

BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email info.bmjopen@bmj.com

BMJ Open

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

Journal:	<i>BMJ Open</i>
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



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

1
2
3 1 **Assessment of the association between individual- and area-level measures of socio-economic**
4
5 2 **deprivation in a cancer patient cohort in England and Wales**
6
7

8 3

9
10
11 4 **Fiona C Ingleby¹, Aurélien Belot¹, Iain M Atherton², Matthew Baker³, Lucy Elliss-Brookes⁴, Laura**
12
13 5 **M Woods¹**
14

15
16 6

17
18
19 7 1. Department of Non-Communicable Disease Epidemiology, Faculty of Epidemiology and Population
20
21 8 Health, London School of Hygiene and Tropical Medicine, London, UK
22

23
24 9 2. School of Health & Social Care, Edinburgh Napier University, Edinburgh, UK
25

26
27 10 3. National Cancer Research Institute Consumer Forum, London, UK
28

29
30 11 4. National Cancer Registration and Analysis Service, Public Health England, London, UK
31

32
33 12

34
35
36 13 Correspondence: fiona.ingleby@lshtm.ac.uk
37

38
39 14
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 15 **ABSTRACT**
4
5
6

7 16 **Objectives**
8
9

10 17 Most research on health inequalities uses aggregated deprivation scores assigned to the small area
11
12 18 where the patient lives; however, the association between aggregate area-level deprivation
13
14 19 measures and personal deprivation experienced by individuals living in the area is poorly
15
16 20 understood. Our objective was to examine the relationship between individual and ecological
17
18 21 deprivation. We tested the association between metrics of income, occupation and education at
19
20 22 individual and area levels, and assessed the ability of area-based deprivation measures to predict
21
22 23 individual deprivation circumstances.
23
24
25
26
27

28 24 **Setting**
29
30

31 25 England and Wales
32
33
34

35 26 **Participants**
36
37
38

39 27 A cancer patient cohort of 9,547 individuals extracted from the ONS Longitudinal Study.
40
41
42

43 28 **Outcomes**
44
45

46 29 We quantified the association between measures of income, occupation and education at individual
47
48 30 and area levels. In addition, we used ROC curves to assess the ability of area-based deprivation
49
50 31 measures to predict individual deprivation circumstances.
51
52
53

54 32 **Results**
55
56
57
58
59
60

1
2
3 33 We found weak associations between individual and area-level indicators of deprivation. The most
4
5 34 commonly used indicator in health inequalities research, area-based income deprivation, was a
6
7 35 particularly poor predictor of individual income status. Education and occupation were marginally
8
9 36 better predictors. The results were consistent across sexes and across six major cancer types.
10
11
12

13 37 **Conclusions**

14
15
16
17 38 Our results indicate that ecological deprivation measures capture only part of the relationship
18
19 39 between deprivation and health outcomes, especially with respect to income measurement. This has
20
21 40 important implications for our understanding of the relationship between deprivation and health,
22
23 41 and, as a consequence, healthcare policy. The results have a wide-reaching impact for the way in
24
25 42 which we measure and monitor inequalities, and in turn, fund and organise current UK healthcare
26
27 43 policy aimed at reducing them.
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44

1
2
3 45 **Strengths and limitations of this study:**
4
5
6

7 46 - This study presents, for the first time, a detailed description of the strength of association between
8
9 47 aggregate area-level deprivation metrics and individual-level deprivation data, enabling a unique and
10
11 48 direct assessment of whether the widely-used aggregate metrics are actually representative of
12
13 49 individual deprivation circumstances or not
14
15

16
17 50 - The study assesses education, occupation and income indicators of deprivation separately, and
18
19 51 compares the associations for each, allowing a much more detailed understanding of deprivation
20
21 52 than has been possible to date
22
23

24
25 53 - The analyses make use of a large population cohort, representative of all patients in England and
26
27 54 Wales, allowing us to draw conclusions about the implications of the results for NHS healthcare
28
29 55 policy aimed at reducing health inequalities
30
31

32
33 56 - The data used is the most recent individual deprivation data available from the UK census in 2011,
34
35 57 but once data is available from the planned 2021 census, the results could be updated in order to
36
37 58 evaluate any changes in these associations
38
39

40
41
42 59
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

60 INTRODUCTION

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

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

74 The association between individual measures, ecological measures and health outcomes is
75 potentially made more complex by the possible existence of contextual effects: that is, that the
76 relationship between individual deprivation and health outcomes might vary by the patient's socio-
77 economic context (ecological deprivation). The degree to which this occurs is likely to depend on the
78 mechanism by which deprivation (either at individual or ecological level) affects outcomes as well as
79 the type of deprivation examined. For example, within oncology a small number of studies have
80 examined the relative effects of individual- and ecological-level deprivation on cancer risk and
81 outcomes; including studies of breast cancer [14] and head and neck cancer risk [15], outcomes for
82 breast and colorectal cancers [16], a meta-analysis of lung cancer outcomes [17], and a study of
83 outcomes for several major common cancers [18]. Generally, these studies have quantified

1
2
3 84 independent effects of both individual and ecological deprivation, and for both, more deprived areas
4
5 85 or individuals have lower survival [16-18]. However, the strength and nature of the associations
6
7 86 varies considerably across factors including sex, level of geographic aggregation, and which type of
8
9 87 deprivation metric is used [17]. Furthermore, these associations are not well understood in a UK
10
11 88 context, especially in terms of making use of recent data, and an improved understanding will be
12
13 89 important in order to reduce inequalities as part of the NHS long-term plan for 2020-2030 [19]. The
14
15 90 research on health inequalities on which the NHS long-term plan is based uses data aggregated to
16
17 91 small area level, and so improving our understanding of how this relates to individual-level
18
19 92 circumstances is important in terms of developing further policies which more specifically target
20
21 93 individual-level variation in health outcomes.
22
23
24
25
26

27 94 Here, we focus on two key research questions: (1) how strong is the association between individual
28
29 95 and ecological socio-economic deprivation measures in a cohort of cancer patients; and (2) how
30
31 96 strong are the associations between different types of deprivation variables? These questions enable
32
33 97 us to comment on the predictive ability of area-level measures to provide information on individual-
34
35 98 level deprivation status in a cancer patient cohort. We discuss the implications of these results in the
36
37 99 context of the existing literature on cancer outcome inequalities.
38
39
40
41

42 100
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

101 **METHODS**

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

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

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

119 **Ecological deprivation metrics**

120 The Indices of Multiple Deprivation (IMD) were used to measure area-based deprivation. The IMD
121 statistics are calculated for each Lower-level Super Output Area (LSOA) in England and Wales and
122 consist of seven domains. We used the income, employment (occupation) and education domains.
123 LSOA codes were recorded directly for individuals in the 2011 census data, whilst in 2001 census,

1
2
3 124 LSOA codes were derived from concatenating district and ward codes. The temporally closest data
4
5 125 were used for each census: for the 2001 census this was the English IMD2004 [23] and Welsh 2005
6
7 126 report [24], and for the 2011 census this was the English IMD2015 [25] and Welsh 2014 report [26].
8
9
10 127 Each domain was included as ventiles (i.e. 20 equal quantile groups) of the national distribution of
11
12 128 areas, as opposed to the raw scores, to avoid LS members being identified in LSOAs with low
13
14 129 population size.

