Assessment of the association between individual- and area-level measures of socio-economic deprivation in a cancer patient cohort in England and Wales

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-041714
Article Type:	Original research
Date Submitted by the Author:	16-Jun-2020

Complete List of Authors:	Ingleby, Fiona; London School of Hygiene & Tropical Medicine, Belot, Aurelien; London School of Hygiene and Tropical Medicine, Atherton, Iain; Edinburgh Napier University, School of Nursing, Midwifery and Social Care Baker, Matthew ; National Cancer Research Institute, Consumer Involvement Advisory Group, Consumer Forum Elliss-Brookes, Lucy; Public Health England Woods, Laura; London School of Hygiene & Tropical Medicine

Keywords:	EPIDEMIOLOGY, Epidemiology < ONCOLOGY, PUBLIC HEALTH

SCHOLARONE™
Manuscripts

Assessment of the association between individual- and area-level measures of socio-economic deprivation in a cancer patient cohort in England and Wales

Fiona C Ingleby [1], Aurélien Belot [1], Iain M Atherton [2], Matthew Baker [3], Lucy Elliss-Brookes [4], Laura M Woods [1]

1. Department of Non-Communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

2. School of Health & Social Care, Edinburgh Napier University, Edinburgh, UK

3. National Cancer Research Institute Consumer Forum, London, UK

4. National Cancer Registration and Analysis Service, Public Health England, London, UK

Correspondence: fiona.ingleby@lshtm.ac.uk

ABSTRACT

Objectives

Most research on health inequalities uses aggregated deprivation scores assigned to the small area where the patient lives; however, the association between aggregate area-level deprivation measures and personal deprivation experienced by individuals living in the area is poorly understood. Our objective was to examine the relationship between individual and ecological deprivation. We tested the association between metrics of income, occupation and education at individual and area levels, and assessed the ability of area-based deprivation measures to predict individual deprivation circumstances.

Setting

England and Wales

Participants

A cancer patient cohort of 9,547 individuals extracted from the ONS Longitudinal Study.

Outcomes

We quantified the association between measures of income, occupation and education at individual and area levels. In addition, we used ROC curves to assess the ability of area-based deprivation measures to predict individual deprivation circumstances.

Results

We found weak associations between individual and area-level indicators of deprivation. The most commonly used indicator in health inequalities research, area-based income deprivation, was a particularly poor predictor of individual income status. Education and occupation were marginally better predictors. The results were consistent across sexes and across six major cancer types.

Conclusions

Our results indicate that ecological deprivation measures capture only part of the relationship between deprivation and health outcomes, especially with respect to income measurement. This has important implications for our understanding of the relationship between deprivation and health, and, as a consequence, healthcare policy. The results have a wide-reaching impact for the way in which we measure and monitor inequalities, and in turn, fund and organise current UK healthcare policy aimed at reducing them.

Strengths and limitations of this study:

- 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, enabling a unique and direct assessment of whether the widely-used aggregate metrics are actually representative of individual deprivation circumstances or not

- The study assesses education, occupation and income indicators of deprivation separately, and compares the associations for each, allowing a much more detailed understanding of deprivation than has been possible to date

- The analyses make use of a large population cohort, representative of all patients in England and Wales, allowing us to draw conclusions about the implications of the results for NHS healthcare policy aimed at reducing health inequalities

- The data used is the most recent individual deprivation data available from the UK census in 2011, but once data is available from the planned 2021 census, the results could be updated in order to evaluate any changes in these associations

INTRODUCTION

There is strong evidence across economically advanced countries that people who live in more socio-economically deprived areas have poorer health outcomes than those living in more advantaged areas [1-8]. These inequalities can be substantial: for example, in England, they account for around 1 in 10 cancer deaths in the first five years after diagnosis [9-11]. There is little evidence of these inequalities narrowing, despite efforts to reduce them [5, 12-13].

Much of the research exploring health inequalities across deprivation groups has been conducted using data aggregated to small geographic areas. These ecological measures represent aggregated individual characteristics for the population. Arguably, attributing these measures to individuals invokes an implicit assumption that area-level measures are at least somewhat representative of an individual's personal deprivation. In reality, whilst these studies have improved our understanding of trends in health outcomes across ecological deprivation groups, they have not directly addressed the association between individual deprivation and mortality because the correlation between ecological measures of deprivation and individual deprivation status is largely unknown.

The association between individual measures, ecological measures and health outcomes is potentially made more complex by the possible existence of contextual effects: that is, that the relationship between individual deprivation and health outcomes might vary by the patient's socio-economic context (ecological deprivation). The degree to which this occurs is likely to depend on the mechanism by which deprivation (either at individual or ecological level) affects outcomes as well as the type of deprivation examined. For example, within oncology a small number of studies have examined the relative effects of individual- and ecological-level deprivation on cancer risk and outcomes; including studies of breast cancer [14] and head and neck cancer risk [15], outcomes for breast and colorectal cancers [16], a meta-analysis of lung cancer outcomes [17], and a study of outcomes for several major common cancers [18]. Generally, these studies have quantified independent effects of both individual and ecological deprivation, and for both, more deprived areas or individuals have lower survival [16-18]. However, the strength and nature of the associations varies considerably across factors including sex, level of geographic aggregation, and which type of deprivation metric is used [17]. Furthermore, these associations are not well understood in a UK context, especially in terms of making use of recent data, and an improved understanding will be important in order to reduce inequalities as part of the NHS long-term plan for 2020-2030 [19]. The research on health inequalities on which the NHS long-term plan is based uses data aggregated to small area level, and so improving our understanding of how this relates to individual-level circumstances is important in terms of developing further policies which more specifically target individual-level variation in health outcomes.

Here, we focus on two key research questions: (1) how strong is the association between individual and ecological socio-economic deprivation measures in a cohort of cancer patients; and (2) how strong are the associations between different types of deprivation variables? These questions enable us to comment on the predictive ability of area-level measures to provide information on individual-level deprivation status in a cancer patient cohort. We discuss the implications of these results in the context of the existing literature on cancer outcome inequalities.

METHODS

We analysed data from the Office for National Statistics Longitudinal Study (LS), individually linked to cancer registrations for England and Wales recorded by the National Cancer Data Repository. The LS is a long-term census-based multi-cohort study using four annual birthdates as the selection criterion. This provides a random 1% sample of the population of England and Wales, clustered by date of birth [20-21]. Data are available for all census variables from the 1971 census through to the most recent 2011 census, as well as for variables derived from external, individual linkage, including cancer registrations and administrative data (births and deaths).

The analysis cohort included LS members present at either or both of the 2001 and 2011 census (Figure 1). We defined the adult cancer patient sub-population as anyone with a first primary malignant cancer diagnosis recorded in the national cancer registry between 1 January 2008 and 30 April 2016 for six of the most common cancer types in England and Wales: breast (ICD-10 code C50), colon (C18), rectum (C19-21), prostate (C61), bladder (C67), and Non-Hodgkin Lymphoma (C82-86). A small number (<20) of sex-site inconsistencies, and also a small number (<30) of men with breast cancer were excluded. Only those aged 18-99 at time of diagnosis were included.

Both at individual and area level, we focussed on three main variables: occupation, education and income; which are commonly used to summarise the broad spectrum of socio-economic status in the social sciences [22].

Ecological deprivation metrics

The Indices of Multiple Deprivation (IMD) were used to measure area-based deprivation. The IMD statistics are calculated for each Lower-level Super Output Area (LSOA) in England and Wales and consist of seven domains. We used the income, employment (occupation) and education domains. LSOA codes were recorded directly for individuals in the 2011 census data, whilst in 2001 census, LSOA codes were derived from concatenating district and ward codes. The temporally closest data were used for each census: for the 2001 census this was the English IMD2004 [23] and Welsh 2005 report [24], and for the 2011 census this was the English IMD2015 [25] and Welsh 2014 report [26]. Each domain was included as ventiles (i.e. 20 equal quantile groups) of the national distribution of areas, as opposed to the raw scores, to avoid LS members being identified in LSOAs with low population size.

Individual-level deprivation metrics

Individual data on age, sex, qualifications and occupation at the 2011 census were extracted for each patient, while individual income was derived using a previously published method (see below). Data were not available from the 2011 census for a small proportion of individuals; in part accounted for by those who were diagnosed with cancer between 2008-2010 and had died prior to the 2011 census (Figure 1). Where possible, data from the 2001 census was used for these individuals. For missing data on qualifications or occupation, data was completed where possible by proxy, using another adult resident in the household (usually household head). We tested the sensitivity of the association statistics to this use of proxy data by comparing results with and without these imputed values, and found very little difference (Table S1). Prior to data completion by proxy, missingness was 12% for occupation data, 2% for education, and 9% for income. After completion of missing data by proxy, missingness was 6%, <1%, and 5% respectively for each of occupation, education, and income individual-level deprivation variables (Figure 1).

Occupation type was derived from the National Statistics Socio-Economic Classification (NS-SEC). The three-group version of the NS-SEC was used, which categorised LS member occupations as *technical, routine and manual occupations*; *intermediate occupations*; or *higher managerial, administrative and professional occupations* [27]. Unlike the finer-scaled versions of the NS-SEC, the three-group version classifies occupations into approximately hierarchical groups. As recommended for the three-group version of the NS-SEC, those without an occupation classification due to long-term unemployment or studentship were treated as missing. We carried out a sensitivity analysis where these individuals were included in the *technical, routine and manual* group, which did not cause any appreciable differences to the strength of associations.

Education level was categorised as one of six groups based on the standard levels of UK qualifications used in the census [28]: *no qualifications*; *1-4 GCSEs or equivalent*; *5+ GCSEs or equivalent*; *apprenticeships and vocational qualifications*; *A-levels or equivalent*; or *degree-level education and higher*.

Weekly income (GBP) was estimated per individual following the method described by Clemens and Dibben [29], which required information on sex, age, and Standard Occupational Classification (SOC) code. We took a data-driven approach to adjust income estimates for those aged over 60 who are most likely to be retired, using observed annualised percentage decreases in income for those aged over 60 reported by the English Longitudinal Study of Ageing (ELSA [30]; see Tables S2 and S3). After applying this correction, LS members were grouped into quintiles by estimated income, from least deprived (Q1) to most deprived (Q5). Quintiles were calculated based on all available LS members (not just cancer patients), separately for each sex.

Patient and public involvement

Due to data protection, we do not have access to individual identifying data from the ONS-LS and so it was not possible to directly involve these participants in the analyses and discussion for this study. Our aim is to share these results with patients and public through publication, in order to address public health issues surrounding health inequalities. In addition, we included cancer patient representatives at each stage of the design, implementation and analysis of this study, as part of the research team.

Data analysis

Males and females were analysed separately, for all cancer types combined and for individual cancers. We tested the strength of the association between each pairwise combination of the six deprivation variables: individual-level income quintile, education and occupation groups; and LSOA-level quintiles for income, education and occupation. Associations were quantified using Cramer's *V* statistic, a measure of the strength of the association between pairs of categorical variables derived from a chi-squared statistic, with 95% confidence intervals also approximated from the chi-squared distribution [31]. The measure has the big advantage of not assuming that categories are ordinal. Cramer's $V<0.10$ are generally interpreted as a weak association and $V>0.30$ strong, although the values depend in part on the number of categories in the variable with the lowest number of groups (*V* can be slightly higher where group numbers are fewer [31]). In most comparisons here, this is the same (five groups), except for comparisons involving individual-level occupation (three groups).

For each type of deprivation metric (i.e. education, income or occupation) we assessed the extent to which the area-level value accurately predicted the 'true' individual-level value. Individuals were considered 'deprived' if their individual-level value was either *no qualifications* or *1-4 GCSEs* (education), *technical, routine and manual* (occupation), or below the 40th centile of income (*quintiles 4* and *5*). A binary classification was applied to the corresponding area-level deprivation variable, which was repeated using each ventile of the area-level variable as the binary threshold. For ventile 1 as threshold, individuals in ventiles 2-20 were categorised as deprived; for ventile 2 as threshold, individuals in ventiles 3-20 were categorised as deprived; and so on. Three aspects of predictive ability were then measured: (1) accuracy, the total proportion of individuals correctly classified; (2) sensitivity, the proportion of 'deprived' individuals correctly classified by the area-level measure; and (3) specificity, the proportion of 'not deprived' individuals correctly classified by the area-level measure. Using these measures, we generated ROC curves [32] for each type of deprivation measure and calculated the area-under-curve (AUC).

All analyses were carried out in R version 3.6.1. Graphs were generated using the package ggplot2 (v3.2.1).

RESULTS

The linked dataset consisted of 4,826 male cancer patients and 4,721 female cancer patients with non-missing individual deprivation data for analysis (Figure 1). The patient cohort tended to include more individuals from the more deprived groups (Table 1).

Our analyses set out first to investigate the association between individual and ecological deprivation measures in cancer patients. We found that the associations between individual- and ecological-level measures were generally weak for both men and women (Figure 2), despite a general trend of the highest proportion of deprived individuals being found in the most deprived areas (Figure 3). We also used binary deprived/not deprived individual and area-level categories to assess how well area-level status predicted individual status and found that none of the area-based measures were strongly reliable predictors of individual-level deprivation status (Figure 4), although occupation performed better than education or income. For occupation, using ventiles 14 (men) and 16 (women) to predict a binary deprivation status yielded the highest predictive accuracy (Figure 4A). The ROC curves showed that for each sex the predictive sensitivity was higher than the 0.5 expected by chance, with AUC values of 0.65 and 0.62 for men and women, respectively (Figure 4B). Predictive sensitivity for education was slightly lower, with an AUC 0.62 for both sexes (Figures 4C and 4D). For income, the predictive sensitivity of area-level income was very low with AUC values of 0.59 for men and 0.56 for women (Figures 4E and 4F), indicating the predictive ability was not much greater than expected by chance.

A secondary aim of the analyses was to test the strength of associations between the different types of deprivation variables included in the study. For both males and females, associations between deprivation variables at the individual level were moderately strong, whilst strong associations were found between the different ecological-level deprivation variables at the LSOA level (Figure 2). There is some evidence of stronger associations between variables at the individual level for women than for men.

The relationships observed in the overall cancer patient cohort were also observed for each cancer when examined separately (Tables S4-S9). There was suggestive evidence of stronger associations for bladder cancer patients than for other cancer types, but small sample size and wide confidence intervals around the estimates make these results hard to interpret.

DISCUSSION

The main aim of this study was to assess the strength of the association between individual and ecological deprivation measures. Overall, the results show that aggregated area-level deprivation metrics are weak predictors of individual-level deprivation status in the cancer patient cohort analysed here. Area-level income displayed a particularly weak association with individual-level income status; whereas area-level occupation, and, to a lesser extent education, appear to have slightly stronger relationships with individual-level measures. These results have important and wide-ranging implications for the interpretation of studies that examine the impact of deprivation on health outcomes, particularly those that form the basis of policies aimed at addressing inequalities. If aggregated area-level deprivation metrics do not fully represent socio-economic variation, then policies based on these measures risk misunderstanding the relationship between health and deprivation.

