PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Assessment of the concordance between individual- and area-level measures of socio-economic deprivation in a cancer patient cohort in England and Wales
AUTHORS	Ingleby, Fiona; Belot, Aurelien; Atherton, Iain; Baker, Matthew; Elliss-Brookes, Lucy; Woods, Laura

VERSION 1 – REVIEW

REVIEWER	Vesna Zadnik Slovenian Cancer Registry
REVIEW RETURNED	15-Jul-2020

GENERAL COMMENTS	My major concern is on the mixing of concepts of association and agreement in the analysis. These are very different concepts (association answers questions related to aetiology, agreement answers questions on reliability between assessment methods), the methods used to describe them appear similar, but the summery indices used are different. I suggest to apply and interpret the agreement tests when checking the accordance of individual and area level SES measurements. The first statement in “strengths” is to strong (lines 46, 47: “This study presents, for the first time, a detailed description of the strength of association between aggregate area-level deprivation metrics and individual-level deprivation data”). Refer for similar studies example to: doi: 10.3390/ijerph16030296. The most common cancers in the UK should be analysed (lines 112, 113). I miss lung cancer and melanoma among them – they both are strongly associated with SES. Some discussion on the third aspect of the of predictive ability (specificity) would be interesting – foreseen in the Methodology section (lines 192, 193).

REVIEWER	Limor Helpman Juravinski Cancer Center, Hamilton Health Sciences, McMaster University Faculty of Health Sciences
REVIEW RETURNED	08-Aug-2020

GENERAL COMMENTS	This study undertakes an important evaluation of the agreement between individual and geographic measures of deprivation, in an attempt to support the broad use of area-level deprivation indices in population health studies.

	It is a well thought-out analysis on a cohort of cancer patients, comparing census data to validated UK area-level deprivation indices. It also purports to study the association between the evaluated deprivation domains (income, occupation and education). I have some minor comments that I would be interested to have the authors address: 1. The authors chose to focus on cancer patients identified from the National Cancer Registry – although the outcomes and variables assessed are not health or cancer associated, and there is not rationale given to limiting the study population to cancer patients. 2. Several limitations due to missing data are noted, namely -Missing 2011 census data was supplemented with 2001 where available. These datapoints are a decade apart and may well differ for individuals and neighbourhoods across this span of time. What information do we have on rates of change in deprivation indices in the domains evaluated between census cohorts? -Further missing information is supplemented by proxy, using another household member. While this may be representative for household income, it can certainly differ for individual occupation and education levels. Although a sensitivity analysis was performed without imputed data, I find this proxy supplementation particularly liable to error, and would be interested in any published data on its level of accuracy. 3. An important limitation that needs to be addressed is the fact that individual income levels were estimated based on a method using age and occupational classification. Using an estimated income as a gold standard to which area-level data is compared is questionable. Furthermore, this method of estimation would render any evaluation of associations with other domains – especially occupation – meaningless. Would the authors discuss why a domain for which direct data was unavailable was chosen for this analysis? Overall, this piece of research addresses an important and timely question and uses mostly clear methodology. The results are interesting and the discussion is clear and addresses most salient questions, generating some intriguing hypotheses.

VERSION 1 – AUTHOR RESPONSE

Reviewer's comments	Author's responses
Reviewer 1:	
My major concern is on the mixing of concepts of association and agreement in the analysis. These are very different concepts (association answers questions related to aetiology, agreement answers questions on reliability between assessment methods), the	We agree with the reviewer that the language used in the original manuscript was misleading with respect to terminology of association and concordance. Our analyses assessed concordance (or agreement) between individual and area-level methods used to measure deprivation. We have revised all sections of the manuscript in order to avoid interpreting the results in terms of association, and instead we present our analyses