18 130 **Individual-level deprivation metrics**

21
22 131 Individual data on age, sex, qualifications and occupation at the 2011 census were extracted for each
23
24 132 patient, while individual income was derived using a previously published method (see below). Data
25
26 133 were not available from the 2011 census for a small proportion of individuals; in part accounted for
27
28 134 by those who were diagnosed with cancer between 2008-2010 and had died prior to the 2011
29
30 135 census (Figure 1). Where possible, data from the 2001 census was used for these individuals. For
31
32 136 missing data on qualifications or occupation, data was completed where possible by proxy, using
33
34 137 another adult resident in the household (usually household head). We tested the sensitivity of the
35
36 138 association statistics to this use of proxy data by comparing results with and without these imputed
37
38 139 values, and found very little difference (Table S1). Prior to data completion by proxy, missingness
39
40 140 was 12% for occupation data, 2% for education, and 9% for income. After completion of missing data
41
42 141 by proxy, missingness was 6%, <1%, and 5% respectively for each of occupation, education, and
43
44 142 income individual-level deprivation variables (Figure 1).

48
49
50 143 Occupation type was derived from the National Statistics Socio-Economic Classification (NS-SEC). The
51
52 144 three-group version of the NS-SEC was used, which categorised LS member occupations as *technical,*
53
54 145 *routine and manual occupations; intermediate occupations; or higher managerial, administrative*
55
56 146 *and professional occupations* [27]. Unlike the finer-scaled versions of the NS-SEC, the three-group
57
58 147 version classifies occupations into approximately hierarchical groups. As recommended for the
59
60

1
2
3 148 three-group version of the NS-SEC, those without an occupation classification due to long-term
4
5 149 unemployment or studentship were treated as missing. We carried out a sensitivity analysis where
6
7
8 150 these individuals were included in the *technical, routine and manual* group, which did not cause any
9
10 151 appreciable differences to the strength of associations.

11
12
13 152 Education level was categorised as one of six groups based on the standard levels of UK
14
15 153 qualifications used in the census [28]: *no qualifications; 1-4 GCSEs or equivalent; 5+ GCSEs or*
16
17
18 154 *equivalent; apprenticeships and vocational qualifications; A-levels or equivalent; or degree-level*
19
20 155 *education and higher.*

21
22
23
24 156 Weekly income (GBP) was estimated per individual following the method described by Clemens and
25
26 157 Dibben [29], which required information on sex, age, and Standard Occupational Classification (SOC)
27
28 158 code. We took a data-driven approach to adjust income estimates for those aged over 60 who are
29
30 159 most likely to be retired, using observed annualised percentage decreases in income for those aged
31
32 160 over 60 reported by the English Longitudinal Study of Ageing (ELSA [30]; see Tables S2 and S3). After
33
34 161 applying this correction, LS members were grouped into quintiles by estimated income, from least
35
36 162 deprived (Q1) to most deprived (Q5). Quintiles were calculated based on all available LS members
37
38 163 (not just cancer patients), separately for each sex.

43 164 **Patient and public involvement**

44
45
46
47 165 Due to data protection, we do not have access to individual identifying data from the ONS-LS and so
48
49 166 it was not possible to directly involve these participants in the analyses and discussion for this study.
50
51 167 Our aim is to share these results with patients and public through publication, in order to address
52
53 168 public health issues surrounding health inequalities. In addition, we included cancer patient
54
55 169 representatives at each stage of the design, implementation and analysis of this study, as part of the
56
57 170 research team.
58
59
60

171 **Data analysis**

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

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

1
2
3 196 All analyses were carried out in R version 3.6.1. Graphs were generated using the package ggplot2
4
5 197 (v3.2.1).
6
7
8
9 198
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

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

1
2
3 222 is some evidence of stronger associations between variables at the individual level for women than
4
5 223 for men.
6
7
8

9 224 The relationships observed in the overall cancer patient cohort were also observed for each cancer
10
11 225 when examined separately (Tables S4-S9). There was suggestive evidence of stronger associations
12
13 226 for bladder cancer patients than for other cancer types, but small sample size and wide confidence
14
15
16 227 intervals around the estimates make these results hard to interpret.
17
18

19
20 228
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

1
2
3 229 **DISCUSSION**
4
5
6

7 230 The main aim of this study was to assess the strength of the association between individual and
8
9 231 ecological deprivation measures. Overall, the results show that aggregated area-level deprivation
10
11 232 metrics are weak predictors of individual-level deprivation status in the cancer patient cohort
12
13 233 analysed here. Area-level income displayed a particularly weak association with individual-level
14
15 234 income status; whereas area-level occupation, and, to a lesser extent education, appear to have
16
17 235 slightly stronger relationships with individual-level measures. These results have important and
18
19 236 wide-ranging implications for the interpretation of studies that examine the impact of deprivation
20
21 237 on health outcomes, particularly those that form the basis of policies aimed at addressing
22
23 238 inequalities. If aggregated area-level deprivation metrics do not fully represent socio-economic
24
25 239 variation, then policies based on these measures risk misunderstanding the relationship between
26
27 240 health and deprivation.
28
29
30
31
32

33 241 The calculation of the IMD income domain is based on the proportion of individuals in an area
34
35 242 eligible for low-income tax credits or benefits. It is therefore principally an estimator of the
36
37 243 distribution of very low incomes, and provides relatively little information about the distribution of
38
39 244 mid- to high-incomes. On the other hand, the individual-level income estimation method we used
40
41 245 generates a continuous scale of income, the quintiles of which separate individuals with higher
42
43 246 incomes from middle and lower incomes more effectively. An additional consideration is the
44
45 247 calculation of an individual's income, which is not directly collected as part of census data in the UK
46
47 248 and we therefore had to use an estimation method [29]. While this method is validated on UK data,
48
49 249 it is nonetheless likely to introduce a degree of error, and perhaps especially so for those individuals
50
51 250 managing periods of insecure employment or unemployment, whose occupations will be the least
52
53 251 well-documented in the census. As such, ecological and individual metrics quantify income variation
54
55 252 in different ways and might not be expected to closely correlate with one another. Income
56
57 253 deprivation carries a major weight in the calculation of the IMD for area-level statistics, but our
58
59
60