The calculation of the IMD income domain is based on the proportion of individuals in an area eligible for low-income tax credits or benefits. It is therefore principally an estimator of the distribution of very low incomes, and provides relatively little information about the distribution of mid- to high-incomes. On the other hand, the individual-level income estimation method we used generates a continuous scale of income, the quintiles of which separate individuals with higher incomes from middle and lower incomes more effectively. An additional consideration is the calculation of an individual's income, which is not directly collected as part of census data in the UK and we therefore had to use an estimation method [29]. While this method is validated on UK data, it is nonetheless likely to introduce a degree of error, and perhaps especially so for those individuals managing periods of insecure employment or unemployment, whose occupations will be the least well-documented in the census. As such, ecological and individual metrics quantify income variation in different ways and might not be expected to closely correlate with one another. Income deprivation carries a major weight in the calculation of the IMD for area-level statistics, but our analyses show that it is not straightforward to translate this to individual circumstances. Differentially targeting healthcare funding towards the poorest communities, based on area-level income metrics, is a sensible policy with important potential benefits in terms of reducing inequalities, but it is nonetheless also important to recognise that this could overlook some individuals, and perhaps especially those with low income but not in the lowest income bracket.

For occupation, the area-level IMD domain is based on the proportion of unemployment in an area. In our individual-level data, unemployed individuals were treated as missing data [27] and would therefore have been categorised by proxy (wherever possible) using the occupational category of another adult in the same household. This approach makes an imperfect assumption that the type of occupation of an unemployed individual can be approximated by the occupation of another adult in the same household (usually a spouse or partner). However, the relatively good predictive accuracy of area- and individual-level occupation variables in our results suggests that there is a fair degree of geographic clustering of levels of unemployment and occupation types. Interestingly, the association between individual and ecological occupation measures was not affected by a sensitivity analysis we carried out with unemployed individuals included in the analysis as part of the *technical, routine and manual* group, which could be explained by levels of unemployment being highest in these types of jobs [33].

Our results showed that the ability of area-level education to predict individual status was similar to occupation, although slightly lower. In the case of education, the area-level IMD domain represents the proportion of people in an area with no qualifications, which was one of the individual-level categories we included for education, and this data was directly available from the census. As such, we might have expected a close association between the two education variables. Although more closely associated than the respective income metrics, the overall weak association and predictive power is consistent with the full picture presented by our results that area-level measures only capture some of the variation in deprivation, and do not fully represent individual deprivation status.

Our results suggest that, at least for cancer patients diagnosed in England and Wales, area-level statistics are not a good proxy for individual-level deprivation status, indeed for income deprivation they are only a small improvement upon the toss of a coin. This is somewhat consistent with a recent study of a French population by Bryere *et al* [34], although we generally found slightly lower predictive power for area-level variables to predict individual-level deprivation. A major difference between the two analyses is that where Bryere *et al* used data that was a random sample of the population, we focussed on a cancer patient cohort.

Data availability has undoubtedly been a limiting factor in the ability of previous research to consider both area- and individual-level effects of deprivation. Aggregated data is typically more easily accessible and therefore predominantly features in inequalities research. Our results have implications for the interpretation of studies that rely solely on area-level measures of deprivation such as the IMD. These are useful tools for summarising geographic trends, but our results suggest that caution is needed in terms of extending the interpretation to individual deprivation circumstances. We are not suggesting that aggregated deprivation statistics should not be used, or that the use of aggregated data produces unreliable results for the effect of ecological deprivation. On the contrary, our results show that area- and individual-level health inequalities should be viewed as independent phenomenon, both of interest, and that their separate effects as well as their interaction are likely to be important for understanding and reducing socio-economic differences. For example, further research could address the extent to which inequalities in cancer outcomes are related to area-level factors such as the availability of health care services and resources, in comparison to individual-level factors such as symptom awareness and individual means to access appointments and treatment. Further, establishing whether or not, for instance, more deprived cancer patients experience better outcomes when living in an affluent area compared to living in a more deprived area, due to increased availability of health care services and resources, is integral to fully understanding these differentials and thus the way in which resources should be deployed to address them.

Our data suggest, in fact, that where interventions such as cancer symptom awareness campaigns or screening have been directed at ecologically deprived areas, a significant minority of deprived patients will have missed out. The policies to reduce health inequalities set out in the NHS long-term plan [19] are based on research using aggregate measures of deprivation. If the mechanism by which deprivation affects cancer survival principally functions at an individual level, it follows that such campaigns may have had limited efficiency. Conversely, if ecological factors are the predominant driver of inequalities this approach will have had greater traction. The fact that inequalities are not significantly reducing, even in the context of policy change [13], suggests the latter is, even if only partially, at work.

In conclusion, we have shown that individual and contextual deprivation are not strongly associated in a cancer patient cohort, and we argue that this shows the potential for individual and contextual factors to have independent effects on health inequalities. Further research will be important to disentangle these factors and enable more targeted policy recommendations, especially in terms of individual-level deprivation effects, which have not received much research attention to date. An improved understanding of how individual deprivation affects health outcomes has potential to inform more effective policies to reduce health inequalities.

Funding: This study was supported by a grant from the Economic and Social Research Council (ES/S001808/1).

Ethics approval: LSHTM Ethics Online Application 14600; approved 01/02/2018.

Competing interests: The authors have no conflicts of interest to declare.

Acknowledgements: This work makes use of data from the National Cancer Data Repository prepared by the National Cancer Intelligence Network in association with the National Disease Registration Service and the cancer registries of England and Wales. The permission of the Office for National Statistics to use the Longitudinal Study is gratefully acknowledged, as is the help provided by staff of the Centre for Longitudinal Study Information & User Support (CeLSIUS). CeLSIUS is supported by the ESRC Census of Population Programme under project ES/R00823X/1. The authors alone are responsible for the interpretation of the data in this paper. This work contains statistical data from ONS which is Crown Copyright. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates. We gratefully acknowledge the participation of all members of the ONS Longitudinal Study.

Contributorship statement: All authors contributed to the study design. FCI, AB, IMA and LMW analysed the data. All authors contributed to the interpretation of the results. FCI, AB, IMA and LMW prepared the manuscript. All authors commented on and approved the final manuscript.

Data sharing statement: Data are not publicly available but can be accessed via appropriate application to the ONS Longitudinal Study.

References

1. Butler DC, Petterson S, Bazemore A, Douglas KA. Use of measures of socioeconomic deprivation in planning primary health care workforce and defining health care need in Australia. *Aus J Rural Health* 2010, **18**, 199-204

2. Rey G, Jougla E, Fouillet A, Hemon D. Ecological association between a deprivation index and mortality in France over the period 1997-2001: variations with spatial scale, degree of urbanicty, age, gender and cause of death. *BMC Public Health* 2009, **9**, 33

3. Coleman MP, Babb P, Sloggett A, Quinn MJ, De Stavola BL. Socio-economic inequalities in cancer survival in England and Wales. *Cancer* 2001, **91**, 208-216

4. Coleman MP, Rachet B, Woods LM, Mitry E, Riga M, Cooper N *et al*. Trends and socio-economic inequalities in cancer survival in England and Wales up to 2001. *Br J Cancer* 2004, **90**, 1367-1373

5. Rachet B, Ellis L, Maringe C, Chu T, Nur U, Quaresma M *et al*. Socio-economic inequalities in cancer survival in England after the NHS cancer plan. *Br J Cancer* 2010, **103**, 446-453

6. Singh GK, Williams SD, Siahpush M, Mulhollen A. Socio-economic, rural-urban, and racial inequalities in US cancer survival. *J Cancer Epidemiol* 2011

7. Hastert TA, Beresford SA, Sheppard L, White E. Disparities in cancer incidence and mortality by area-level socio-economic status: a multi-level analysis. *J Epidemiol Community Health* 2015, **69**, 168-176

8. Hagedoorn P, Vandenheede H, Vanthomme K, Willaert D, Gadeyne S. A cohort study into head and neck cancer mortality in Belgium (2001-11). *Oral Oncol* 2016, **61**, 76-82

9. Ellis L, Coleman MP, Rachet B. How many deaths would be avoidable if socio-economic inequalities in cancer survival in England were eliminated ? *Eur J Cancer* 2012, **48**, 270-278

10. Woods LM, Rachet B, O'Connell D, Lawrence G, Tracey E, Willmore A *et al*. Large differences in patterns of breast cancer survival between Australia and England : a comparative study using cancer registry data. *Int J Cancer* 2009, **124**, 2391-2399

11. Rutherford MJ, Ironmonger L, Ormiston-Smith N, Abel GA, Greenberg DC, Lyratzopoulos G *et al*. Estimating the potential survival gains by eliminating socio-economic and sex inequalities in stage of diagnosis of melanoma. *Br J Cancer* 2015, **112**, S116-123

12. Herbert A, Abel GA, Winters S, McPhail S, Elliss-Brookes L, Lyratzopoulos G. Are inequalties in cancer diagnosis through emergency presentation narrowing, widening, or remaining unchanged? Longitudinal analysis of English population-based data 2006-2013. *J Epidemiol Community Health* 2018, **73**, 3-10

13. Exarchakou A, Rachet B, Belot A, Maringe C, Coleman MP. Impact of national cancer policies on cancer survival trends and socio-economic inequalities in England, 1996-2013. *BMJ* 2018, **360**, k764

14. Webster TF, Hoffman K, Weinberg J, Vieira V, Aschengrau A. Community- and individual-level socio-economic status and breast cancer risk: multi-level modelling on Cape Cod, Massachusetts. *Environ Health Perspect* 2008, **116**, 1125-1129

15. Bryere J, Menvielle G, Dejardin O, Launay L, Molinie F, Stucker I *et al*. Neighborhood deprivation and risk of head and neck cancer: a multilevel analysis from France. *Oral Oncol* 2017, **71**, 144-149

16. Lamont EB, Zaslavsky AM, Subramanian SV, Meilleur AE, He Y, Landrum MB. Elderly breast and colorectal cancer patients' clinical course: patient and contexual influences. *Med Care* 2014*,* **52**, 809-817

17. Finke I, Behrens G, Weisser L, Brenner H, Jansen L. Socio-economic differences and lung cancer survival – systematic review and meta-analysis. *Frontiers in Oncology* 2018, **8**, 1-20

18. Sloggett A, Young H, Grundy E. The association of cancer survival with four socio-economic indicators. *BMC Cancer* 2007*,* **7**, 20

19. NHS. 2019. The NHS Long Term Plan. https://www.longtermplan.nhs.uk/ *Accessed: March 2020*

20. Shelton N, Marshall CE, Stuchbury R, Grundy E, Dennet A, Tomlinson A *et al*. Cohort profile: the Office for National Statistics Longitudinal Study. *Intl J Epidemiol* 2019, **48**, 383-384

21. Hattersley L, Creeser R. 1995. Longitudinal Study 1971-1991: History, organisation and quality of data. HMSO, London.

22. Winkleby MA, Jatulis DE, Frank E, Fortmann SP. Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. *Am J Public Health* 1992, **82**, 816-20

23. Office for National Statistics. 2004. The English Indices of Deprivation Statistical Release. https://webarchive.nationalarchives.gov.uk/20100407164233/http://www.communities.gov.uk/archived/general-content/communities/indicesofdeprivation/216309/ *Accessed: January 2020*

24. StatsWales. 2005. Welsh Indices of Multiple Deprivation 2005. https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-Multiple-Deprivation/Archive/WIMD-2005 *Accessed: January 2020*

25. Office for National Statistics. 2015. The English Indices of Deprivation Statistical Release. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015 *Accessed: January 2020*

26. StatsWales. 2014. Welsh Indices of Multiple Deprivation 2014. https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-Multiple-Deprivation/Archive/WIMD-2014 *Accessed: January 2020*

27. Office for National Statistics. 2010. The National Statistics Socio-economic Classification: User Manual. Palgrave Macmillan, London. https://www.ons.gov.uk/methodology/classificationsandstandards/otherclassifications/thenationalstatisticssocioeconomicclassificationnssecrebasedonsoc2010 *Accessed: January 2020*

28. Centre for Longitudinal Study Information and User Support. 2020. Data Dictionary https://www.ucl.ac.uk/infostudies/silva-php-resources/researchProjects/celsius/standalone/index.php *Accessed: January 2020*

29. Clemens T, Dibben C. A method for estimating wage, using standardised occupational classifications, for use in medical research in the place of self-reported income. *BMC Medical Research Methodology* 2014, **14**, 59

30. Banks J, Nazroo J, Steptoe A. *The Dynamics of Ageing: Evidence from the English Longitudinal Study of Ageing*. The Institute for Fiscal Studies: London, UK, 2014.

31. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Lawrence Erlbaum Associates: New Jersey, USA, 1988.

32. Janssens ACJW, Martens FK. Reflection on modern methods: revisiting the area under the ROC curve. *Intl J Epidemiol* 2020, doi: 10.1093/ije/dyz274

33. Office for National Statistics. 2018. Unemployment by previous occupation. https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/datasets/unemploymentbypreviousoccupationunem02 *Accessed: March 2020*

34. Bryere J, Pornet C, Copin N, Launay L, Gusto G, Grosclaude P *et al*. Assessment of the ecological bias of seven aggregate social deprivation indices. *BMC Public Health* 2017*,* **17,** 86

Table 1. Numbers and percentages of cancer patients included in the analysis, by sex; showing distribution across deprivation groups at both individual- and LSOA-level and across cancer types. Data source: ONS LS.

	Men	%	Women	%
Occupation (individual)				
Managerial/Professional	1769	37%	1430	30%
Intermediate	1114	23%	1449	31%
Manual/Technical/Routine	1943	40%	1842	39%
Education (individual)				
Degree-level or higher	1212	25%	1108	23%
A-levels	333	7%	320	7%
Apprenticeship/Vocational training	846	19%	327	7%
5+ GCSEs	372	8%	653	14%
1-4 GCSEs	334	7%	570	12%
No qualifications	1729	34%	1743	37%
Income (individual)*				
Least deprived	627	12%	732	16%
Q2	818	17%	940	20%
Q3	1134	24%	941	20%
Q4	1113	23%	1201	25%
Most deprived	1134	24%	907	19%
Occupation (LSOA)*				
Least deprived	732	15%	760	16%
Q2	863	18%	899	19%
Q3	1051	22%	966	21%
Q4	1048	22%	1005	21%
Most deprived	1132	23%	1091	23%
Education (LSOA)*				
Least deprived	773	16%	755	16%
Q2	878	18%	928	20%
Q3	1014	21%	926	20%
Q4	1060	22%	1030	22%
Most deprived	1101	23%	1082	23%
Income (LSOA)*				
Least deprived	710	15%	725	15%
Q2	820	17%	823	18%
Q3	989	20%	1018	22%
Q4	1137	24%	1049	22%
Most deprived	1170	24%	1106	23%
Cancer type				
Breast (C50)	-	-	3330	71%
Colon (C18)	692	14%	608	13%
Rectal (C19-21)	521	11%	349	7%
Prostate (C61)	2840	59%	-	-

Bladder (C67)	395	8%	130	3%
NHL (C82-86)	378	8%	304	6%
Total	**4826**		**4721**	

466 * Note that quintiles are calculated across the whole population, therefore numbers of cancer
467 patients in each quintile are not necessary evenly divided.