methods used to describe them appear similar, but the summery indices used are different. I suggest to apply and interpret the agreement tests when checking the accordance of individual and area level SES measurements.	in terms of concordance between individual and area-level measures. We also assessed the ability of area-based measures to predict individual-based measures. For this objective, we used the terminology used in the development of prognosis models.
The first statement in “strengths” is too strong (lines 46, 47: “This study presents, for the first time, a detailed description of the strength of association between aggregate area-level deprivation metrics and individual-level deprivation data”). Refer for similar studies example to: doi: 10.3390/ijerph16030296.	Thank you for highlighting this study. We have removed this statement from the revised manuscript, and we have also incorporated the suggested reference into the revised manuscript (Introduction, paragraph 3, reference [14]).
The most common cancers in the UK should be analysed (lines 112, 113). I miss lung cancer and melanoma among them – they both are strongly associated with SES.	We apologise for this error in our cohort description. This should have specified that the cancer types included in the analysis were common in the UK (to provide adequate patient numbers for analysis), as well as being chosen based on evidence of large socio-economic inequalities in terms of cancer survival (see Rachet et al [5]). We have clarified this in the revised manuscript (Methods paragraph 2). Survival is of particular interest as it can be considered a proxy for both healthcare performance and provision as well as early diagnostic interventions such as screening. Incidence, in contrast, reflects long-term socio-economic variations in health behaviours as well as historic occupational exposure. While lung cancer and melanoma show wide socio-economic inequalities in terms of incidence, socio-economic differences in terms of survival for these cancers are relatively narrow (Rachet et al [5]). As well as providing a fuller explanation of cohort selection in the Methods, we have also discussed this aspect of cohort selection as a limitation of the study in the revised Discussion (paragraph 5) and in the ‘Strengths and Limitations’ section.
Some discussion on the third aspect of the predictive ability (specificity) would be interesting – foreseen in the Methodology section (lines 192, 193).	In the Methods, we describe the analyses to calculate accuracy, sensitivity and specificity, and state that these measures are used to generate ROC curves and area-under-curve (AUC) estimates (second paragraph of the sub-section ‘Data analysis’). We present our results in this way because it is a commonly-used and useful method with which to illustrate these analyses (see for example Bryere et al [38]). For each type of deprivation separately, Figure 4 shows a graph of accuracy, and a second graph of sensitivity plotted against 1-specificty (ROC curve). The ROC curve is described using the estimated AUC. In this way, the ROC curves and AUC estimates summarise the combined

	information of the sensitivity and specificity estimates, and the AUC quantifies discrimination/predictive ability. These analyses are key to interpreting the predictive ability of the area-level deprivation variables, and as a key result, this is reported in the Results (second paragraph) and discussed throughout the Discussion (e.g. paragraphs 1, 3, 4). However, we would not be comfortable in interpreting nor commenting solely on specificity, because this has to be interpreted in combination with sensitivity. Therefore, we prefer our results to be based on both aspects of predictive abilities (i.e. sensitivity and specificity combined).
Reviewer 2:	
The authors chose to focus on cancer patients identified from the National Cancer Registry – although the outcomes and variables assessed are not health or cancer associated, and there is not rationale given to limiting the study population to cancer patients.	We have included more detail in the revised manuscript about the cohort selection for this analysis (Methods paragraph 2). Our rationale was based on a broader project for which this cohort is designed. The broader focus is on socio-economic inequalities in cancer survival, therefore the cohort was chosen to include cancer patients from a range of common cancers with known socio-economic differentials in terms of disease outcome. We have stated this in the revised manuscript (Methods paragraph 2). We agree with the reviewer that this cohort is limited and we have revised the Discussion (paragraph 5) to note that it would be interesting to repeat these analyses with a whole-population sample of the ONS-LS (which would require a new project data request and so is not possible with our current dataset).
Missing 2011 census data was supplemented with 2001 where available. These datapoints are a decade apart and may well differ for individuals and neighbourhoods across this span of time. What information do we have on rates of change in deprivation indices in the domains evaluated between census cohorts?	We had no missing data for area-level deprivation measures. Individual-level deprivation data was missing for only a small proportion, and we note that almost half of this missing data was explained by individuals who died prior to the 2011 census (noted in Figure 1 and in Methods paragraph 5), therefore for these individuals, the 2001 census data was the most recent and most appropriate to use. For the remaining missing individual-level data, we used the hierarchical strategy described in the Methods to impute the missing values where possible, using 2001 data where available, then using another household adult as proxy (more detail on this in response to the reviewer's next comment). Unfortunately due to data restrictions to protect cohort member confidentiality, we would not be able to analyse rate of change directly for these variables between census years, and we acknowledge that the imputation method will introduce a degree of error (noted in the Discussion paragraph 3), and that this is a potential limitation of the study (which we have added to the revised 'Strengths and limitations' section of the manuscript). However, this is minimal since the proportion of imputed data was very low (details in Figure 1) and the sensitivity analysis results are included in the supplementary materials (Table S1) in order to