1
2
3 254 analyses show that it is not straightforward to translate this to individual circumstances.
4
5 255 Differentially targeting healthcare funding towards the poorest communities, based on area-level
6
7 256 income metrics, is a sensible policy with important potential benefits in terms of reducing
8
9
10 257 inequalities, but it is nonetheless also important to recognise that this could overlook some
11
12 258 individuals, and perhaps especially those with low income but not in the lowest income bracket.
13
14
15
16 259 For occupation, the area-level IMD domain is based on the proportion of unemployment in an area.
17
18 260 In our individual-level data, unemployed individuals were treated as missing data [27] and would
19
20 261 therefore have been categorised by proxy (wherever possible) using the occupational category of
21
22 262 another adult in the same household. This approach makes an imperfect assumption that the type of
23
24 263 occupation of an unemployed individual can be approximated by the occupation of another adult in
25
26 264 the same household (usually a spouse or partner). However, the relatively good predictive accuracy
27
28 265 of area- and individual-level occupation variables in our results suggests that there is a fair degree of
29
30 266 geographic clustering of levels of unemployment and occupation types. Interestingly, the association
31
32 267 between individual and ecological occupation measures was not affected by a sensitivity analysis we
33
34 268 carried out with unemployed individuals included in the analysis as part of the *technical, routine and*
35
36 269 *manual* group, which could be explained by levels of unemployment being highest in these types of
37
38 270 jobs [33].
39
40
41
42
43
44 271 Our results showed that the ability of area-level education to predict individual status was similar to
45
46 272 occupation, although slightly lower. In the case of education, the area-level IMD domain represents
47
48 273 the proportion of people in an area with no qualifications, which was one of the individual-level
49
50 274 categories we included for education, and this data was directly available from the census. As such,
51
52 275 we might have expected a close association between the two education variables. Although more
53
54 276 closely associated than the respective income metrics, the overall weak association and predictive
55
56 277 power is consistent with the full picture presented by our results that area-level measures only
57
58 278 capture some of the variation in deprivation, and do not fully represent individual deprivation status.
59
60

1
2
3 279 Our results suggest that, at least for cancer patients diagnosed in England and Wales, area-level
4
5 280 statistics are not a good proxy for individual-level deprivation status, indeed for income deprivation
6
7 281 they are only a small improvement upon the toss of a coin. This is somewhat consistent with a
8
9
10 282 recent study of a French population by Bryere *et al* [34], although we generally found slightly lower
11
12 283 predictive power for area-level variables to predict individual-level deprivation. A major difference
13
14 284 between the two analyses is that where Bryere *et al* used data that was a random sample of the
15
16 285 population, we focussed on a cancer patient cohort.

17
18
19
20 286 Data availability has undoubtedly been a limiting factor in the ability of previous research to consider
21
22 287 both area- and individual-level effects of deprivation. Aggregated data is typically more easily
23
24 288 accessible and therefore predominantly features in inequalities research. Our results have
25
26 289 implications for the interpretation of studies that rely solely on area-level measures of deprivation
27
28 290 such as the IMD. These are useful tools for summarising geographic trends, but our results suggest
29
30 291 that caution is needed in terms of extending the interpretation to individual deprivation
31
32 292 circumstances. We are not suggesting that aggregated deprivation statistics should not be used, or
33
34 293 that the use of aggregated data produces unreliable results for the effect of ecological deprivation.
35
36 294 On the contrary, our results show that area- and individual-level health inequalities should be
37
38 295 viewed as independent phenomenon, both of interest, and that their separate effects as well as
39
40 296 their interaction are likely to be important for understanding and reducing socio-economic
41
42 297 differences. For example, further research could address the extent to which inequalities in cancer
43
44 298 outcomes are related to area-level factors such as the availability of health care services and
45
46 299 resources, in comparison to individual-level factors such as symptom awareness and individual
47
48 300 means to access appointments and treatment. Further, establishing whether or not, for instance,
49
50 301 more deprived cancer patients experience better outcomes when living in an affluent area
51
52 302 compared to living in a more deprived area, due to increased availability of health care services and
53
54
55
56
57
58
59
60

1
2
3 303 resources, is integral to fully understanding these differentials and thus the way in which resources
4
5 304 should be deployed to address them.
6
7
8

9 305 Our data suggest, in fact, that where interventions such as cancer symptom awareness campaigns or
10
11 306 screening have been directed at ecologically deprived areas, a significant minority of deprived
12
13 307 patients will have missed out. The policies to reduce health inequalities set out in the NHS long-term
14
15 308 plan [19] are based on research using aggregate measures of deprivation. If the mechanism by which
16
17 309 deprivation affects cancer survival principally functions at an individual level, it follows that such
18
19 310 campaigns may have had limited efficiency. Conversely, if ecological factors are the predominant
20
21 311 driver of inequalities this approach will have had greater traction. The fact that inequalities are not
22
23 312 significantly reducing, even in the context of policy change [13], suggests the latter is, even if only
24
25 313 partially, at work.
26
27
28
29

30
31 314 In conclusion, we have shown that individual and contextual deprivation are not strongly associated
32
33 315 in a cancer patient cohort, and we argue that this shows the potential for individual and contextual
34
35 316 factors to have independent effects on health inequalities. Further research will be important to
36
37 317 disentangle these factors and enable more targeted policy recommendations, especially in terms of
38
39 318 individual-level deprivation effects, which have not received much research attention to date. An
40
41 319 improved understanding of how individual deprivation affects health outcomes has potential to
42
43 320 inform more effective policies to reduce health inequalities.
44
45
46
47

48 321
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 322 **Funding:** This study was supported by a grant from the Economic and Social Research Council
4
5 323 (ES/S001808/1).
6
7
8

9 324 **Ethics approval:** LSHTM Ethics Online Application 14600; approved 01/02/2018.
10
11
12

13 325 **Competing interests:** The authors have no conflicts of interest to declare.
14
15

16 326 **Acknowledgements:** This work makes use of data from the National Cancer Data Repository
17
18 327 prepared by the National Cancer Intelligence Network in association with the National Disease
19
20 328 Registration Service and the cancer registries of England and Wales. The permission of the Office for
21
22 329 National Statistics to use the Longitudinal Study is gratefully acknowledged, as is the help provided
23
24 330 by staff of the Centre for Longitudinal Study Information & User Support (CeLSIUS). CeLSIUS is
25
26 331 supported by the ESRC Census of Population Programme under project ES/R00823X/1. The authors
27
28 332 alone are responsible for the interpretation of the data in this paper. This work contains statistical
29
30 333 data from ONS which is Crown Copyright. The use of the ONS statistical data in this work does not
31
32 334 imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data.
33
34 335 This work uses research datasets which may not exactly reproduce National Statistics aggregates.
35
36 336 We gratefully acknowledge the participation of all members of the ONS Longitudinal Study.
37
38
39
40
41
42

43 337 **Contributorship statement:** All authors contributed to the study design. FCI, AB, IMA and LMW
44
45 338 analysed the data. All authors contributed to the interpretation of the results. FCI, AB, IMA and LMW
46
47 339 prepared the manuscript. All authors commented on and approved the final manuscript.
48
49
50

51 340 **Data sharing statement:** Data are not publicly available but can be accessed via appropriate
52
53 341 application to the ONS Longitudinal Study.
54
55