468

Figure legends

Figure 1. Consort diagram describing the dataset linkage and variables used in the analysis, as well as the flow of LS members through the data processing steps: overall numbers, cancer patient sub-population filtering, and missing data exclusions. Data source: ONS LS.

Figure 2. Cramer's *V* ±95% CI for all pairwise combinations of deprivation metrics. Strength of association is indicated by darker shading for men in top half (green; N=4,826), and women in bottom half (purple; N=4,721). Data source: ONS LS.

Figure 3. Stacked barplots showing proportions of men and women in each combination of categories for (A) individual occupation vs. LSOA occupation quintiles; (B) individual education vs. LSOA education quintiles; and (C) individual income vs. LSOA income quintiles. Data source: ONS LS.

Figure 4. Predictive accuracy of LSOA-level variables to predict deprived/not deprived individual deprivation status (left); and ROC curves (right) plotted as sensitivity (true positive rate) against 1-specificity (false positive rate). A/B: occupation; C/D: education; and E/F: income. Dashed lines indicate LSOA ventile value with maximum predictive accuracy when used as the threshold value to differentiate between deprived/not deprived, where deprived are those above this threshold. AUC values are shown next to ROC curves. Data source: ONS LS.

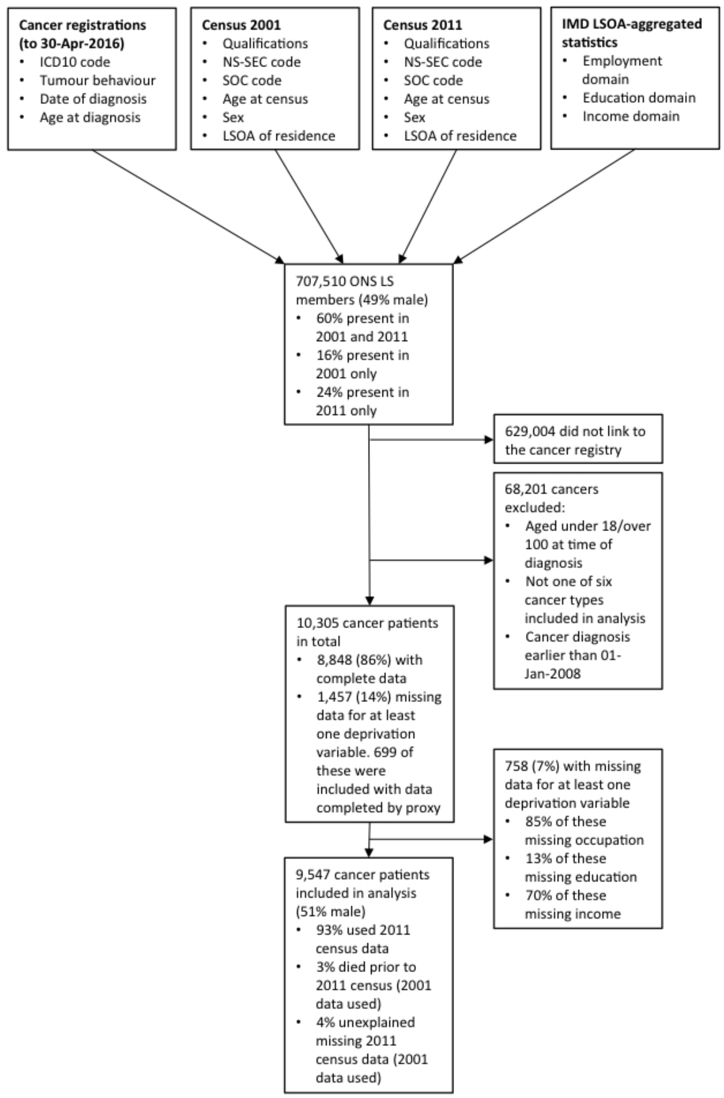

Figure 1. Consort diagram describing the dataset linkage and variables used in the analysis, as well as the flow of LS members through the data processing steps: overall numbers, cancer patient sub-population filtering, and missing data exclusions. Data source: ONS LS.

254x366mm (72 x 72 DPI)

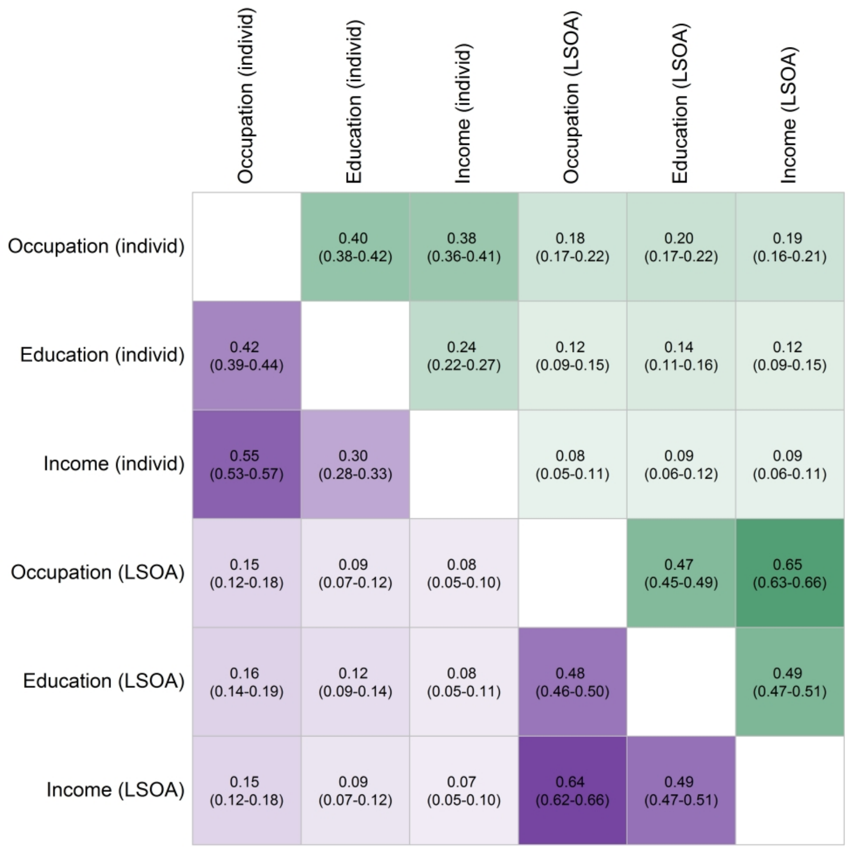

Figure 2. Cramer's V ±95% CI for all pairwise combinations of deprivation metrics. Strength of association is indicated by darker shading for men in top half (green; N=4,826), and women in bottom half (purple; N=4,721). Data source: ONS LS.

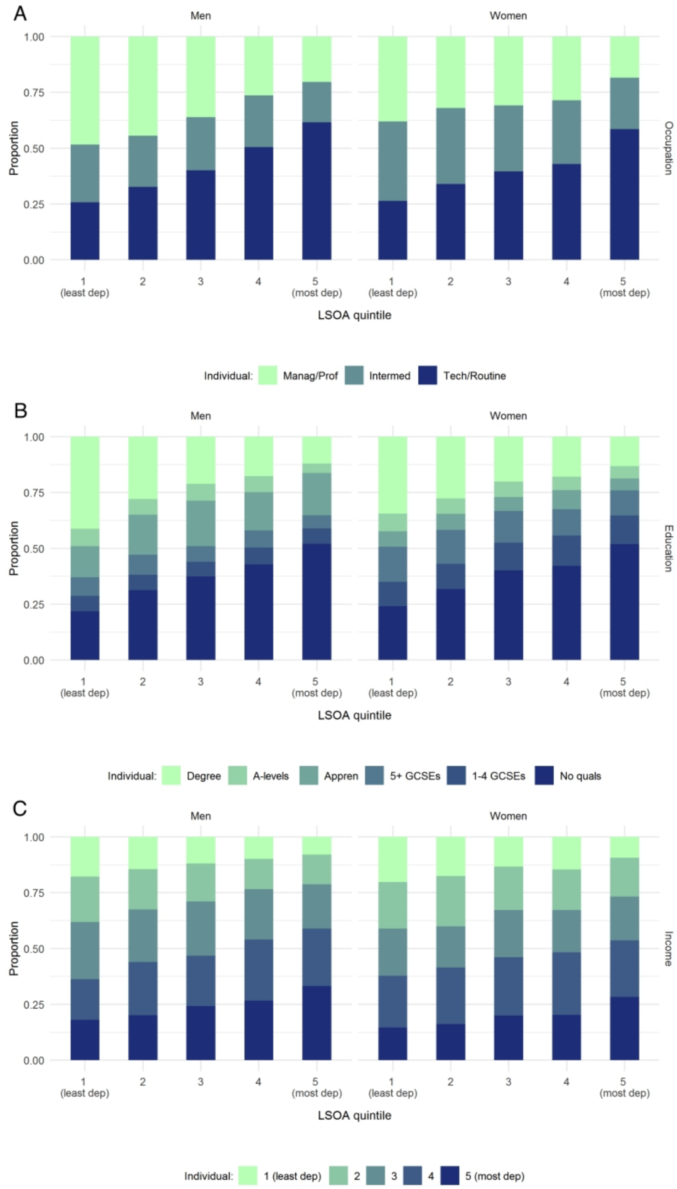

Figure 3. Stacked barplots showing proportions of men and women in each combination of categories for (A) individual occupation vs. LSOA occupation quintiles; (B) individual education vs. LSOA education quintiles; and (C) individual income vs. LSOA income quintiles. Data source: ONS LS.

846x1481mm (72 x 72 DPI)

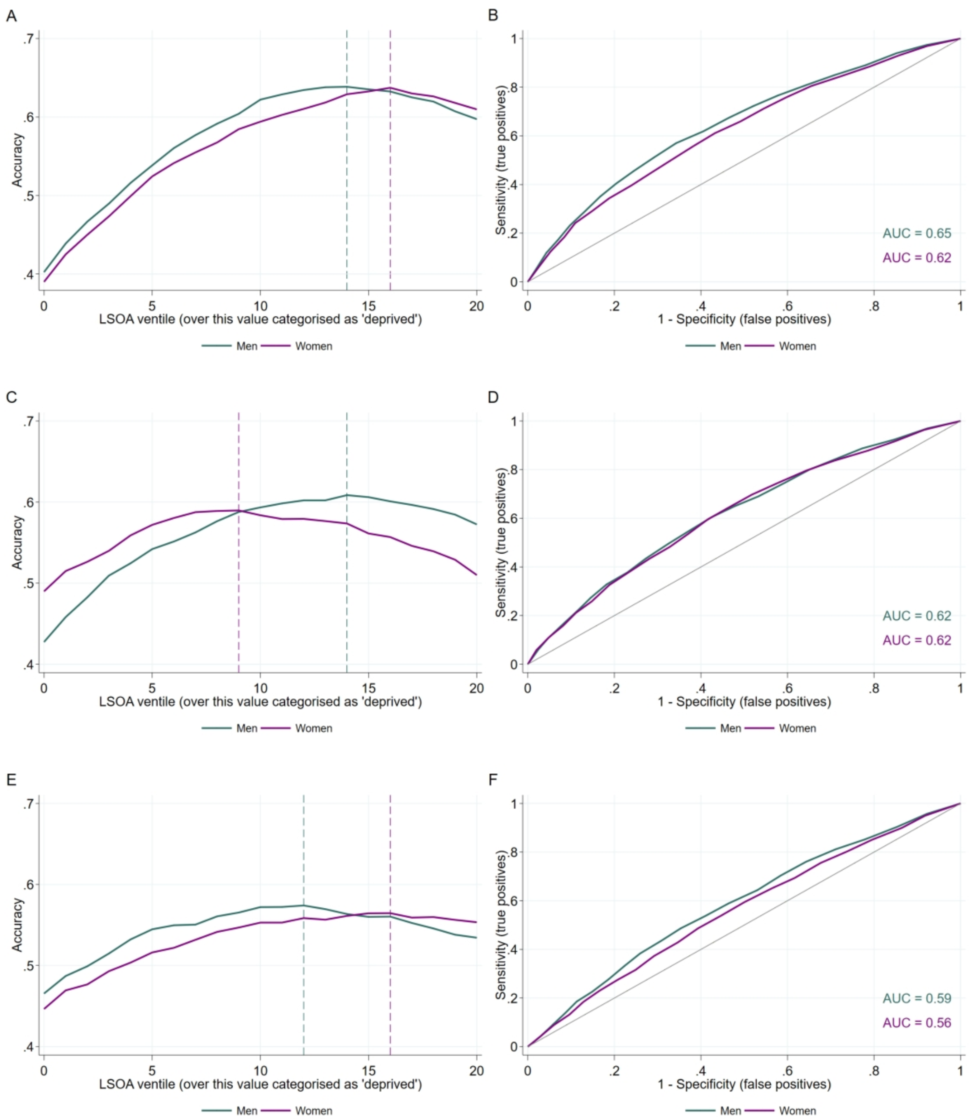

Figure 4. Predictive accuracy of LSOA-level variables to predict deprived/not deprived individual deprivation status (left); and ROC curves (right) plotted as sensitivity (true positive rate) against 1-specificity (false positive rate). A/B: occupation; C/D: education; and E/F: income. Dashed lines indicate LSOA ventile value with maximum predictive accuracy when used as the threshold value to differentiate between deprived/not deprived, where deprived are those above this threshold. AUC values are shown next to ROC curves. Data source: ONS LS.

881x1058mm (72 x 72 DPI)

Supplementary Information

Table S1. Cramer's *V* ±95% CI for all pairwise combinations of deprivation metrics – men in top half (shaded; N=4516), women in bottom half (unshaded; N=4332). These estimates were generated as a sensitivity analysis for the imputation used to complete missing deprivation data by proxy using other household adults, therefore these estimates exclude any individuals with imputed data. Data source: ONS LS.