	show that the imputation does not impact on the overall results.
Further missing information is supplemented by proxy, using another household member. While this may be representative for household income, it can certainly differ for individual occupation and education levels. Although a sensitivity analysis was performed without imputed data, I find this proxy supplementation particularly liable to error, and would be interested in any published data on its level of accuracy.	We agree that using one household member as proxy for another will introduce a degree of error. We chose to include the imputed data in our main results since the imputed fraction of the dataset was small, and since the sensitivity analyses designed to test the suitability of the imputation had demonstrated no impact on the concordance estimates (results summarised in Table S1). We are not aware of any published study using this method, perhaps due to the paucity of studies using individual-level data of this type in the UK. However, the rationale for this imputationspan style="font-family:Calibri"> is based on extensive literature showing that partners tend to have similar incomes (Nakosteen et al 2001 Economic Inquiry 39:201-213), occupations (Mansour et al 2018 J Population Economics 31:1005-1033) and educational attainment (Domanski et al 2007 J Euro Societies 9:495-526; Schwartz et al 2005 Demography 42:621-646). We have included this rationale for the imputation, and the relevant references, in the revised Methods (paragraph 5).
An important limitation that needs to be addressed is the fact that individual income levels were estimated based on a method using age and occupational classification. Using an estimated income as a gold standard to which area-level data is compared is questionable. Furthermore, this method of estimation would render any evaluation of associations with other domains – especially occupation – meaningless. Would the authors discuss why a domain for which direct data was unavailable was chosen for this analysis?	The income estimation method is an important point to raise, and we have discussed various implications of this estimation method in the Discussion (paragraph 2). The lack of direct data on individual income is an issue for all studies addressing socio-economic inequalities in the UK population, because income is one of the most commonly-used indicators of socio-economic status but individual level data on taxable income is not routinely collected at a population level in the UK. This shortcoming is a major motivation behind the derivation of the estimation method by Clemens and Dibben [33], as used in our study. This method has been validated as a reliable indicator for UK data and is widely used in published studies. We note that the occupational coding (NS-SEC) which is used to derive the occupational groups in terms of occupational type is different from the more fine-grained Standard Occupation Code (SOC) used as a component part of the income estimation (the income estimate relies on additional individual data, not only the SOC), which gives specific codes for different jobs. Therefore while these variables are clearly linked, they are not directly dependent on each other. It would be true of any dataset that income will largely depend on occupation, so we would expect these variables to be strongly linked even in a dataset with income information collected directly. Despite the lack of direct data on income in the ONS-LS, we consider it very important to be able to include income in our analysis, given how widely it is used in the literature on socio-economic health inequalities. We justify our choice of deprivation variables

	in the Methods (paragraph 3), and the Discussion includes detailed consideration of the differences between the estimation of income at individual and area level and potential limitations of the estimation method (paragraph 2). In addition, we have made sure that the revised manuscript highlights the implicit link between income and occupation variables (Methods, paragraph 4 under sub-section 'Individual-level deprivation metrics').

VERSION 2 – REVIEW

REVIEWER	Limor Helpman Division of Gynecologic Oncology, McMaster University, Canada
REVIEW RETURNED	16-Oct-2020

GENERAL COMMENTS	Thank you for addressing our comments thoroughly and making appropriate modifications. Congratulations on a well written paper!