56
57 342 **References**
58
59
60

- 1
2
3 343 1. Butler DC, Petterson S, Bazemore A, Douglas KA. Use of measures of socioeconomic deprivation in
4
5 344 planning primary health care workforce and defining health care need in Australia. *Aus J Rural*
6
7 345 *Health* 2010, **18**, 199-204
8
9 346
10
11 347 2. Rey G, Jouglu E, Fouillet A, Hemon D. Ecological association between a deprivation index and
12
13 348 mortality in France over the period 1997-2001: variations with spatial scale, degree of urbanicity,
14
15 349 age, gender and cause of death. *BMC Public Health* 2009, **9**, 33
16
17 350
18
19 351 3. Coleman MP, Babb P, Sloggett A, Quinn MJ, De Stavola BL. Socio-economic inequalities in cancer
20
21 352 survival in England and Wales. *Cancer* 2001, **91**, 208-216
22
23 353
24
25 354 4. Coleman MP, Rachet B, Woods LM, Mitry E, Riga M, Cooper N *et al*. Trends and socio-economic
26
27 355 inequalities in cancer survival in England and Wales up to 2001. *Br J Cancer* 2004, **90**, 1367-1373
28
29 356
30
31 357 5. Rachet B, Ellis L, Maringe C, Chu T, Nur U, Quaresma M *et al*. Socio-economic inequalities in cancer
32
33 358 survival in England after the NHS cancer plan. *Br J Cancer* 2010, **103**, 446-453
34
35 359
36
37 360 6. Singh GK, Williams SD, Siahpush M, Mulhollen A. Socio-economic, rural-urban, and racial
38
39 361 inequalities in US cancer survival. *J Cancer Epidemiol* 2011
40
41 362
42
43 363 7. Hastert TA, Beresford SA, Sheppard L, White E. Disparities in cancer incidence and mortality by
44
45 364 area-level socio-economic status: a multi-level analysis. *J Epidemiol Community Health* 2015, **69**,
46
47 365 168-176
48
49 366
50
51 367 8. Hagedoorn P, Vandenneede H, Vanthomme K, Willaert D, Gadeyne S. A cohort study into head
52
53 368 and neck cancer mortality in Belgium (2001-11). *Oral Oncol* 2016, **61**, 76-82
54
55
56
57
58
59
60

- 1
2
3 369
4
5 370 9. Ellis L, Coleman MP, Rachet B. How many deaths would be avoidable if socio-economic
6
7 371 inequalities in cancer survival in England were eliminated ? *Eur J Cancer* 2012, **48**, 270-278
8
9
10 372
11
12 373 10. Woods LM, Rachet B, O'Connell D, Lawrence G, Tracey E, Willmore A *et al*. Large differences in
13
14 374 patterns of breast cancer survival between Australia and England : a comparative study using cancer
15
16 375 registry data. *Int J Cancer* 2009, **124**, 2391-2399
17
18
19 376
20
21 377 11. Rutherford MJ, Ironmonger L, Ormiston-Smith N, Abel GA, Greenberg DC, Lyratzopoulos G *et al*.
22
23 378 Estimating the potential survival gains by eliminating socio-economic and sex inequalities in stage of
24
25 379 diagnosis of melanoma. *Br J Cancer* 2015, **112**, S116-123
26
27
28 380
29
30 381 12. Herbert A, Abel GA, Winters S, McPhail S, Elliss-Brookes L, Lyratzopoulos G. Are inequalties in
31
32 382 cancer diagnosis through emergency presentation narrowing, widening, or remaining unchanged?
33
34 383 Longitudinal analysis of English population-based data 2006-2013. *J Epidemiol Community Health*
35
36 384 2018, **73**, 3-10
37
38
39 385
40
41 386 13. Exarchakou A, Rachet B, Belot A, Maringe C, Coleman MP. Impact of national cancer policies on
42
43 387 cancer survival trends and socio-economic inequalities in England, 1996-2013. *BMJ* 2018, **360**, k764
44
45 388
46
47
48 389 14. Webster TF, Hoffman K, Weinberg J, Vieira V, Aschengrau A. Community- and individual-level
49
50 390 socio-economic status and breast cancer risk: multi-level modelling on Cape Cod, Massachusetts.
51
52 391 *Environ Health Perspect* 2008, **116**, 1125-1129
53
54 392
55
56
57 393 15. Bryere J, Menvielle G, Dejardin O, Launay L, Molinie F, Stucker I *et al*. Neighborhood deprivation
58
59 394 and risk of head and neck cancer: a multilevel analysis from France. *Oral Oncol* 2017, **71**, 144-149
60

- 1
2
3 395
4
5 396 16. Lamont EB, Zaslavsky AM, Subramanian SV, Meilleur AE, He Y, Landrum MB. Elderly breast and
6
7 397 colorectal cancer patients' clinical course: patient and contextual influences. *Med Care* 2014, **52**, 809-
8
9 398 817
10
11 399
12
13 400 17. Finke I, Behrens G, Weisser L, Brenner H, Jansen L. Socio-economic differences and lung cancer
14
15 401 survival – systematic review and meta-analysis. *Frontiers in Oncology* 2018, **8**, 1-20
16
17 402
18
19 403 18. Sloggett A, Young H, Grundy E. The association of cancer survival with four socio-economic
20
21 404 indicators. *BMC Cancer* 2007, **7**, 20
22
23 405
24
25 406 19. NHS. 2019. The NHS Long Term Plan. <https://www.longtermplan.nhs.uk/> Accessed: March 2020
26
27 407
28
29 408 20. Shelton N, Marshall CE, Stuchbury R, Grundy E, Dennet A, Tomlinson A *et al.* Cohort profile: the
30
31 409 Office for National Statistics Longitudinal Study. *Int J Epidemiol* 2019, **48**, 383-384
32
33 410
34
35 411 21. Hattersley L, Creeser R. 1995. Longitudinal Study 1971-1991: History, organisation and quality of
36
37 412 data. HMSO, London.
38
39 413
40
41 414 22. Winkleby MA, Jatulis DE, Frank E, Fortmann SP. Socioeconomic status and health: how
42
43 415 education, income, and occupation contribute to risk factors for cardiovascular disease. *Am J Public*
44
45 416 *Health* 1992, **82**, 816-20
46
47 417
48
49 418 23. Office for National Statistics. 2004. The English Indices of Deprivation Statistical Release.
50
51 419 <https://webarchive.nationalarchives.gov.uk/20100407164233/http://www.communities.gov.uk/archived/general-content/communities/indicesofdeprivation/216309/> Accessed: January 2020
52
53 420
54
55
56
57
58
59
60

