of patients from a single hospital system"

CANCER

# 2. Question about the methods

# 2.1 Fail to include adequate potential confounders

3. While the authors duly recognize the potential for confounding in such observational data repositories, they have not adequately discussed known potential confounding factors not captured in the OPTUM data, they have not adequately discussed the likely impact of social determinants of health not adequately measured as well as the potential for confounding by factors as yet not identified.

Cancer stage, recent cancer treatment, type of therapy, progression vs stable/responsive disease

4. The potential risk factors chosen by the authors are inadequate. They utilized those provided by the CDC for the general population but did not include many of the risk factors demonstrated repeated in previously published studies in patients with cancer and COVID-19 including cancer stage, recent cancer treatment, type of therapy, progression vs stable/responsive disease, etc.

1) I believe the most critical issue related to this analysis are data missing on stage of disease. Can authors provide how many patients had early vs metastatic cancer?

2) Can authors report about mortality in both cohort (early vs metastatic)?

3) Severity of COVID 19 illness should be also reported according to cancer stage (stage4, 3, 2, …no cancer)

4) Do we have any information about treatment patients were receiving (chemotherapy, endocrine therapy, targeted therapy or hormonal treatment)?

The major flaw in this study is that the model does not include: 1) cancer status, consistently shown in other studies to be associated with outcomes; 2) performance status, consistently shown in other studies to be strongly associated with outcomes; 3) timing of COVID-19 diagnosis as a matching criteria (see below, #4).

3. How were recent systemic and radiation therapy captured? Radiation was likely through CPT codes, but how about systemic therapy, including oral antineoplastics? The methods are very scant here.

Include time of diagnosis as a covariate

4. It is necessary to match not just on the demographic variables and region, but also on the timing of the COVID-19 diagnosis. This has become evident to be a critical confounder both for patients with cancer and without. See Grivas et al.

# 2.2 Fail to provide details about propensity score matching or question matching strategy

Include details of PSM

4. The description of the propensity score matching is inadequately detailed. The nearest neighbour method that has been used is the default in 'matchit', but it comes with a range of options. Although it appears that the authors have simply used this package in default mode, what that default mode entails should be specified.

Solution to collinearity: Collinearity is not a problem in PSM

The authors have not adequately discussed the propensity matching process that was employed as well as the potential collinearity between factors utilized in the matching process. A step by step description to their efforts to address collinearity and the associated variance inflation factors for the variables included.

Just use the number of comorbidities as a covariate could be too crude: if possible, can be represented by few more covariates

5. Unclear why the authors did not match on individual comorbidities, as they are far from equal in their usual clinical implication or in their associated with adverse COVID-19 outcomes.

Include balance diagnostics

5. Table 1 (and Table A1): Please provide a measure of imbalance (such as standardized mean difference [SMD]) before and after matching and specify which threshold was used to consider whether the groups were balanced or not.

6. "Although we did not match at individual comorbidity level, the distributions of comorbidities became more similar after matching (Table A.1)."

This claim is not substantiated by the table as there are substantial differences in individual comorbidities after matching. It is unclear what threshold the authors use to define balance.

Rephrase method part: provide more details about why PSM is used and which outcomes from which methods

Reviewer #6: I have been asked to look at this paper as a statistical referee. I am afraid that I find the design of the study and its analysis to be badly matched to each other.

7. It is unclear to me from which analysis the following main finding of the paper is derived from:

"However, the hazard ratio of mortality for individuals with cancer diagnosis versus those without was 1.04 (95% CI: 0.97-1.11) with a p-value of 0.260 after adjustment. This indicates a non-significant effect of cancer history on COVID-19 mortality."

Is this from a multivariable model, since the authors refer to the results being "after adjustment"? If so, which variables are included in the model?

Generally, it is unclear how the authors chose the variables to match on versus the variables to adjust for using a multivariable model.

8. Table 3: Why do the authors perform this analysis on a matched cohort? Is the matching relevant to the associations between these variables and outcomes? It is unclear which patients are included in this analysis? Presumably these are only the patients with cancer?

Use method for paired data is appropriate for analysis after matching

1. The objective of the study was to determine whether cancer is an independent risk factor for poorer outcomes between patients with and without a cancer diagnosis. In order to facilitate this, the authors have performed a successful matching of 31219 of the 31880 cancer patients to individual controls. Having done so, it should have been a straightforward application of methods for paired data, including conditional logistic regression, to answer the questions posed in this study. Conditional logistic regression can also answer wider questions on the influence of the characteristics of cancer, as tabulated in Table A2, and whether the impact of cancer on the outcomes is modified by any covariates.

????

Instead the authors have used unconditional logistic regression or multinomial regression on the successfully matched patients. In doing so the authors have included results on risk factors for COVID-19 outcomes that would have been better estimated on the full dataset. Arguably, these findings are irrelevant as there is a wealth of information on risk factors for COVID-19 outcomes. They have also included relevant estimates relating to the questions being posed by the study, though these estimates will be less precise than those from a matched analysis. Also, the questions being answered are more narrowly focussed than would have been possible through a well-conducted set of conditional logistic regressions.