	Occupation (individ)	**Education (individ)**	**Income (individ)**	**Occupation (LSOA)**	**Education (LSOA)**	**Income (LSOA)**
Occupation (individ)		0.41 (0.39 – 0.43)	0.39 (0.36 – 0.41)	0.18 (0.15 – 0.21)	0.19 (0.17 – 0.22)	0.18 (0.15 – 0.21)
Education (individ)	0.42 (0.40 – 0.45)		0.25 (0.22 – 0.27)	0.12 (0.09 – 0.15)	0.14 (0.11 – 0.17)	0.12 (0.09 – 0.15)
Income (individ)	0.56 (0.54 – 0.58)	0.31 (0.28 – 0.34)		0.08 (0.05 – 0.11)	0.09 (0.06 – 0.11)	0.08 (0.06 – 0.11)
Occupation (LSOA)	0.15 (0.12 – 0.18)	0.09 (0.06 – 0.12)	0.08 (0.05 – 0.11)		0.46 (0.45 – 0.49)	0.63 (0.61 – 0.67)
Education (LSOA)	0.17 (0.14 – 0.19)	0.11 (0.08 – 0.14)	0.08 (0.05 – 0.11)	0.48 (0.45 – 0.50)		0.48 (0.46 – 0.51)
Income (LSOA)	0.15 (0.12 – 0.18)	0.09 (0.06 – 0.12)	0.08 (0.05 – 0.12)	0.63 (0.61 – 0.66)	0.49 (0.46 – 0.51)	

Table S2. Data on average total weekly income per age group in each wave of the ELSA study, taken directly from Table EL2a in the ELSA study report [30]. The shading has been added to illustrate each age cohort moving through the waves of the study (as mid-point age of each age category at two-year intervals of the waves of the study).

Age group	Wave 1 (2002-2003)	Wave 2 (2004-2005)	Wave 3 (2006-2007)	Wave 4 (2008-2009)	Wave 5 (2010-2011)	Wave 6 (2012-2013)
50-54	464.11	453.76	434.42	432.07	399.10	474.18
55-59	422.60	415.02	391.35	385.86	369.92	366.09
60-64	394.19	385.33	369.41	348.70	332.15	339.47
65-69	345.51	313.67	313.08	307.48	296.21	313.03
70-74	297.62	308.96	287.19	292.42	303.03	281.56
75+	275.11	269.58	257.37	266.03	274.18	272.99

The annualised change in income was calculated per age group (taken over the widest possible period for each age group in the given data), and the calculated annual percentage decrease in income was applied to the current dataset for every year after the age of 60. Age groups were assigned according to the age at the start of the study (i.e census year 2001). The actual percentage decreases which were used are shown in **Table S3**.

Table S3. Calculated annualised percentage decreases in income, per age group. Shading is applied per age group to match **Table S2**.

Age group	Observed decrease (years of data)	Annualised decrease
50-54	27% (10)	2.7%
55-59	26% (10)	2.6%
60-64	29% (10)	2.9%
65-69	21% (10)	2.1%
70-74	10.6% (6)	1.8%
75+	2.0% (2)	1.0%

Table S4. Breast cancer (C50) patients only: Cramer's *V* ±95% CI for all pairwise combinations of deprivation metrics – women only, bottom half (unshaded; N=3330). Data source: ONS LS.

	Occupation (individ)	**Education (individ)**	**Income (individ)**	**Occupation (LSOA)**	**Education (LSOA)**	**Income (LSOA)**
Occupation (individ)						
Education (individ)	0.42 (0.39 – 0.45)					
Income (individ)	0.56 (0.53 – 0.58)	0.30 (0.27 – 0.33)				
Occupation (LSOA)	0.16 (0.13 – 0.19)	0.11 (0.07 – 0.14)	0.08 (0.05 – 0.12)			
Education (LSOA)	0.17 (0.14 – 0.20)	0.13 (0.10 – 0.16)	0.09 (0.05 – 0.12)	0.48 (0.46 – 0.51)		
Income (LSOA)	0.16 (0.13 – 0.20)	0.11 (0.08 – 0.14)	0.08 (0.05 – 0.12)	0.64 (0.62 – 0.66)	0.50 (0.47 – 0.52)	

Table S5. Colon cancer (C18) patients only: Cramer's *V* ±95% CI for all pairwise combinations of deprivation metrics – men in top half (shaded; N=692), women in bottom half (unshaded; N=608). Data source: ONS LS.

	Occupation (individ)	**Education (individ)**	**Income (individ)**	**Occupation (LSOA)**	**Education (LSOA)**	**Income (LSOA)**
Occupation (individ)		0.38 (0.32 – 0.44)	0.35 (0.28 – 0.41)	0.19 (0.12 – 0.26)	0.22 (0.15 – 0.29)	0.19 (0.12 – 0.26)
Education (individ)	0.42 (0.35 – 0.48)		0.25 (0.18 – 0.32)	0.11 (0.03 – 0.18)	0.15 (0.08 – 0.23)	0.12 (0.04 – 0.19)
Income (individ)	0.53 (0.48 – 0.59)	0.32 (0.25 – 0.39)		0.09 (0.02 – 0.17)	0.10 (0.03 – 0.18)	0.09 (0.01 – 0.16)
Occupation (LSOA)	0.14 (0.06 – 0.21)	0.12 (0.04 – 0.19)	0.09 (0.01 – 0.17)		0.47 (0.41 – 0.53)	0.65 (0.60 – 0.69)
Education (LSOA)	0.18 (0.10 – 0.17)	0.13 (0.05 – 0.21)	0.09 (0.01 – 0.17)	0.48 (0.41 – 0.54)		0.49 (0.44 – 0.55)
Income (LSOA)	0.16 (0.08 – 0.24)	0.12 (0.04 – 0.19)	0.08 (0.00 – 0.16)	0.64 (0.59 – 0.68)	0.48 (0.42 – 0.54)	

Table S6. Rectal cancer (C19-21) patients only: Cramer's *V* ±95% CI for all pairwise combinations of deprivation metrics – men in top half (shaded; N=521), women in bottom half (unshaded; N=349). Data source: ONS LS.

	Occupation (individ)	Education (individ)	Income (individ)	Occupation (LSOA)	Education (LSOA)	Income (LSOA)
Occupation (individ)		0.36 (0.28 – 0.43)	0.38 (0.31 – 0.45)	0.18 (0.09 – 0.26)	0.19 (0.10 – 0.27)	0.19 (0.10 – 0.27)
Education (individ)	0.38 (0.29 – 0.47)		0.26 (0.18 – 0.34)	0.15 (0.06 – 0.23)	0.15 (0.06 – 0.23)	0.16 (0.07 – 0.24)
Income (individ)	0.54 (0.46 – 0.61)	0.31 (0.21 – 0.40)		0.10 (0.02 – 0.19)	0.10 (0.01 – 0.18)	0.12 (0.03 – 0.20)
Occupation (LSOA)	0.18 (0.08 – 0.28)	0.11 (0.01 – 0.21)	0.13 (0.03 – 0.23)		0.45 (0.38 – 0.51)	0.66 (0.61 – 0.71)
Education (LSOA)	0.16 (0.06 – 0.26)	0.10 (0.00 – 0.21)	0.11 (0.00 – 0.21)	0.47 (0.39 – 0.55)		0.49 (0.42 – 0.55)
Income (LSOA)	0.16 (0.05 – 0.26)	0.08 (0.00 – 0.19)	0.09 (0.00 – 0.20)	0.65 (0.59 – 0.71)	0.53 (0.45 – 0.60)	

Table S7. Prostate cancer (C61) patients only: Cramer's *V* ±95% CI for all pairwise combinations of deprivation metrics – men only, top half (shaded; N=2840). Data source: ONS LS.

	Occupation (individ)	Education (individ)	Income (individ)	Occupation (LSOA)	Education (LSOA)	Income (LSOA)
Occupation (individ)		0.42 (0.39 – 0.45)	0.39 (0.36 – 0.42)	0.18 (0.14 – 0.21)	0.19 (0.15 – 0.22)	0.18 (0.14 – 0.21)
Education (individ)			0.25 (0.21 – 0.28)	0.13 (0.09 – 0.16)	0.14 (0.11 – 0.18)	0.12 (0.09 – 0.16)
Income (individ)				0.08 (0.05 – 0.12)	0.09 (0.05 – 0.12)	0.09 (0.05 – 0.13)
Occupation (LSOA)					0.47 (0.45 – 0.50)	0.64 (0.62 – 0.67)
Education (LSOA)						0.49 (0.46 – 0.52)
Income (LSOA)						

Table S8. Bladder cancer (C67) patients only: Cramer's *V* ±95% CI for all pairwise combinations of deprivation metrics – men in top half (shaded; N=395), women in bottom half (unshaded; N=130). Data source: ONS LS.

	Occupation (individ)	Education (individ)	Income (individ)	Occupation (LSOA)	Education (LSOA)	Income (LSOA)
Occupation (individ)		0.40 (0.32 – 0.48)	0.39 (0.30 – 0.47)	0.22 (0.12 – 0.31)	0.23 (0.13 – 0.32)	0.21 (0.12 – 0.31)
Education (individ)	0.49 (0.35 – 0.61)		0.24 (0.14 – 0.33)	0.15 (0.05 – 0.25)	0.15 (0.06 – 0.25)	0.15 (0.05 – 0.24)
Income (individ)	0.57 (0.44 – 0.67)	0.35 (0.19 – 0.49)		0.16 (0.06 – 0.25)	0.14 (0.05 – 0.24)	0.13 (0.04 – 0.23)
Occupation (LSOA)	0.16 (0.00 – 0.32)	0.25 (0.08 – 0.41)	0.19 (0.02 – 0.35)		0.48 (0.40 – 0.55)	0.66 (0.60 – 0.71)
Education (LSOA)	0.25 (0.08 – 0.40)	0.25 (0.08 – 0.40)	0.23 (0.05 – 0.38)	0.50 (0.35 – 0.62)		0.50 (0.42 – 0.57)
Income (LSOA)	0.23 (0.06 – 0.39)	0.21 (0.04 – 0.37)	0.21 (0.04 – 0.37)	0.60 (0.48 – 0.70)	0.47 (0.32 – 0.60)	

Table S9. NHL cancer (C82-86) patients only: Cramer's *V* ±95% CI for all pairwise combinations of deprivation metrics – men in top half (shaded; N=378), women in bottom half (unshaded; N=304). Data source: ONS LS.

	Occupation (individ)	Education (individ)	Income (individ)	Occupation (LSOA)	Education (LSOA)	Income (LSOA)
Occupation (individ)		0.41 (0.32 – 0.49)	0.40 (0.31 – 0.48)	0.24 (0.14 – 0.33)	0.25 (0.15 – 0.34)	0.25 (0.15 – 0.34)
Education (individ)	0.41 (0.32 – 0.50)		0.27 (0.18 – 0.36)	0.18 (0.08 – 0.27)	0.18 (0.08 – 0.27)	0.17 (0.07 – 0.27)
Income (individ)	0.55 (0.47 – 0.63)	0.30 (0.19 – 0.40)		0.20 (0.10 – 0.29)	0.16 (0.06 – 0.26)	0.19 (0.09 – 0.28)
Occupation (LSOA)	0.17 (0.06 – 0.28)	0.13 (0.02 – 0.24)	0.13 (0.02 – 0.24)		0.46 (0.37 – 0.53)	0.65 (0.58 – 0.70)
Education (LSOA)	0.16 (0.04 – 0.26)	0.15 (0.04 – 0.26)	0.12 (0.01 – 0.23)	0.45 (0.35 – 0.53)		0.46 (0.37 – 0.54)
Income (LSOA)	0.17 (0.05 – 0.27)	0.14 (0.03 – 0.25)	0.12 (0.00 – 0.23)	0.67 (0.61 – 0.73)	0.44 (0.34 – 0.53)	

STROBE guidelines checklist:

Section	Item #	Recommendation	Check
Title/abstract	1	Indicate study design	Term 'cohort' used in both title and abstract
		Abstract summarises what was done and what was found	Structured abstract has this information in relevant sections
Introduction	2	Scientific background and rationale reported	This is described in detail in introduction
Objectives	3	State specific objectives	Listed clearly in abstract and in full in final paragraph of introduction
Methods	4	Present key elements of study design early in manuscript	In both abstract and methods
	5	Describe setting, locations, dates, follow-up, data collection	In first section of methods
	6	Cohort study to include eligibility, patient selection, method of follow-up	In first section of methods
	7	Define all variables	In methods detail
	8	Give sources of data and derivation of all variables	In methods detail
	9	Describe any efforts to address potential sources of bias	Sensitivity analyses described in full
	10	Study size described in full	Described in methods and consort diagram in figure 1
	11	Explain how quantitative variables were handled in analysis	In methods detail
	12	Describe all statistical methods	In statistical methods section
		Describe any methods used for sub-groups or interactions	Not applicable
		Explain how missing data were addressed	In methods detail
		Cohort study to include loss to follow-up if applicable	Not applicable
		Describe any sensitivity analysis	In methods detail
Results	13	Report numbers of individuals at each stage	Consort diagram, figure 1
		Give reasons for non-participation	Consort diagram, figure 1
	14	Characteristics of study cohort	Table 1
		Give numbers with missing data	Consort diagram, figure 1
		Summarise follow-up time	In description of Longitudinal Study in methods
	15	Cohort study to include numbers of outcomes	Table 1
	16	Give unadjusted estimates and 95% CI	In results
	17	Report other analyses	Sensitivity analyses reported, and analyses repeated separately for all cancer types in supplement
Discussion	18	Summarise key results with reference to study objectives	First paragraph of discussion
	19	Discuss limitations and sources of bias	In discussion main text
	20	Give interpretation with acknowledgement of limitations, possible bias, other relevant studies	In discussion main text
	21	Discuss the generalisability	In discussion main text
	22	Give funding information	In funding statement

Assessment of the concordance between individual- and area-level measures of socio-economic deprivation in a cancer patient cohort in England and Wales

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-041714.R1
Article Type:	Original research
Date Submitted by the Author:	24-Sep-2020

Complete List of Authors:	Ingleby, Fiona; London School of Hygiene & Tropical Medicine, Belot, Aurelien; London School of Hygiene and Tropical Medicine, Atherton, Iain; Edinburgh Napier University, School of Nursing, Midwifery and Social Care Baker, Matthew ; National Cancer Research Institute, Consumer Involvement Advisory Group, Consumer Forum Elliss-Brookes, Lucy; Public Health England Woods, Laura; London School of Hygiene & Tropical Medicine

Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Epidemiology, Oncology, Public health
Keywords:	EPIDEMIOLOGY, Epidemiology < ONCOLOGY, PUBLIC HEALTH

SCHOLARONE™
Manuscripts

Assessment of the concordance between individual- and area-level measures of socio-economic deprivation in a cancer patient cohort in England and Wales

Fiona C Ingleby [1], Aurélien Belot [1], Iain M Atherton [2], Matthew Baker [3], Lucy Elliss-Brookes [4], Laura M Woods [1]

1. Department of Non-Communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

2. School of Health & Social Care, Edinburgh Napier University, Edinburgh, UK

3. National Cancer Research Institute Consumer Forum, London, UK

4. National Cancer Registration and Analysis Service, Public Health England, London, UK

Correspondence: fiona.ingleby@lshtm.ac.uk

ABSTRACT

Objectives

Most research on health inequalities uses aggregated deprivation scores assigned to the small area where the patient lives; however, the concordance between aggregate area-level deprivation measures and personal deprivation experienced by individuals living in the area is poorly understood. Our objective was to examine the agreement between individual and ecological deprivation. We tested the concordance between metrics of income, occupation and education at individual and area levels, and assessed the reliability of area-based deprivation measures to predict individual deprivation circumstances.