- 1
2
3 421
4
5 422 24. StatsWales. 2005. Welsh Indices of Multiple Deprivation 2005.
6
7 423 <https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of->
8
9 424 [Multiple-Deprivation/Archive/WIMD-2005](https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-Multiple-Deprivation/Archive/WIMD-2005) Accessed: January 2020
10
11 425
12
13 426 25. Office for National Statistics. 2015. The English Indices of Deprivation Statistical Release.
14
15 427 <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015> Accessed: January
16
17 428 2020
18
19 429
20
21 430 26. StatsWales. 2014. Welsh Indices of Multiple Deprivation 2014.
22
23 431 <https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of->
24
25 432 [Multiple-Deprivation/Archive/WIMD-2014](https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-Multiple-Deprivation/Archive/WIMD-2014) Accessed: January 2020
26
27 433
28
29 434 27. Office for National Statistics. 2010. The National Statistics Socio-economic Classification: User
30
31 435 Manual. Palgrave Macmillan, London.
32
33 436 <https://www.ons.gov.uk/methodology/classificationsandstandards/otherclassifications/thenationals>
34
35 437 [tistics/socioeconomicclassificationnssecrebasedonsoc2010](https://www.ons.gov.uk/methodology/classificationsandstandards/otherclassifications/thenationalsstatistics/socioeconomicclassificationnssecrebasedonsoc2010) Accessed: January 2020
36
37 438
38
39 439 28. Centre for Longitudinal Study Information and User Support. 2020. Data Dictionary
40
41 440 [https://www.ucl.ac.uk/infostudies/silva-php-](https://www.ucl.ac.uk/infostudies/silva-php-resources/researchProjects/celsius/standalone/index.php)
42
43 441 [resources/researchProjects/celsius/standalone/index.php](https://www.ucl.ac.uk/infostudies/silva-php-resources/researchProjects/celsius/standalone/index.php) Accessed: January 2020
44
45 442
46
47 443 29. Clemens T, Dibben C. A method for estimating wage, using standardised occupational
48
49 444 classifications, for use in medical research in the place of self-reported income. *BMC Medical*
50
51 445 *Research Methodology* 2014, **14**, 59
52
53 446
54
55
56
57
58
59
60

- 1
2
3 447 30. Banks J, Nazroo J, Steptoe A. *The Dynamics of Ageing: Evidence from the English Longitudinal*
4
5 448 *Study of Ageing*. The Institute for Fiscal Studies: London, UK, 2014.
6
7 449
8
9
10 450 31. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Lawrence Erlbaum
11
12 451 Associates: New Jersey, USA, 1988.
13
14 452
15
16 453 32. Janssens ACJW, Martens FK. Reflection on modern methods: revisiting the area under the ROC
17
18 454 curve. *Intl J Epidemiol* 2020, doi: 10.1093/ije/dyz274
19
20
21 455
22
23 456 33. Office for National Statistics. 2018. Unemployment by previous occupation.
24
25 457 [https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/datasets/](https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/datasets/unemploymentbypreviousoccupationunem02)
26
27 458 [unemploymentbypreviousoccupationunem02](https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/datasets/unemploymentbypreviousoccupationunem02) Accessed: March 2020
28
29
30 459
31
32 460 34. Bryere J, Pornet C, Copin N, Launay L, Gusto G, Grosclaude P *et al*. Assessment of the ecological
33
34 461 bias of seven aggregate social deprivation indices. *BMC Public Health* 2017, **17**, 86
35
36
37 462
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

463 **Table 1.** Numbers and percentages of cancer patients included in the analysis, by sex; showing
 464 distribution across deprivation groups at both individual- and LSOA-level and across cancer types.
 465 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

For peer review only

1
2
3 469 **Figure legends**
4
5

6 470 **Figure 1.** Consort diagram describing the dataset linkage and variables used in the analysis, as well as
7
8 471 the flow of LS members through the data processing steps: overall numbers, cancer patient sub-
9
10 472 population filtering, and missing data exclusions. Data source: ONS LS.
11
12

13 473
14
15

16 474 **Figure 2.** Cramer's $V \pm 95\%$ CI for all pairwise combinations of deprivation metrics. Strength of
17
18 475 association is indicated by darker shading for men in top half (green; N=4,826), and women in
19
20 476 bottom half (purple; N=4,721). Data source: ONS LS.
21
22

23 477
24
25

26 478 **Figure 3.** Stacked barplots showing proportions of men and women in each combination of
27
28 479 categories for (A) individual occupation vs. LSOA occupation quintiles; (B) individual education vs.
29
30 480 LSOA education quintiles; and (C) individual income vs. LSOA income quintiles. Data source: ONS LS.
31
32

33 481
34
35

36 482 **Figure 4.** Predictive accuracy of LSOA-level variables to predict deprived/not deprived individual
37
38 483 deprivation status (left); and ROC curves (right) plotted as sensitivity (true positive rate) against 1-
39
40 484 specificity (false positive rate). A/B: occupation; C/D: education; and E/F: income. Dashed lines
41
42 485 indicate LSOA ventile value with maximum predictive accuracy when used as the threshold value to
43
44 486 differentiate between deprived/not deprived, where deprived are those above this threshold. AUC
45
46 487 values are shown next to ROC curves. Data source: ONS LS.
47
48

49 488
50
51
52
53
54
55
56
57
58
59
60

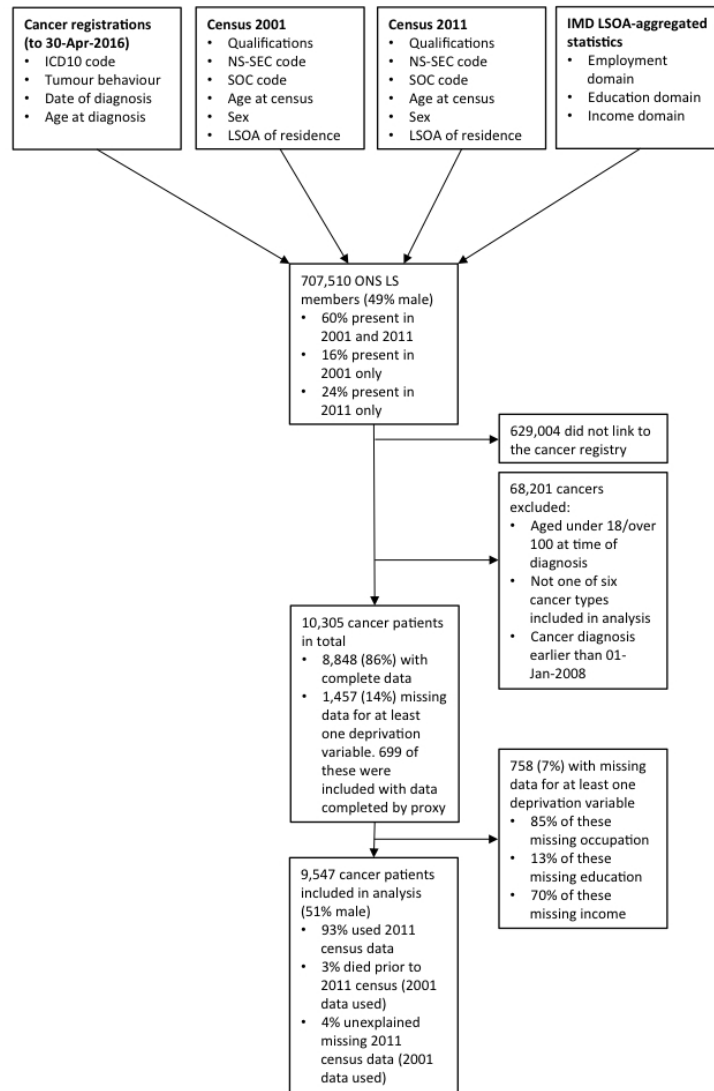


Figure 1. Consort diagram describing the dataset linkage and variables used in the analysis, as well as the flow of LS members through the data processing steps: overall numbers, cancer patient sub-population filtering, and missing data exclusions. Data source: ONS LS.