# 2.3 issues about the regression model

8. Was the model fit robust? Was there evidence of collinearity between variables?

# 2.4 Fail to provide details about missing value handling and potential outcome

6. There is not adequate discussion of how missing values were handled as these are an inherent problem in such data.

Moreover, the statistical analyses used in the study are sometimes confusing or insufficiently detailed (such as how the authors dealt with missing data).

4. From the methods section, it is unclear how missing data was addressed?

# 2. Questions about the data

# 2.1 The representability of the data

~~My only major comment relates to the representability of the Optum database compare to the US population (and beyond). Even if the number of patients is large, this point has to be emphasized.~~

# 2.2 Potential shortage of the database

Do the authors have an explanation for the relatively low rate of CRC and lung cancer patients in their database?

6. It is reported that only 9.4% patients received systemic therapy, and that only 1.2% received the therapy within 4 weeks of COVID-19 diagnosis. This almost assures that the authors have missed significant swathes of treatment exposure, such as oral endocrine and targeted therapies.

11. "9.4% received systemic therapy and 1.2% received the therapy within 4 weeks before COVID-19 diagnosis."

Why did so few patients receive systemic therapies (9.4%)? Is this at any time or within a set time frame? How many patients developed metastatic disease? In which setting was systemic therapy given? Is it possible that oral medications were not captured in this billing-based dataset?

For instance, it is unclear what proportion of patients in this study have active cancer versus just remote diagnoses of cancer (with the low proportion of patients on systemic therapies suggesting that many patients may have indolent disease or be in remission).

# 2.3 Query issue

2. Case identification using ICD-9/10 codes similarly runs the risk of bringing in large numbers of false positives, especially given that these codes are often used to bill for a diagnostic biopsy before the result is known (whether benign or malignant). It is unclear from the methods which ICD-9 and ICD-10 codes were used; some are listed - C50/C61/C18/C34 but these would be far from exhaustive. How many occurrences of a code were required to declare the patient as a cancer case?

# 3. Questions about writing

9. Table 3 and Table 6 are identical.

9. Tables 6 and 3 appear to be entirely duplicated.

10. Figure A1 is not easily readable, please edit the colours or change the layout.

2. Within the limitations of the regression analyses that have been performed, I find the accompanying labelling and description in the Results section less than easy to follow. For example, in the second paragraph of the Mortality sub-section, the text describes logistic regression analysis using matched controls, and references Table 3. The title for Table 3 indicates that it is the mortality and severity for individuals with cancer history that is being shown. The confusion is added to by a reference to the 'hazard ratio' being 1.04 and p=0.260, which is not presented in Tables. Greater care is needed to ensure that the reader is fully informed about precisely what is being shown, and where results are obtained from.

3. Table 3 and Table 6 are identical, apart from different headings. What is going on here?

5. The 'hazard ratio', mentioned in point 2, is not the same as an odds ratio, which is what has been estimated.

6. Age groupings have been formed inappropriately. For example, in Table 1, does a patient aged 50 appear in the first or second age category? Similarly, problems exist for ages 65 and 75. In Table 5, a cancer diagnosis at ages 20, 40, 60, and 80 are all ambiguous as to the category they belong to.

7. P-values should be presented in the format described in Instructions for Authors.

8. The authors occasionally present point estimates and 95% confidence limits to a different number of decimal places. It would be preferable for these to be harmonised.

# Others

7. There is no discussion whether interaction between model covariates were assessed and, if needed, mitigated.

8. It is important for the authors to point out that while the large sample size reported may enhance the precision of the estimates provided, sample size, in and of itself does not address the systematic error that bias the results due to any residual confounding after partial adjustment.

- Detail the impact of systematic therapies (same role for chemo and IO or targeted therapies?)

- Detail the impact of radiotherapy (curative intent versus palliative); indeed, giving the high rate of prostate cancer patients, the impact of the use of (palliative bone) radiotherapy might translate more the stage of the disease rather than a biological effect of the radiotherapy itself on the COVID19 outcomes

1. COVID-19 cohort identification using both laboratory results and ICD-10-CM codes (U07x). A recent study found that billing codes are rife with false positives, as is often the case for ICD-10-CM (Khera et al. medRxiv https://doi.org/10.1101/2021.03.16.21253770). It is very late in the manuscript discussion that the authors suddenly introduce a subgroup analysis based only on PCR, and in this subgroup analysis the primary result is much closer to significance.

7. Considering that survival of more than 5 years is associated with better outcomes, it seems likely that patients with a more recent diagnosis of cancer, who are also presumably more likely to be on active treatment, should be considered separately in a subgroup analysis.