Setting

England and Wales

Participants

A cancer patient cohort of 9,547 individuals extracted from the ONS Longitudinal Study.

Outcomes

We quantified the concordance between measures of income, occupation and education at individual and area level. In addition, we used ROC curves and the area under the curve (AUC) to assess the reliability of area-based deprivation measures to predict individual deprivation circumstances.

Results

We found low concordance between individual and area-level indicators of deprivation (Cramer's *V* statistics range between 0.07 and 0.20). The most commonly used indicator in health inequalities research, area-based income deprivation, was a poor predictor of individual income status (AUC between 0.56 and 0.59), whereas education and occupation were slightly better predictors (AUC between 0.62 and 0.65). The results were consistent across sexes and across six major cancer types.

Conclusions

Our results indicate that ecological deprivation measures capture only part of the relationship between deprivation and health outcomes, especially with respect to income measurement. This has important implications for our understanding of the relationship between deprivation and health, and, as a consequence, healthcare policy. The results have a wide-reaching impact for the way in which we measure and monitor inequalities, and in turn, fund and organise current UK healthcare policy aimed at reducing them.

Strengths and limitations of this study:

- This study presents a detailed description of concordance between aggregate area-level deprivation metrics and individual-level deprivation data, enabling an assessment of whether the widely-used aggregate metrics are actually representative of individual deprivation circumstances or not

- The study assesses education, occupation and income indicators of deprivation separately, and quantifies concordance between individual and area-level measures for each, allowing a more detailed understanding of deprivation than has been possible to date

- The cohort focusses on cancer types known to have significant socio-economic inequalities in terms of cancer survival, meaning that extension to a broader population (other cancers or the general population) would be of interest in future work

- The data used is the most recent individual deprivation data available from the UK census, and are therefore limited to year 2011, but once data is available from the planned 2021 census, the results could be updated

- A small proportion of individual-level deprivation data was missing and so we completed this information where possible using another household adult, which could have led to a very small number of individuals being misclassified

INTRODUCTION

There is strong evidence across economically advanced countries that people who live in more socio-economically deprived areas have poorer health outcomes than those living in more advantaged areas [1-8]. These inequalities can be substantial: for example, in England, they account for around 1 in 10 cancer deaths in the first five years after diagnosis [9-11]. There is little evidence of these inequalities narrowing, despite efforts to reduce them [5, 12-13].

Much of the research exploring health inequalities across deprivation groups has been conducted using data aggregated to small geographic areas. These ecological measures represent aggregated individual characteristics for the population. Arguably, attributing these measures to individuals invokes an implicit assumption that area-level measures are at least somewhat representative of an individual's personal deprivation. In reality, whilst these studies have improved our understanding of trends in health outcomes across ecological deprivation groups, they have not directly addressed the relationship between individual deprivation and mortality because the concordance between ecological measures of deprivation and individual deprivation status is not well understood.

The relationship between individual measures, ecological measures and health outcomes is potentially made more complex by the possible existence of contextual effects: that is, that the relationship between individual deprivation and health outcomes might vary by the patient's socio-economic context (ecological deprivation). The degree to which this occurs is likely to depend on the mechanism by which deprivation (either at individual or ecological level) affects outcomes as well as the type of deprivation examined. For example, within oncology a small number of studies have examined the relative effects of individual- and ecological-level deprivation on both cancer risk [14-16] and outcomes [17-19]. Generally, these studies have quantified independent effects of both individual and ecological deprivation, and for both, more deprived areas or individuals have higher risk and lower survival [14, 17-19]. However, the strength and nature of these trends varies considerably across factors including sex, level of geographic aggregation, and which type of deprivation metric is used [18]. Furthermore, these associations are not well understood in a UK context, especially in terms of making use of recent data, and an improved understanding will be important in order to reduce inequalities as part of the NHS long-term plan for 2020-2030 [20]. The research on health inequalities on which the NHS long-term plan is based uses data aggregated to small area level, and so improving our understanding of how reliably this matches individual-level circumstances is important in terms of developing further policies which more specifically target individual-level variation in health outcomes.

Here, we focus on two key research questions: (1) how strong is the concordance between individual and ecological socio-economic deprivation measures in a cohort of cancer patients; and (2) how strong is the concordance between different types of deprivation variables? These questions enable us to comment on the predictive ability of area-level measures to provide information on individual-level deprivation status in a cancer patient cohort. We discuss the implications of these results in the context of the existing literature on cancer outcome inequalities.

METHODS

We analysed data from the Office for National Statistics Longitudinal Study (LS), individually linked to cancer registrations for England and Wales recorded by the National Cancer Data Repository. The LS is a long-term census-based multi-cohort study using four annual birthdates as the selection criterion. This provides a random 1% sample of the population of England and Wales, clustered by date of birth [21-22]. Data are available for all census variables from the 1971 census through to the most recent 2011 census, as well as for variables derived from external, individual linkage, including cancer registrations and administrative data (births and deaths).

The analysis cohort included LS members present at either or both of the 2001 and 2011 census (Figure 1). We defined the adult cancer patient sub-population as anyone with a first primary malignant cancer diagnosis recorded in the national cancer registry between 1 January 2008 and 30 April 2016 for six common cancer types in England and Wales: breast (ICD-10 code C50), colon (C18), rectum (C19-21), prostate (C61), bladder (C67), and Non-Hodgkin Lymphoma (C82-86). These cancers were selected for analysis based on evidence of wide socio-economic inequalities in cancer survival in the UK [5]. A small number (<20) of sex-site inconsistencies, and also a small number (<30) of men with breast cancer were excluded. Only those aged 18-99 at time of diagnosis were included.

Both at individual and area level, we focussed on three main variables: occupation, education and income; which are commonly used to summarise the broad spectrum of socio-economic status in the social sciences [23].

Ecological deprivation metrics

The Indices of Multiple Deprivation (IMD) were used to measure area-based deprivation. The IMD statistics are calculated for each Lower-level Super Output Area (LSOA) in England and Wales and consist of seven domains. We used the income, employment (occupation) and education domains. LSOA codes were recorded directly for individuals in the 2011 census data, whilst in 2001 census, LSOA codes were derived from concatenating district and ward codes. The temporally closest data were used for each census: for the 2001 census this was the English IMD2004 [24] and Welsh 2005 report [25], and for the 2011 census this was the English IMD2015 [26] and Welsh 2014 report [27]. Each domain was included as ventiles (i.e. 20 equal quantile groups) of the national distribution of areas, as opposed to the raw scores, to avoid LS members being identified in LSOAs with low population size.

Individual-level deprivation metrics

Individual data on age, sex, qualifications and occupation at the 2011 census were extracted for each patient, while individual income was derived using a previously published method (see below). Individual data were not available from the 2011 census for a small proportion of individuals; in part accounted for by those who were diagnosed with cancer between 2008-2010 and had died prior to the 2011 census (Figure 1). Where possible, data from the 2001 census was used for these individuals. For missing data on qualifications or occupation, data was completed where possible by proxy, using another adult resident in the household (usually household head or spouse). The rationale for this use of information by proxy is based on evidence that partners tend to have similar incomes [28], occupations [29] and educational attainment [30]. We tested the sensitivity of the estimated concordance statistics to this use of proxy data by comparing results with and without these imputed values, and found very little difference (Table S1). Prior to data completion by proxy, missingness was 12% for occupation data, 2% for education, and 9% for income. After completion of missing data by proxy, missingness was 6%, <1%, and 5% respectively for each of occupation, education, and income individual-level deprivation variables (Figure 1).

150 Occupation type was derived from the National Statistics Socio-Economic Classification (NS-SEC). The three-group version of the NS-SEC was used, which categorised LS member occupations as 1) *technical, routine and manual occupations*; 2) *intermediate occupations*; or 3) *higher managerial, administrative and professional occupations* [31]. Unlike the finer-scaled versions of the NS-SEC, the three-group version classifies occupations into approximately hierarchical groups. As recommended for the three-group version of the NS-SEC, those without an occupation classification due to long-term unemployment or studentship were treated as missing [31]. We carried out a sensitivity analysis where these individuals were included in the *technical, routine and manual* group, which did not cause any appreciable differences to the concordance estimates.

Education level was categorised as one of six groups based on the standard levels of UK qualifications used in the census [32]: 1) *no qualifications*; 2) *1-4 GCSEs or equivalent*; 3) *5+ GCSEs or equivalent*; 4) *apprenticeships and vocational qualifications*; 5) *A-levels or equivalent*; or 6) *degree-level education and higher*.

Weekly income (GBP) was estimated per individual following the method described by Clemens and Dibben [33], which required information on sex, age, and Standard Occupational Classification (SOC) code. Income is therefore linked to occupation. The SOC codes used, however, capture specific detail not available within the NS-SEC codes used for the occupation variable, which more broadly classifies types of occupation. We took a data-driven approach to adjust income estimates for those aged over 60 who are most likely to be retired, using observed annualised percentage decreases in income for those aged over 60 reported by the English Longitudinal Study of Ageing (ELSA [34]; see Tables S2 and S3). After applying this correction, LS members were grouped into quintiles by estimated income, from least deprived (Q1) to most deprived (Q5). Quintiles were calculated based on all available LS members (not just cancer patients), separately for each sex.

Patient and public involvement

Due to data protection, we do not have access to individual identifying data from the ONS-LS and so it was not possible to directly involve these participants in the analyses and discussion for this study. Our aim is to share these results with patients and public through publication, in order to address public health issues surrounding health inequalities. In addition, we included cancer patient representatives at each stage of the design, implementation and analysis of this study, as part of the research team.

Data analysis

Males and females were analysed separately, for all cancer types combined and for individual cancers. We tested the degree of concordance between each pairwise combination of the six deprivation variables: individual-level income quintile, education and occupation groups; and LSOA-level quintiles for income, education and occupation. Concordance was quantified using Cramer's *V* statistic, a measure of the concordance between pairs of categorical variables derived from a chi-squared statistic, with 95% confidence intervals also approximated from the chi-squared distribution [35]. The measure has the big advantage of not assuming that categories are ordinal. Cramer's $V<0.10$ are generally interpreted as low concordance and $V>0.30$ high, although the values depend in part on the number of categories in the variable with the lowest number of groups (*V* can be slightly higher where group numbers are fewer [35]). In most comparisons here, this is the same (five groups), except for comparisons involving individual-level occupation (three groups).

For each type of deprivation metric (i.e. education, income or occupation) we assessed the extent to which the area-level value accurately predicted the 'true' individual-level value. Individuals were considered 'deprived' if their individual-level value was either *no qualifications* or *1-4 GCSEs* (education), *technical, routine and manual* (occupation), or below the 40th centile of income (*quintiles 4* and *5*). A binary classification was applied to the corresponding area-level deprivation variable, which was repeated using each ventile of the area-level variable as the binary threshold.

198 For ventile 1 as threshold, individuals in ventiles 2-20 were categorised as deprived; for ventile 2 as threshold, individuals in ventiles 3-20 were categorised as deprived; and so on. Three aspects of predictive ability were then measured: (1) accuracy, the total proportion of individuals correctly classified; (2) sensitivity, the proportion of 'deprived' individuals correctly classified by the area-level measure; and (3) specificity, the proportion of 'not deprived' individuals correctly classified by the area-level measure. Using these measures, we generated ROC curves [36] for each type of deprivation measure and calculated the area-under-curve (AUC) to summarize the ability of the area-based measure to predict individual-level deprivation.

All analyses were carried out in R version 3.6.1. Graphs were generated using the package ggplot2 (v3.2.1).

RESULTS

The linked dataset consisted of 4,826 male cancer patients and 4,721 female cancer patients with non-missing individual deprivation data for analysis (Figure 1). The patient cohort tended to include more individuals from the more deprived groups (Table 1).

Our analyses set out first to investigate concordance between individual and ecological deprivation measures in cancer patients. We found that concordance between individual- and ecological-level measures was generally low for both men and women (Figure 2), despite a general trend of the highest proportion of deprived individuals being found in the most deprived areas (Figure 3). We also used binary deprived/not deprived individual and area-level categories to assess how well area-level status predicted individual status and found that none of the area-based measures were strongly reliable predictors of individual-level deprivation status (Figure 4), although occupation performed better than education or income. For occupation, using ventiles 14 (men) and 16 (women) to predict a binary deprivation status yielded the highest predictive accuracy (Figure 4A). The ROC curves showed that for each sex the ability to discriminate was higher than the 0.5 expected by chance, with AUC values of 0.65 and 0.62 for men and women, respectively (Figure 4B). Predictive ability for education was slightly lower, with an AUC 0.62 for both sexes (Figures 4C and 4D). For income, the predictive ability of area-level income was very low with AUC values of 0.59 for men and 0.56 for women (Figures 4E and 4F), indicating the predictive ability was not much greater than expected by chance.

A secondary aim of the analyses was to test the concordance between the different types of deprivation variables included in the study. For both males and females, concordance between deprivation variables at the individual level was moderately high, whilst high concordance was found between the different ecological-level deprivation variables at the LSOA level (Figure 2). There is some evidence of higher concordance between variables at the individual level for women than for men.

The patterns observed in the overall cancer patient cohort were also observed for each cancer when examined separately (Tables S4-S9). There was suggestive evidence of higher concordance between deprivation variables for bladder cancer patients than for other cancer types, but small sample size and wide confidence intervals around the estimates make these results hard to interpret.

DISCUSSION

The main aim of this study was to assess the concordance between individual and ecological deprivation measures. Area-level income displayed particularly low concordance with individual-level income status; whereas area-level occupation, and, to a lesser extent education, appear to have slightly higher concordance with individual-level measures. Additionally, the results showed that aggregated area-level deprivation metrics are weak predictors of individual-level deprivation status in the cancer patient cohort analysed here. These results have important and wide-ranging implications for the interpretation of studies that examine the impact of deprivation on health outcomes, particularly those that form the basis of policies aimed at addressing inequalities. If aggregated area-level deprivation metrics do not fully represent socio-economic variation between individuals, then policies based on these measures risk misunderstanding the relationship between health and deprivation.

The calculation of the IMD income domain is based on the proportion of individuals in an area eligible for low-income tax credits or benefits. It is therefore principally an estimator of the distribution of very low incomes, and provides relatively little information about the distribution of mid- to high-incomes. On the other hand, the individual-level income estimation method we used generates a continuous scale of income, the quintiles of which separate individuals with higher incomes from middle and lower incomes more effectively. An additional consideration is the calculation of an individual's income, which is not directly collected as part of census data in the UK and we therefore had to use an estimation method [33]. While this method is validated on UK data, it is nonetheless likely to introduce a degree of error, and perhaps especially so for those individuals managing periods of insecure employment or unemployment, whose occupations will be the least well-documented in the census. As such, ecological and individual metrics quantify income variation in different ways and might not be expected to closely match with one another. Income deprivation carries a major weight in the calculation of the IMD for area-level statistics, but our analyses show that it is not straightforward to translate this to individual circumstances. Differentially targeting healthcare funding towards the poorest communities, based on area-level income metrics, is a sensible policy with important potential benefits in terms of reducing inequalities, but it is nonetheless also important to recognise that this could overlook some individuals, and perhaps especially those with low income but not in the lowest income bracket.