254x366mm (72 x 72 DPI)

	Occupation (individ)	Education (individ)	Income (individ)	Occupation (LSOA)	Education (LSOA)	Income (LSOA)
Occupation (individ)		0.40 (0.38-0.42)	0.38 (0.36-0.41)	0.18 (0.17-0.22)	0.20 (0.17-0.22)	0.19 (0.16-0.21)
Education (individ)	0.42 (0.39-0.44)		0.24 (0.22-0.27)	0.12 (0.09-0.15)	0.14 (0.11-0.16)	0.12 (0.09-0.15)
Income (individ)	0.55 (0.53-0.57)	0.30 (0.28-0.33)		0.08 (0.05-0.11)	0.09 (0.06-0.12)	0.09 (0.06-0.11)
Occupation (LSOA)	0.15 (0.12-0.18)	0.09 (0.07-0.12)	0.08 (0.05-0.10)		0.47 (0.45-0.49)	0.65 (0.63-0.66)
Education (LSOA)	0.16 (0.14-0.19)	0.12 (0.09-0.14)	0.08 (0.05-0.11)	0.48 (0.46-0.50)		0.49 (0.47-0.51)
Income (LSOA)	0.15 (0.12-0.18)	0.09 (0.07-0.12)	0.07 (0.05-0.10)	0.64 (0.62-0.66)	0.49 (0.47-0.51)	

Figure 2. Cramer's V \pm 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.

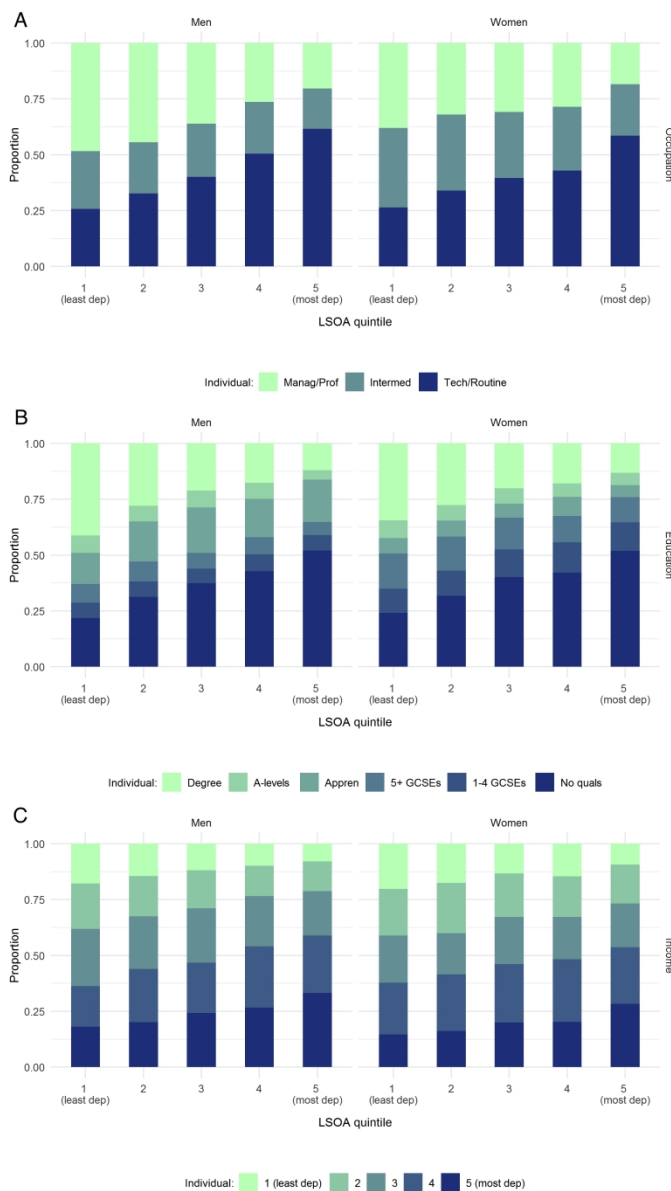


Figure 3. Stacked barplots showing proportions of men and women in each combination of categories for (A) individual occupation vs. LSOA occupation quintiles; (B) individual education vs. LSOA education quintiles; and (C) individual income vs. LSOA income quintiles. Data source: ONS LS.

846x1481mm (72 x 72 DPI)

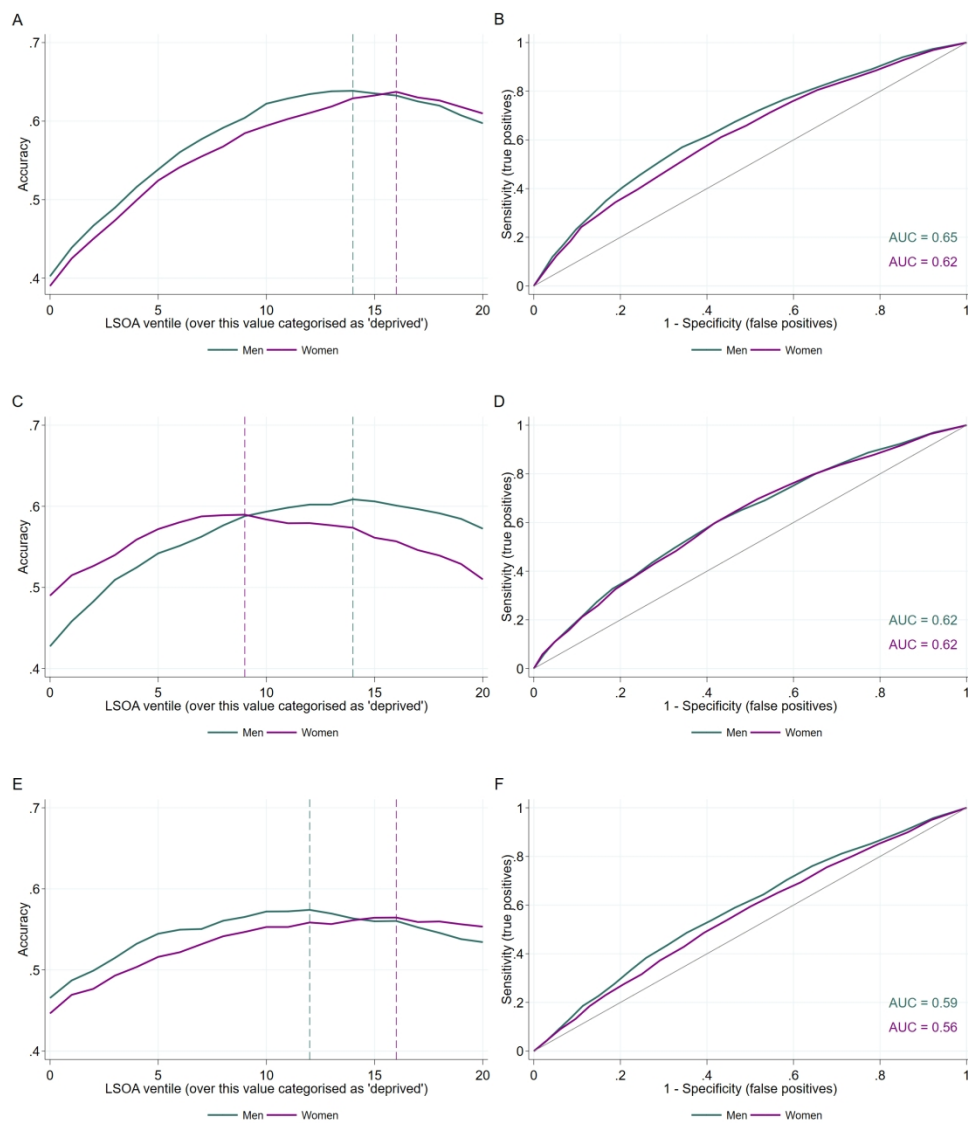


Figure 4. Predictive accuracy of LSOA-level variables to predict deprived/not deprived individual deprivation status (left); and ROC curves (right) plotted as sensitivity (true positive rate) against 1-specificity (false positive rate). A/B: occupation; C/D: education; and E/F: income. Dashed lines indicate 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)

1
2
3 **1 Supplementary Information**
4
5

6 **2 Table S1.** Cramer's $V \pm 95\%$ CI for all pairwise combinations of deprivation metrics – men in top half
7
8 (shaded; N=4516), women in bottom half (unshaded; N=4332). These estimates were generated as a
9
10 sensitivity analysis for the imputation used to complete missing deprivation data by proxy using
11
12 other household adults, therefore these estimates exclude any individuals with imputed data. Data
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
6 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%

21 **Table S4.** Breast cancer (C50) patients only: Cramer's $V \pm 95\%$ CI for all pairwise combinations of
 22 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)	

23
 24 **Table S5.** Colon cancer (C18) patients only: Cramer's $V \pm 95\%$ CI for all pairwise combinations of
 25 deprivation metrics – men in top half (shaded; N=692), women in bottom half (unshaded; N=608).
 26 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)	

27

1
2
3 **Table S6.** Rectal cancer (C19-21) patients only: Cramer's $V \pm 95\%$ CI for all pairwise combinations of
4
5 deprivation metrics – men in top half (shaded; N=521), women in bottom half (unshaded; N=349).
6
7
8 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)	

31
32 **Table S7.** Prostate cancer (C61) patients only: Cramer's $V \pm 95\%$ CI for all pairwise combinations of
33 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)						

34

35 **Table S8.** Bladder cancer (C67) patients only: Cramer’s $V \pm 95\%$ CI for all pairwise combinations of
 36 deprivation metrics – men in top half (shaded; N=395), women in bottom half (unshaded; N=130).
 37 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)	

39 **Table S9.** NHL cancer (C82-86) patients only: Cramer’s $V \pm 95\%$ CI for all pairwise combinations of
 40 deprivation metrics – men in top half (shaded; N=378), women in bottom half (unshaded; N=304).
 41 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

BMJ Open

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

Journal:	<i>BMJ Open</i>
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



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

1
2
3 **1 Assessment of the concordance between individual- and area-level measures of socio-economic**
4
5 **2 deprivation in a cancer patient cohort in England and Wales**
6
7

8
9
3

10
11 **4 Fiona C Ingleby¹, Aurélien Belot¹, Iain M Atherton², Matthew Baker³, Lucy Elliss-Brookes⁴, Laura**
12
13 **5 M Woods¹**
14
15

16
17
6

18
19 7 1. Department of Non-Communicable Disease Epidemiology, Faculty of Epidemiology and Population
20
21 8 Health, London School of Hygiene and Tropical Medicine, London, UK
22

23
24 9 2. School of Health & Social Care, Edinburgh Napier University, Edinburgh, UK
25
26

27 10 3. National Cancer Research Institute Consumer Forum, London, UK
28
29

30 11 4. National Cancer Registration and Analysis Service, Public Health England, London, UK
31
32

33
34
12

35
36 13 Correspondence: fiona.ingleby@lshtm.ac.uk
37
38

39
40
14
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 15 **ABSTRACT**
4
5
6

7 16 **Objectives**
8
9

10 17 Most research on health inequalities uses aggregated deprivation scores assigned to the small area
11
12 18 where the patient lives; however, the concordance between aggregate area-level deprivation
13
14 19 measures and personal deprivation experienced by individuals living in the area is poorly
15
16 20 understood. Our objective was to examine the agreement between individual and ecological
17
18 21 deprivation. We tested the concordance between metrics of income, occupation and education at
19
20 22 individual and area levels, and assessed the reliability of area-based deprivation measures to predict
21
22 23 individual deprivation circumstances.
23
24
25
26
27

28 24 **Setting**
29
30

31 25 England and Wales
32
33
34

35 26 **Participants**
36
37
38

39 27 A cancer patient cohort of 9,547 individuals extracted from the ONS Longitudinal Study.
40
41
42

43 28 **Outcomes**
44
45

46 29 We quantified the concordance between measures of income, occupation and education at
47
48 30 individual and area level. In addition, we used ROC curves and the area under the curve (AUC) to
49
50 31 assess the reliability of area-based deprivation measures to predict individual deprivation
51
52 32 circumstances.
53
54
55
56

57 33 **Results**
58
59
60

1
2
3 34 We found low concordance between individual and area-level indicators of deprivation (Cramer's V
4
5 35 statistics range between 0.07 and 0.20). The most commonly used indicator in health inequalities
6
7 36 research, area-based income deprivation, was a poor predictor of individual income status (AUC
8
9 37 between 0.56 and 0.59), whereas education and occupation were slightly better predictors (AUC
10
11 38 between 0.62 and 0.65). The results were consistent across sexes and across six major cancer types.
12
13
14
15

16 39 **Conclusions**

17
18
19 40 Our results indicate that ecological deprivation measures capture only part of the relationship
20
21 41 between deprivation and health outcomes, especially with respect to income measurement. This has
22
23 42 important implications for our understanding of the relationship between deprivation and health,
24
25 43 and, as a consequence, healthcare policy. The results have a wide-reaching impact for the way in
26
27 44 which we measure and monitor inequalities, and in turn, fund and organise current UK healthcare
28
29 45 policy aimed at reducing them.
30
31
32
33
34
35 46
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 **47 Strengths and limitations of this study:**
4
5
6

7 48 - This study presents a detailed description of concordance between aggregate area-level
8
9 49 deprivation metrics and individual-level deprivation data, enabling an assessment of whether the
10
11 50 widely-used aggregate metrics are actually representative of individual deprivation circumstances or
12
13
14 51 not
15
16

17 52 - The study assesses education, occupation and income indicators of deprivation separately, and
18
19 53 quantifies concordance between individual and area-level measures for each, allowing a more
20
21
22 54 detailed understanding of deprivation than has been possible to date
23
24