For occupation, the area-level IMD domain is based on the proportion of unemployment in an area. In our individual-level data, unemployed individuals were treated as missing data [31] and would therefore have been categorised by proxy (wherever possible) using the occupational category of another adult in the same household. This approach makes an imperfect assumption that the type of occupation of an unemployed individual can be approximated by the occupation of another adult in the same household (usually a spouse or partner). However, the relatively good predictive accuracy of area- and individual-level occupation variables in our results suggests that there is a fair degree of geographic clustering of levels of unemployment and occupation types. Interestingly, concordance between individual and ecological occupation measures was not affected by a sensitivity analysis we carried out with unemployed individuals included in the analysis as part of the *technical, routine and manual* group, which could be explained by levels of unemployment being highest in these types of jobs [37].

Our results showed that the ability of area-level education to predict individual status was similar to occupation, although slightly lower. In the case of education, the area-level IMD domain represents the proportion of people in an area with no qualifications, which was one of the individual-level categories we included for education, and this data was directly available from the census. As such, we might have expected close concordance between the two education variables. Although concordance is higher than for the respective income metrics, concordance is low overall and the predictive ability is consistent with the full picture presented by our results that area-level measures only capture some of the variation in deprivation, and do not fully represent individual deprivation status.

Our results suggest that, at least for cancer patients diagnosed in England and Wales, area-level statistics are not a very good proxy for individual-level deprivation status, indeed for income deprivation they are only a small improvement upon the toss of a coin. This is somewhat consistent with a recent study of a French population by Bryere *et al* [38], although we generally found slightly lower predictive power for area-level variables to predict individual-level deprivation. A major difference between the two analyses is that where Bryere *et al* used data that was a random sample of the population, we focussed on a cancer patient cohort. In particular, the cohort focussed on cancer types with wide socio-economic inequalities in survival [5], and survival inequalities were of interest as survival differences can be readily interpreted in terms of healthcare provision and performance. However, it may be interesting for further research to validate these results on the overall population cohort in the ONS-LS.

Data availability has undoubtedly been a limiting factor in the ability of previous research to consider both area- and individual-level effects of deprivation. Aggregated data is typically more easily accessible and therefore predominantly features in inequalities research. Our results have implications for the interpretation of studies that rely solely on area-level measures of deprivation such as the IMD. These are useful tools for summarising geographic trends, but our results suggest that caution is needed in terms of extending the interpretation to individual deprivation circumstances. We are not suggesting that aggregated deprivation statistics should not be used, or that the use of aggregated data produces unreliable results for the effect of ecological deprivation. On the contrary, our results show that area- and individual-level health inequalities should be viewed as independent phenomenon, both of interest, and that their separate effects as well as their interaction are likely to be important for understanding and reducing socio-economic differences. For example, further research could address the extent to which inequalities in cancer outcomes are related to area-level factors such as the availability of health care services and resources, in comparison to individual-level factors such as symptom awareness and individual means to access appointments and treatment. Further, establishing whether or not, for instance, more deprived cancer patients experience better outcomes when living in an affluent area compared to living in a more deprived area, due to increased availability of health care services and resources, is integral to fully understanding these differentials and thus the way in which resources should be deployed to address them.

Our data suggest, in fact, that where interventions such as cancer symptom awareness campaigns or screening have been directed at ecologically deprived areas, a significant minority of deprived patients will have missed out. The policies to reduce health inequalities set out in the NHS long-term plan [20] are based on research using aggregate measures of deprivation. If the mechanism by which deprivation affects cancer survival principally functions at an individual level, it follows that such campaigns may have had limited efficiency. Conversely, if ecological factors are the predominant driver of inequalities this approach will have had greater traction. The fact that inequalities are not significantly reducing, even in the context of policy change [13], suggests the latter is, even if only partially, at work.

In conclusion, we have shown that individual and contextual deprivation are not highly concordant with each other in a cancer patient cohort, and we argue that this shows the potential for individual and contextual factors to have independent effects on health inequalities. Further research will be important to disentangle these factors and enable more targeted policy recommendations, especially in terms of individual-level deprivation effects, which have not received much research attention to date. An improved understanding of how individual deprivation affects health outcomes has potential to inform more effective policies to reduce health inequalities.

Funding: This study was supported by a grant from the Economic and Social Research Council (ES/S001808/1) and partially by a programme grant from Cancer Research UK (grant number C7923/A18525).

Ethics approval: London School of Hygiene and Tropical Medicine Research Ethics Committee: online application 14600; approved 01/02/2018.

Competing interests: The authors have no conflicts of interest to declare.

Acknowledgements: This work makes use of data from the National Cancer Data Repository prepared by the National Cancer Intelligence Network in association with the National Disease Registration Service and the cancer registries of England and Wales. The permission of the Office for National Statistics to use the Longitudinal Study is gratefully acknowledged, as is the help provided by staff of the Centre for Longitudinal Study Information & User Support (CeLSIUS). CeLSIUS is supported by the ESRC Census of Population Programme under project ES/R00823X/1. The authors alone are responsible for the interpretation of the data in this paper. This work contains statistical data from ONS which is Crown Copyright. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates. We gratefully acknowledge the participation of all members of the ONS Longitudinal Study.

Contributorship statement: All authors (FCI, AB, IMA, MB, LEB and LMW) contributed to the study design. FCI, AB, IMA and LMW analysed the data. All authors (FCI, AB, IMA, MB, LEB and LMW) contributed to the interpretation of the results. FCI, AB, IMA and LMW prepared the manuscript. All authors (FCI, AB, IMA, MB, LEB and LMW) commented on and approved the final manuscript.

Data sharing statement: Data are not publicly available but can be accessed via appropriate application to the ONS Longitudinal Study.

References

1. Butler DC, Petterson S, Bazemore A, Douglas KA. Use of measures of socioeconomic deprivation in planning primary health care workforce and defining health care need in Australia. *Aus J Rural Health* 2010, **18**, 199-204

2. Rey G, Jougla E, Fouillet A, Hemon D. Ecological association between a deprivation index and mortality in France over the period 1997-2001: variations with spatial scale, degree of urbanicty, age, gender and cause of death. *BMC Public Health* 2009, **9**, 33

3. Coleman MP, Babb P, Sloggett A, Quinn MJ, De Stavola BL. Socio-economic inequalities in cancer survival in England and Wales. *Cancer* 2001, **91**, 208-216

4. Coleman MP, Rachet B, Woods LM, Mitry E, Riga M, Cooper N *et al*. Trends and socio-economic inequalities in cancer survival in England and Wales up to 2001. *Br J Cancer* 2004, **90**, 1367-1373

5. Rachet B, Ellis L, Maringe C, Chu T, Nur U, Quaresma M *et al*. Socio-economic inequalities in cancer survival in England after the NHS cancer plan. *Br J Cancer* 2010, **103**, 446-453

6. Singh GK, Williams SD, Siahpush M, Mulhollen A. Socio-economic, rural-urban, and racial inequalities in US cancer survival. *J Cancer Epidemiol* 2011

7. Hastert TA, Beresford SA, Sheppard L, White E. Disparities in cancer incidence and mortality by area-level socio-economic status: a multi-level analysis. *J Epidemiol Community Health* 2015, **69**, 168-176

8. Hagedoorn P, Vandenheede H, Vanthomme K, Willaert D, Gadeyne S. A cohort study into head and neck cancer mortality in Belgium (2001-11). *Oral Oncol* 2016, **61**, 76-82

9. Ellis L, Coleman MP, Rachet B. How many deaths would be avoidable if socio-economic inequalities in cancer survival in England were eliminated ? *Eur J Cancer* 2012, **48**, 270-278

10. Woods LM, Rachet B, O'Connell D, Lawrence G, Tracey E, Willmore A *et al*. Large differences in patterns of breast cancer survival between Australia and England : a comparative study using cancer registry data. *Int J Cancer* 2009, **124**, 2391-2399

11. Rutherford MJ, Ironmonger L, Ormiston-Smith N, Abel GA, Greenberg DC, Lyratzopoulos G *et al*. Estimating the potential survival gains by eliminating socio-economic and sex inequalities in stage of diagnosis of melanoma. *Br J Cancer* 2015, **112**, S116-123

12. Herbert A, Abel GA, Winters S, McPhail S, Elliss-Brookes L, Lyratzopoulos G. Are inequalties in cancer diagnosis through emergency presentation narrowing, widening, or remaining unchanged? Longitudinal analysis of English population-based data 2006-2013. *J Epidemiol Community Health* 2018, **73**, 3-10

13. Exarchakou A, Rachet B, Belot A, Maringe C, Coleman MP. Impact of national cancer policies on cancer survival trends and socio-economic inequalities in England, 1996-2013. *BMJ* 2018, **360**, k764

14. Lokar K, Zagar T, Zadnik V. Estimation of the ecological fallacy in the geographical analysis of the association of socio-economic deprivation and cancer incidence. *Int J Environ Res Public Health* 2019, **16**, 296

15. Webster TF, Hoffman K, Weinberg J, Vieira V, Aschengrau A. Community- and individual-level socio-economic status and breast cancer risk: multi-level modelling on Cape Cod, Massachusetts. *Environ Health Perspect* 2008, **116**, 1125-1129

16. Bryere J, Menvielle G, Dejardin O, Launay L, Molinie F, Stucker I *et al*. Neighborhood deprivation and risk of head and neck cancer: a multilevel analysis from France. *Oral Oncol* 2017, **71**, 144-149

17. Lamont EB, Zaslavsky AM, Subramanian SV, Meilleur AE, He Y, Landrum MB. Elderly breast and colorectal cancer patients' clinical course: patient and contexual influences. *Med Care* 2014, **52**, 809-817

18. Finke I, Behrens G, Weisser L, Brenner H, Jansen L. Socio-economic differences and lung cancer survival – systematic review and meta-analysis. *Frontiers in Oncology* 2018, **8**, 1-20

19. Sloggett A, Young H, Grundy E. The association of cancer survival with four socio-economic indicators. *BMC Cancer* 2007, **7**, 20

20. NHS. 2019. The NHS Long Term Plan. https://www.longtermplan.nhs.uk/ *Accessed: March 2020*

21. Shelton N, Marshall CE, Stuchbury R, Grundy E, Dennet A, Tomlinson A *et al*. Cohort profile: the Office for National Statistics Longitudinal Study. *Intl J Epidemiol* 2019, **48**, 383-384

22. Hattersley L, Creeser R. 1995. Longitudinal Study 1971-1991: History, organisation and quality of data. HMSO, London.

23. Winkleby MA, Jatulis DE, Frank E, Fortmann SP. Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. *Am J Public Health* 1992, **82**, 816-20

24. Office for National Statistics. 2004. The English Indices of Deprivation Statistical Release. https://webarchive.nationalarchives.gov.uk/20100407164233/http://www.communities.gov.uk/archived/general-content/communities/indicesofdeprivation/216309/ *Accessed: January 2020*

25. StatsWales. 2005. Welsh Indices of Multiple Deprivation 2005. https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-Multiple-Deprivation/Archive/WIMD-2005 *Accessed: January 2020*

26. Office for National Statistics. 2015. The English Indices of Deprivation Statistical Release. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015 *Accessed: January 2020*

27. StatsWales. 2014. Welsh Indices of Multiple Deprivation 2014. https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-Multiple-Deprivation/Archive/WIMD-2014 *Accessed: January 2020*

28. Nakosteen RA, Zimmer MA. Spouse selection and earnings: evidence of marital sorting. *Economic Inquiry* 2001, **39**, 201-213

29. Mansour H, McKinnish T. Same-occupation spouses: preferences or search costs? *J Population Economics* 2018, **31**, 1005-1033

30. Schwartz CR, Mare RD. Trends in educational assortative marriage from 1940 to 2003. *Demography* 2005, **42**, 621-646

31. Office for National Statistics. 2010. The National Statistics Socio-economic Classification: User Manual. Palgrave Macmillan, London. https://www.ons.gov.uk/methodology/classificationsandstandards/otherclassifications/thenationalstatisticssocioeconomicclassificationnssecrebasedonsoc2010 *Accessed: January 2020*

32. Centre for Longitudinal Study Information and User Support. 2020. Data Dictionary https://www.ucl.ac.uk/infostudies/silva-php-resources/researchProjects/celsius/standalone/index.php *Accessed: January 2020*

33. Clemens T, Dibben C. A method for estimating wage, using standardised occupational classifications, for use in medical research in the place of self-reported income. *BMC Medical Research Methodology* 2014, **14**, 59

34. Banks J, Nazroo J, Steptoe A. *The Dynamics of Ageing: Evidence from the English Longitudinal Study of Ageing*. The Institute for Fiscal Studies: London, UK, 2014.

35. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Lawrence Erlbaum Associates: New Jersey, USA, 1988.

36. Janssens ACJW, Martens FK. Reflection on modern methods: revisiting the area under the ROC curve. *Intl J Epidemiol* 2020, doi: 10.1093/ije/dyz274

37. Office for National Statistics. 2018. Unemployment by previous occupation. https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/datasets/unemploymentbypreviousoccupationunem02 *Accessed: March 2020*

38. Bryere J, Pornet C, Copin N, Launay L, Gusto G, Grosclaude P *et al*. Assessment of the ecological bias of seven aggregate social deprivation indices. *BMC Public Health* 2017*,* **17,** 86

494 **Table 1.** Numbers and percentages of cancer patients included in the analysis, by sex; showing
495 distribution across deprivation groups at both individual- and LSOA-level and across cancer types.
496 Data source: ONS LS.

	Men	%	Women	%
Occupation (individual)				
Managerial/Professional	1769	37%	1430	30%
Intermediate	1114	23%	1449	31%
Manual/Technical/Routine	1943	40%	1842	39%
Education (individual)				
Degree-level or higher	1212	25%	1108	23%
A-levels	333	7%	320	7%
Apprenticeship/Vocational training	846	19%	327	7%
5+ GCSEs	372	8%	653	14%
1-4 GCSEs	334	7%	570	12%
No qualifications	1729	34%	1743	37%
Income (individual)*				
Least deprived	627	12%	732	16%
Q2	818	17%	940	20%
Q3	1134	24%	941	20%
Q4	1113	23%	1201	25%
Most deprived	1134	24%	907	19%
Occupation (LSOA)*				
Least deprived	732	15%	760	16%
Q2	863	18%	899	19%
Q3	1051	22%	966	21%
Q4	1048	22%	1005	21%
Most deprived	1132	23%	1091	23%
Education (LSOA)*				
Least deprived	773	16%	755	16%
Q2	878	18%	928	20%
Q3	1014	21%	926	20%
Q4	1060	22%	1030	22%
Most deprived	1101	23%	1082	23%
Income (LSOA)*				
Least deprived	710	15%	725	15%
Q2	820	17%	823	18%
Q3	989	20%	1018	22%
Q4	1137	24%	1049	22%
Most deprived	1170	24%	1106	23%
Cancer type				
Breast (C50)	-	-	3330	71%
Colon (C18)	692	14%	608	13%
Rectal (C19-21)	521	11%	349	7%
Prostate (C61)	2840	59%	-	-

Bladder (C67)	395	8%	130	3%
NHL (C82-86)	378	8%	304	6%
Total	**4826**		**4721**	

* Note that quintiles are calculated across the whole population, therefore numbers of cancer patients in each quintile are not necessary evenly divided.

500 **Figure legends**

501 **Figure 1.** Consort diagram describing the dataset linkage and variables used in the analysis, as well as
502 the flow of LS members through the data processing steps: overall numbers, cancer patient sub-
503 population filtering, and missing data exclusions. Data source: ONS LS.

504

505 **Figure 2.** Cramer's *V* ±95% CI for all pairwise combinations of deprivation metrics. Strength of
506 concordance is indicated by darker shading for men in top half (green; N=4,826), and women in
507 bottom half (purple; N=4,721). Data source: ONS LS.

508

509 **Figure 3.** Stacked barplots showing proportions of men and women in each combination of
510 categories for (A) individual occupation vs. LSOA occupation quintiles; (B) individual education vs.
511 LSOA education quintiles; and (C) individual income vs. LSOA income quintiles. Data source: ONS LS.

512

513 **Figure 4.** Predictive accuracy of LSOA-level variables to predict deprived/not deprived individual
514 deprivation status (left); and ROC curves (right) plotted as sensitivity (true positive rate) against 1-
515 specificity (false positive rate). A/B: occupation; C/D: education; and E/F: income. Dashed lines
516 indicate LSOA ventile value with maximum predictive accuracy when used as the threshold value to
517 differentiate between deprived/not deprived, where deprived are those above this threshold. AUC
518 values are shown next to ROC curves. Data source: ONS LS.

519

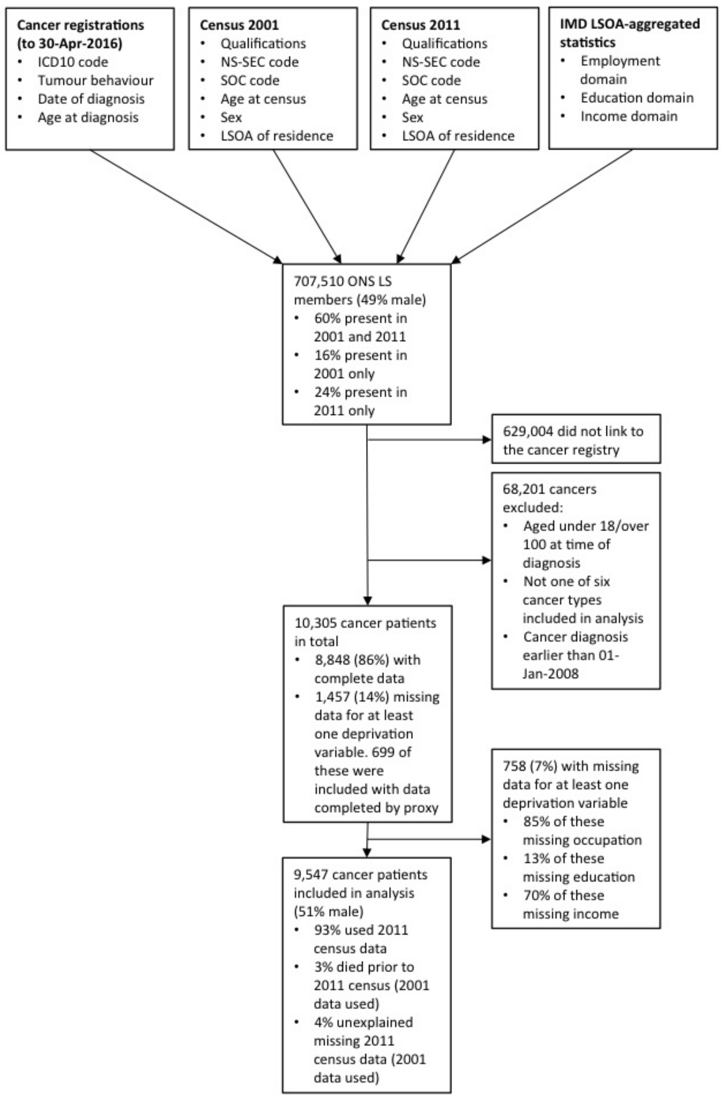

Figure 1. Consort diagram describing the dataset linkage and variables used in the analysis, as well as the flow of LS members through the data processing steps: overall numbers, cancer patient sub-population filtering, and missing data exclusions. Data source: ONS LS.

190x275mm (96 x 96 DPI)

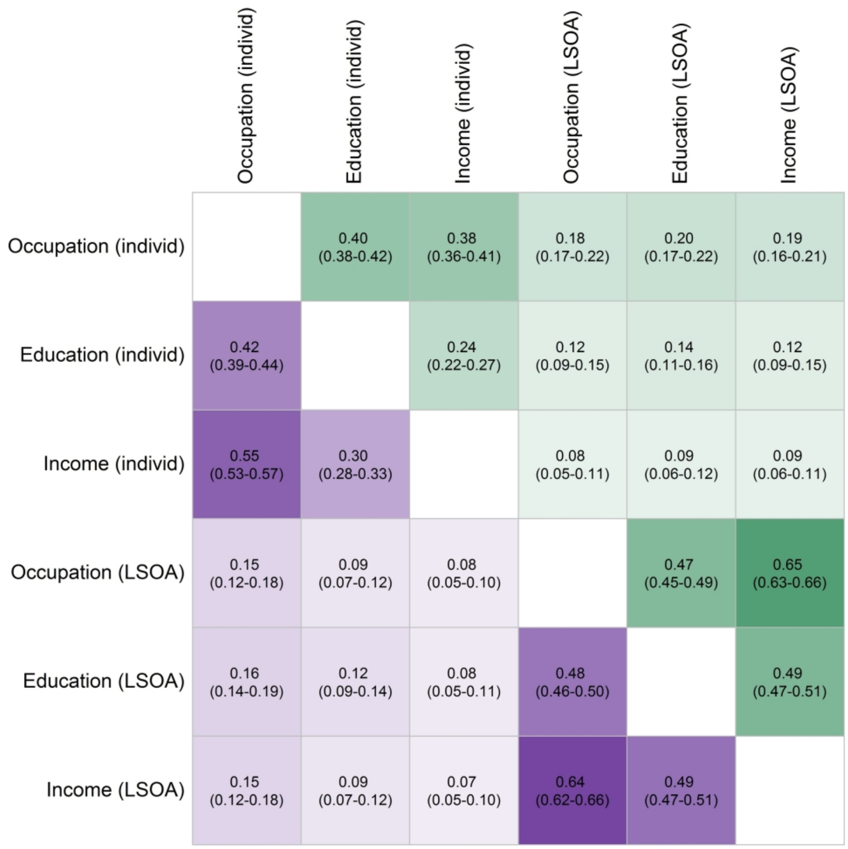

Figure 2. Cramer's V ±95% CI for all pairwise combinations of deprivation metrics. Strength of association is indicated by darker shading for men in top half (green; N=4,826), and women in bottom half (purple; N=4,721). Data source: ONS LS.

254x254mm (300 x 300 DPI)

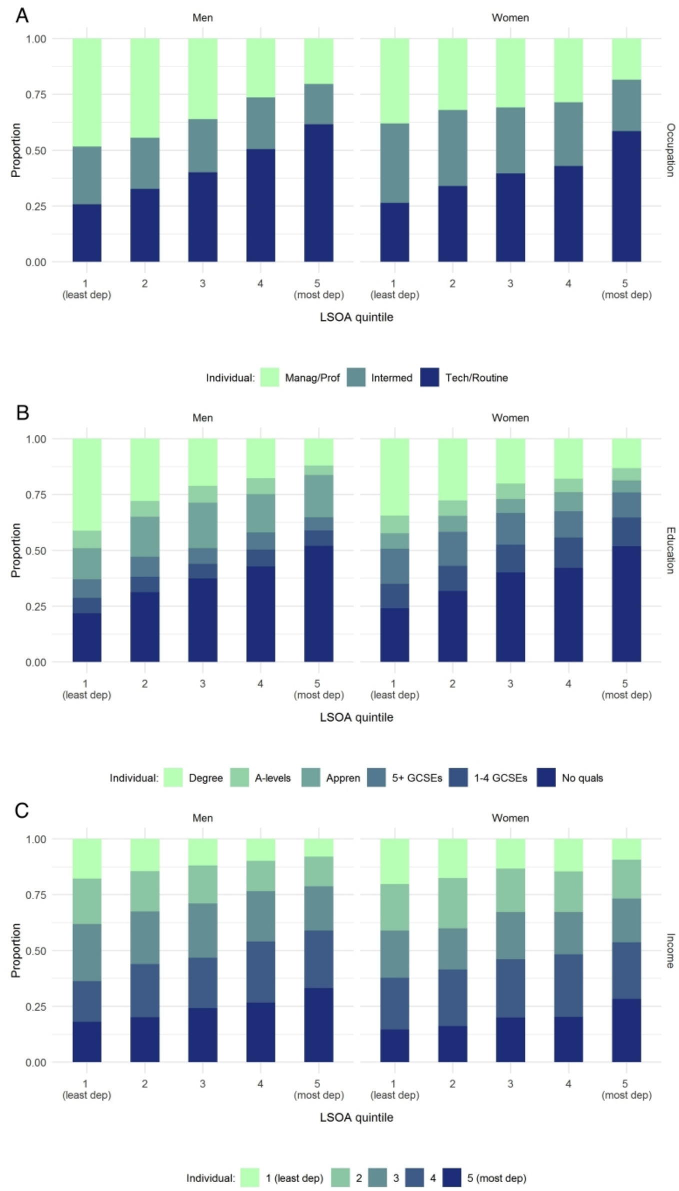

Figure 3. Stacked barplots showing proportions of men and women in each combination of categories for (A) individual occupation vs. LSOA occupation quintiles; (B) individual education vs. LSOA education quintiles; and (C) individual income vs. LSOA income quintiles. Data source: ONS LS.

635x1111mm (96 x 96 DPI)

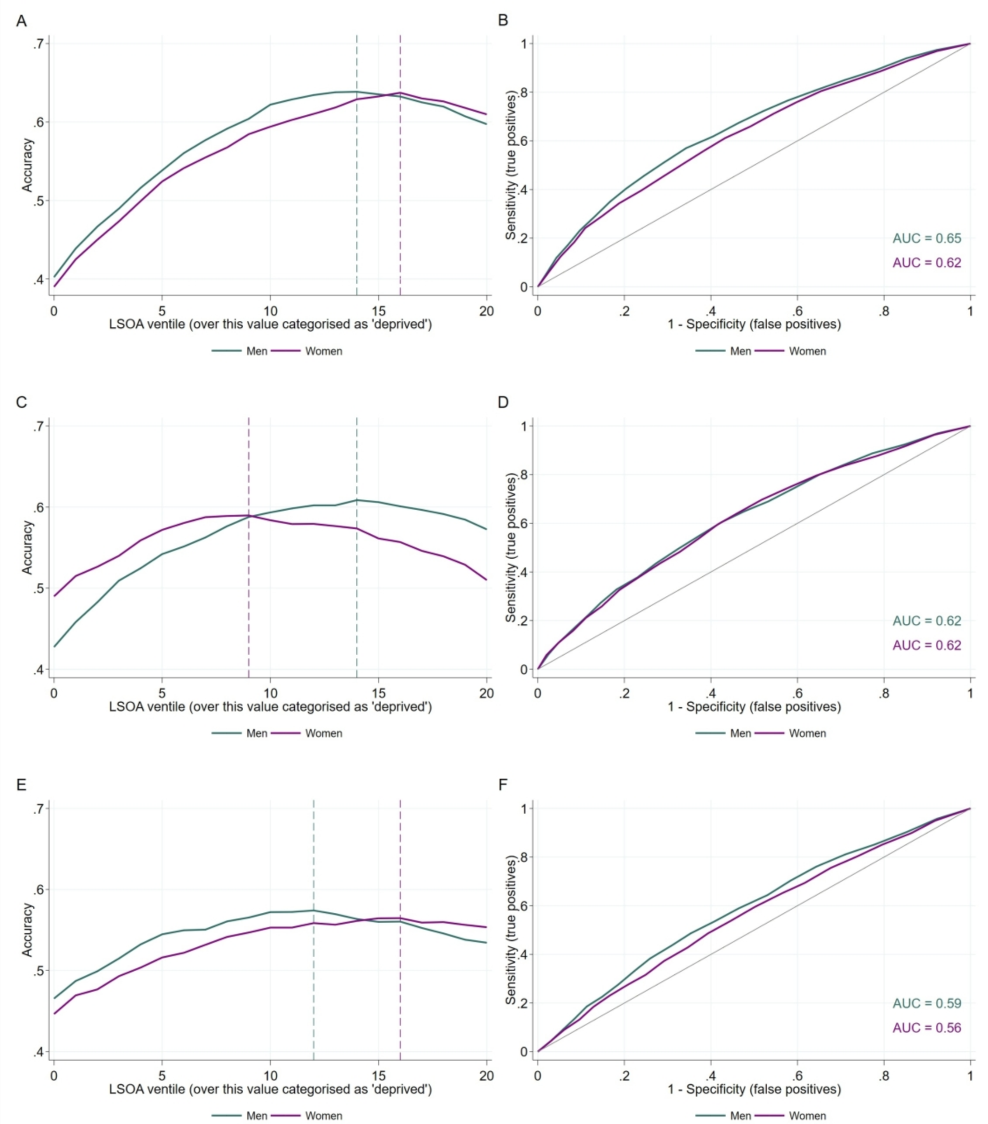

Figure 4. Predictive accuracy of LSOA-level variables to predict deprived/not deprived individual deprivation status (left); and ROC curves (right) plotted as sensitivity (true positive rate) against 1-specificity (false positive rate). A/B: occupation; C/D: education; and E/F: income. Dashed lines indicate LSOA ventile value with maximum predictive accuracy when used as the threshold value to differentiate between deprived/not deprived, where deprived are those above this threshold. AUC values are shown next to ROC curves. Data source: ONS LS.

661x793mm (96 x 96 DPI)

Supplementary Information

Table S1. Cramer's *V* ±95% CI for all pairwise combinations of deprivation metrics – men in top half (shaded; N=4516), women in bottom half (unshaded; N=4332). These estimates were generated as a sensitivity analysis for the imputation used to complete missing deprivation data by proxy using other household adults, therefore these estimates exclude any individuals with imputed data. Data source: ONS LS.

	Occupation (individ)	Education (individ)	Income (individ)	Occupation (LSOA)	Education (LSOA)	Income (LSOA)
Occupation (individ)		0.41 (0.39 – 0.43)	0.39 (0.36 – 0.41)	0.18 (0.15 – 0.21)	0.19 (0.17 – 0.22)	0.18 (0.15 – 0.21)
Education (individ)	0.42 (0.40 – 0.45)		0.25 (0.22 – 0.27)	0.12 (0.09 – 0.15)	0.14 (0.11 – 0.17)	0.12 (0.09 – 0.15)
Income (individ)	0.56 (0.54 – 0.58)	0.31 (0.28 – 0.34)		0.08 (0.05 – 0.11)	0.09 (0.06 – 0.11)	0.08 (0.06 – 0.11)
Occupation (LSOA)	0.15 (0.12 – 0.18)	0.09 (0.06 – 0.12)	0.08 (0.05 – 0.11)		0.46 (0.45 – 0.49)	0.63 (0.61 – 0.67)
Education (LSOA)	0.17 (0.14 – 0.19)	0.11 (0.08 – 0.14)	0.08 (0.05 – 0.11)	0.48 (0.45 – 0.50)		0.48 (0.46 – 0.51)
Income (LSOA)	0.15 (0.12 – 0.18)	0.09 (0.06 – 0.12)	0.08 (0.05 – 0.12)	0.63 (0.61 – 0.66)	0.49 (0.46 – 0.51)	

Table S2. Data on average total weekly income per age group in each wave of the ELSA study, taken directly from Table EL2a in the ELSA study report [34]. The shading has been added to illustrate each age cohort moving through the waves of the study (as mid-point age of each age category at two-year intervals of the waves of the study).

Age group	Wave 1 (2002-2003)	Wave 2 (2004-2005)	Wave 3 (2006-2007)	Wave 4 (2008-2009)	Wave 5 (2010-2011)	Wave 6 (2012-2013)
50-54	464.11	453.76	434.42	432.07	399.10	474.18
55-59	422.60	415.02	391.35	385.86	369.92	366.09
60-64	394.19	385.33	369.41	348.70	332.15	339.47
65-69	345.51	313.67	313.08	307.48	296.21	313.03
70-74	297.62	308.96	287.19	292.42	303.03	281.56
75+	275.11	269.58	257.37	266.03	274.18	272.99

The annualised change in income was calculated per age group (taken over the widest possible period for each age group in the given data), and the calculated annual percentage decrease in income was applied to the current dataset for every year after the age of 60. Age groups were assigned according to the age at the start of the study (i.e census year 2001). The actual percentage decreases which were used are shown in **Table S3**.

Table S3. Calculated annualised percentage decreases in income, per age group. Shading is applied per age group to match **Table S2**.

Age group	Observed decrease (years of data)	Annualised decrease
50-54	27% (10)	2.7%
55-59	26% (10)	2.6%
60-64	29% (10)	2.9%
65-69	21% (10)	2.1%
70-74	10.6% (6)	1.8%
75+	2.0% (2)	1.0%

Table S4. Breast cancer (C50) patients only: Cramer's *V* ±95% CI for all pairwise combinations of deprivation metrics – women only, bottom half (unshaded; N=3330). Data source: ONS LS.

	Occupation (individ)	**Education (individ)**	**Income (individ)**	**Occupation (LSOA)**	**Education (LSOA)**	**Income (LSOA)**
Occupation (individ)						
Education (individ)	0.42 (0.39 – 0.45)					
Income (individ)	0.56 (0.53 – 0.58)	0.30 (0.27 – 0.33)				
Occupation (LSOA)	0.16 (0.13 – 0.19)	0.11 (0.07 – 0.14)	0.08 (0.05 – 0.12)			
Education (LSOA)	0.17 (0.14 – 0.20)	0.13 (0.10 – 0.16)	0.09 (0.05 – 0.12)	0.48 (0.46 – 0.51)		
Income (LSOA)	0.16 (0.13 – 0.20)	0.11 (0.08 – 0.14)	0.08 (0.05 – 0.12)	0.64 (0.62 – 0.66)	0.50 (0.47 – 0.52)	

Table S5. Colon cancer (C18) patients only: Cramer's *V* ±95% CI for all pairwise combinations of deprivation metrics – men in top half (shaded; N=692), women in bottom half (unshaded; N=608). Data source: ONS LS.

	Occupation (individ)	**Education (individ)**	**Income (individ)**	**Occupation (LSOA)**	**Education (LSOA)**	**Income (LSOA)**
Occupation (individ)		0.38 (0.32 – 0.44)	0.35 (0.28 – 0.41)	0.19 (0.12 – 0.26)	0.22 (0.15 – 0.29)	0.19 (0.12 – 0.26)
Education (individ)	0.42 (0.35 – 0.48)		0.25 (0.18 – 0.32)	0.11 (0.03 – 0.18)	0.15 (0.08 – 0.23)	0.12 (0.04 – 0.19)
Income (individ)	0.53 (0.48 – 0.59)	0.32 (0.25 – 0.39)		0.09 (0.02 – 0.17)	0.10 (0.03 – 0.18)	0.09 (0.01 – 0.16)
Occupation (LSOA)	0.14 (0.06 – 0.21)	0.12 (0.04 – 0.19)	0.09 (0.01 – 0.17)		0.47 (0.41 – 0.53)	0.65 (0.60 – 0.69)
Education (LSOA)	0.18 (0.10 – 0.17)	0.13 (0.05 – 0.21)	0.09 (0.01 – 0.17)	0.48 (0.41 – 0.54)		0.49 (0.44 – 0.55)
Income (LSOA)	0.16 (0.08 – 0.24)	0.12 (0.04 – 0.19)	0.08 (0.00 – 0.16)	0.64 (0.59 – 0.68)	0.48 (0.42 – 0.54)	

Table S6. Rectal cancer (C19-21) patients only: Cramer's *V* ±95% CI for all pairwise combinations of deprivation metrics – men in top half (shaded; N=521), women in bottom half (unshaded; N=349). Data source: ONS LS.

	Occupation (individ)	Education (individ)	Income (individ)	Occupation (LSOA)	Education (LSOA)	Income (LSOA)
Occupation (individ)		0.36 (0.28 – 0.43)	0.38 (0.31 – 0.45)	0.18 (0.09 – 0.26)	0.19 (0.10 – 0.27)	0.19 (0.10 – 0.27)
Education (individ)	0.38 (0.29 – 0.47)		0.26 (0.18 – 0.34)	0.15 (0.06 – 0.23)	0.15 (0.06 – 0.23)	0.16 (0.07 – 0.24)
Income (individ)	0.54 (0.46 – 0.61)	0.31 (0.21 – 0.40)		0.10 (0.02 – 0.19)	0.10 (0.01 – 0.18)	0.12 (0.03 – 0.20)
Occupation (LSOA)	0.18 (0.08 – 0.28)	0.11 (0.01 – 0.21)	0.13 (0.03 – 0.23)		0.45 (0.38 – 0.51)	0.66 (0.61 – 0.71)
Education (LSOA)	0.16 (0.06 – 0.26)	0.10 (0.00 – 0.21)	0.11 (0.00 – 0.21)	0.47 (0.39 – 0.55)		0.49 (0.42 – 0.55)
Income (LSOA)	0.16 (0.05 – 0.26)	0.08 (0.00 – 0.19)	0.09 (0.00 – 0.20)	0.65 (0.59 – 0.71)	0.53 (0.45 – 0.60)	

Table S7. Prostate cancer (C61) patients only: Cramer's *V* ±95% CI for all pairwise combinations of deprivation metrics – men only, top half (shaded; N=2840). Data source: ONS LS.

	Occupation (individ)	Education (individ)	Income (individ)	Occupation (LSOA)	Education (LSOA)	Income (LSOA)
Occupation (individ)		0.42 (0.39 – 0.45)	0.39 (0.36 – 0.42)	0.18 (0.14 – 0.21)	0.19 (0.15 – 0.22)	0.18 (0.14 – 0.21)
Education (individ)			0.25 (0.21 – 0.28)	0.13 (0.09 – 0.16)	0.14 (0.11 – 0.18)	0.12 (0.09 – 0.16)
Income (individ)				0.08 (0.05 – 0.12)	0.09 (0.05 – 0.12)	0.09 (0.05 – 0.13)
Occupation (LSOA)					0.47 (0.45 – 0.50)	0.64 (0.62 – 0.67)
Education (LSOA)						0.49 (0.46 – 0.52)
Income (LSOA)						

Table S8. Bladder cancer (C67) patients only: Cramer's *V* ±95% CI for all pairwise combinations of deprivation metrics – men in top half (shaded; N=395), women in bottom half (unshaded; N=130).

Data source: ONS LS.

	Occupation (individ)	**Education (individ)**	**Income (individ)**	**Occupation (LSOA)**	**Education (LSOA)**	**Income (LSOA)**
Occupation (individ)		0.40 (0.32 – 0.48)	0.39 (0.30 – 0.47)	0.22 (0.12 – 0.31)	0.23 (0.13 – 0.32)	0.21 (0.12 – 0.31)
Education (individ)	0.49 (0.35 – 0.61)		0.24 (0.14 – 0.33)	0.15 (0.05 – 0.25)	0.15 (0.06 – 0.25)	0.15 (0.05 – 0.24)
Income (individ)	0.57 (0.44 – 0.67)	0.35 (0.19 – 0.49)		0.16 (0.06 – 0.25)	0.14 (0.05 – 0.24)	0.13 (0.04 – 0.23)
Occupation (LSOA)	0.16 (0.00 – 0.32)	0.25 (0.08 – 0.41)	0.19 (0.02 – 0.35)		0.48 (0.40 – 0.55)	0.66 (0.60 – 0.71)
Education (LSOA)	0.25 (0.08 – 0.40)	0.25 (0.08 – 0.40)	0.23 (0.05 – 0.38)	0.50 (0.35 – 0.62)		0.50 (0.42 – 0.57)
Income (LSOA)	0.23 (0.06 – 0.39)	0.21 (0.04 – 0.37)	0.21 (0.04 – 0.37)	0.60 (0.48 – 0.70)	0.47 (0.32 – 0.60)	

Table S9. NHL cancer (C82-86) patients only: Cramer's *V* ±95% CI for all pairwise combinations of deprivation metrics – men in top half (shaded; N=378), women in bottom half (unshaded; N=304).

Data source: ONS LS.

	Occupation (individ)	**Education (individ)**	**Income (individ)**	**Occupation (LSOA)**	**Education (LSOA)**	**Income (LSOA)**
Occupation (individ)		0.41 (0.32 – 0.49)	0.40 (0.31 – 0.48)	0.24 (0.14 – 0.33)	0.25 (0.15 – 0.34)	0.25 (0.15 – 0.34)
Education (individ)	0.41 (0.32 – 0.50)		0.27 (0.18 – 0.36)	0.18 (0.08 – 0.27)	0.18 (0.08 – 0.27)	0.17 (0.07 – 0.27)
Income (individ)	0.55 (0.47 – 0.63)	0.30 (0.19 – 0.40)		0.20 (0.10 – 0.29)	0.16 (0.06 – 0.26)	0.19 (0.09 – 0.28)
Occupation (LSOA)	0.17 (0.06 – 0.28)	0.13 (0.02 – 0.24)	0.13 (0.02 – 0.24)		0.46 (0.37 – 0.53)	0.65 (0.58 – 0.70)
Education (LSOA)	0.16 (0.04 – 0.26)	0.15 (0.04 – 0.26)	0.12 (0.01 – 0.23)	0.45 (0.35 – 0.53)		0.46 (0.37 – 0.54)
Income (LSOA)	0.17 (0.05 – 0.27)	0.14 (0.03 – 0.25)	0.12 (0.00 – 0.23)	0.67 (0.61 – 0.73)	0.44 (0.34 – 0.53)	

STROBE guidelines checklist:
(note that line numbers refer to the clean version of the revised manuscript without tracked changes)

Section	Item #	Recommendation	Check
Title/abstract	1	Indicate study design	Term 'cohort' used in both title (line 2) and abstract (line 27)
		Abstract summarises what was done and what was found	Structured abstract has this information in relevant objectives, outcomes and results sections (lines 17-38)
Introduction	2	Scientific background and rationale reported	This is described in detail in introduction (pages 5-6)
Objectives	3	State specific objectives	Listed clearly in abstract (lines 17-23) and in full in final paragraph of introduction (page 6: lines 97-102)
Methods	4	Present key elements of study design early in manuscript	In abstract (lines 17-32), introduction (page 6: lines 97-102) and methods (throughout pages 7-11)
	5	Describe setting, locations, dates, follow-up, data collection	In first section of methods (lines 105-120)
	6	Cohort study to include eligibility, patient selection, method of follow-up	In first section of methods (lines 105-120)
	7	Define all variables	In methods lines 124-172
	8	Give sources of data and derivation of all variables	In methods lines 124-172
	9	Describe any efforts to address potential sources of bias	Sensitivity analyses described in lines 144-146, as well as rationale for missing data handling in lines 141-149
	10	Study size described in full	Described in consort diagram (figure 1)
	11	Explain how quantitative variables were handled in analysis	In methods lines 181-207
	12	Describe all statistical methods	In statistical methods section, lines 181-207
		Describe any methods used for sub-groups or interactions	Not applicable to this study; no sub-groups or interactions analysed
		Explain how missing data were addressed	In methods lines 141-149
		Cohort study to include loss to follow-up if applicable	Not applicable to this study
		Describe any sensitivity analysis	Sensitivity analyses described in lines 144-146
Results	13	Report numbers of individuals at each stage	Consort diagram (figure 1)
		Give reasons for non-participation	Consort diagram (figure 1)
	14	Characteristics of study cohort	Table 1 (page 25-26)
		Give numbers with missing data	Consort diagram (figure 1)
		Summarise follow-up time	In description of Longitudinal Study in lines 105-120
	15	Cohort study to include numbers of outcomes	Table 1 (page 25-26)
	16	Give unadjusted estimates and 95% CI	All statistics presented throughout the Results section are unadjusted (as appropriate for our analyses), and 95% CI for all estimates are shown in Figure 2
	17	Report other analyses	Analyses repeated separately for all cancer types in Supplementary tables S4-S9
Discussion	18	Summarise key results with reference to study objectives	First paragraph of discussion, page 14: lines 240-250
	19	Discuss limitations and sources of bias	In discussion main text (lines 258-264, 272-274, 296-300)
	20	Give interpretation with acknowledgement of limitations, possible bias, other relevant studies	Throughout all of Discussion main text, e.g. lines 258-264, 270-274, 292-294
	21	Discuss the generalisability	In discussion main text (lines 290-300)
	22	Give funding information	In funding statement (page 18: lines 337-339)