25 55 - The cohort focusses on cancer types known to have significant socio-economic inequalities in terms
26
27
28 56 of cancer survival, meaning that extension to a broader population (other cancers or the general
29
30 57 population) would be of interest in future work
31
32

33
34 58 - The data used is the most recent individual deprivation data available from the UK census, and are
35
36 59 therefore limited to year 2011, but once data is available from the planned 2021 census, the results
37
38 60 could be updated
39
40

41
42 61 - A small proportion of individual-level deprivation data was missing and so we completed this
43
44 62 information where possible using another household adult, which could have led to a very small
45
46 63 number of individuals being misclassified
47
48

49
50 64
51
52
53
54
55
56
57
58
59
60

65 INTRODUCTION

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

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

79 The relationship between individual measures, ecological measures and health outcomes is
80 potentially made more complex by the possible existence of contextual effects: that is, that the
81 relationship between individual deprivation and health outcomes might vary by the patient's socio-
82 economic context (ecological deprivation). The degree to which this occurs is likely to depend on the
83 mechanism by which deprivation (either at individual or ecological level) affects outcomes as well as
84 the type of deprivation examined. For example, within oncology a small number of studies have
85 examined the relative effects of individual- and ecological-level deprivation on both cancer risk [14-
86 16] and outcomes [17-19]. Generally, these studies have quantified independent effects of both
87 individual and ecological deprivation, and for both, more deprived areas or individuals have higher
88 risk and lower survival [14, 17-19]. However, the strength and nature of these trends varies

1
2
3 89 considerably across factors including sex, level of geographic aggregation, and which type of
4
5 90 deprivation metric is used [18]. Furthermore, these associations are not well understood in a UK
6
7 91 context, especially in terms of making use of recent data, and an improved understanding will be
8
9 92 important in order to reduce inequalities as part of the NHS long-term plan for 2020-2030 [20]. The
10
11 93 research on health inequalities on which the NHS long-term plan is based uses data aggregated to
12
13 94 small area level, and so improving our understanding of how reliably this matches individual-level
14
15 95 circumstances is important in terms of developing further policies which more specifically target
16
17 96 individual-level variation in health outcomes.
18
19
20
21
22

23 97 Here, we focus on two key research questions: (1) how strong is the concordance between individual
24
25 98 and ecological socio-economic deprivation measures in a cohort of cancer patients; and (2) how
26
27 99 strong is the concordance between different types of deprivation variables? These questions enable
28
29 100 us to comment on the predictive ability of area-level measures to provide information on individual-
30
31 101 level deprivation status in a cancer patient cohort. We discuss the implications of these results in the
32
33 102 context of the existing literature on cancer outcome inequalities.
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

104 **METHODS**

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

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

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

124 **Ecological deprivation metrics**

125 The Indices of Multiple Deprivation (IMD) were used to measure area-based deprivation. The IMD
126 statistics are calculated for each Lower-level Super Output Area (LSOA) in England and Wales and

1
2
3 127 consist of seven domains. We used the income, employment (occupation) and education domains.
4
5 128 LSOA codes were recorded directly for individuals in the 2011 census data, whilst in 2001 census,
6
7 129 LSOA codes were derived from concatenating district and ward codes. The temporally closest data
8
9
10 130 were used for each census: for the 2001 census this was the English IMD2004 [24] and Welsh 2005
11
12 131 report [25], and for the 2011 census this was the English IMD2015 [26] and Welsh 2014 report [27].
13
14 132 Each domain was included as ventiles (i.e. 20 equal quantile groups) of the national distribution of
15
16 133 areas, as opposed to the raw scores, to avoid LS members being identified in LSOAs with low
17
18 134 population size.
19
20
21
22

23 135 **Individual-level deprivation metrics**

24
25
26 136 Individual data on age, sex, qualifications and occupation at the 2011 census were extracted for each
27
28 137 patient, while individual income was derived using a previously published method (see below).
29
30 138 Individual data were not available from the 2011 census for a small proportion of individuals; in part
31
32 139 accounted for by those who were diagnosed with cancer between 2008-2010 and had died prior to
33
34 140 the 2011 census (Figure 1). Where possible, data from the 2001 census was used for these
35
36 141 individuals. For missing data on qualifications or occupation, data was completed where possible by
37
38 142 proxy, using another adult resident in the household (usually household head or spouse). The
39
40 143 rationale for this use of information by proxy is based on evidence that partners tend to have similar
41
42 144 incomes [28], occupations [29] and educational attainment [30]. We tested the sensitivity of the
43
44 145 estimated concordance statistics to this use of proxy data by comparing results with and without
45
46 146 these imputed values, and found very little difference (Table S1). Prior to data completion by proxy,
47
48 147 missingness was 12% for occupation data, 2% for education, and 9% for income. After completion of
49
50 148 missing data by proxy, missingness was 6%, <1%, and 5% respectively for each of occupation,
51
52 149 education, and income individual-level deprivation variables (Figure 1).
53
54
55
56
57
58
59
60

1
2
3 150 Occupation type was derived from the National Statistics Socio-Economic Classification (NS-SEC). The
4
5 151 three-group version of the NS-SEC was used, which categorised LS member occupations as 1)
6
7 152 *technical, routine and manual occupations*; 2) *intermediate occupations*; or 3) *higher managerial,*
8
9 153 *administrative and professional occupations* [31]. Unlike the finer-scaled versions of the NS-SEC, the
10
11 154 three-group version classifies occupations into approximately hierarchical groups. As recommended
12
13 155 for the three-group version of the NS-SEC, those without an occupation classification due to long-
14
15 156 term unemployment or studentship were treated as missing [31]. We carried out a sensitivity
16
17 157 analysis where these individuals were included in the *technical, routine and manual* group, which did
18
19 158 not cause any appreciable differences to the concordance estimates.
20
21
22
23
24

25 159 Education level was categorised as one of six groups based on the standard levels of UK
26
27 160 qualifications used in the census [32]: 1) *no qualifications*; 2) *1-4 GCSEs or equivalent*; 3) *5+ GCSEs or*
28
29 161 *equivalent*; 4) *apprenticeships and vocational qualifications*; 5) *A-levels or equivalent*; or 6) *degree-*
30
31 162 *level education and higher*.
32
33
34

35 163 Weekly income (GBP) was estimated per individual following the method described by Clemens and
36
37 164 Dibben [33], which required information on sex, age, and Standard Occupational Classification (SOC)
38
39 165 code. Income is therefore linked to occupation. The SOC codes used, however, capture specific detail
40
41 166 not available within the NS-SEC codes used for the occupation variable, which more broadly
42
43 167 classifies types of occupation. We took a data-driven approach to adjust income estimates for those
44
45 168 aged over 60 who are most likely to be retired, using observed annualised percentage decreases in
46
47 169 income for those aged over 60 reported by the English Longitudinal Study of Ageing (ELSA [34]; see
48
49 170 Tables S2 and S3). After applying this correction, LS members were grouped into quintiles by
50
51 171 estimated income, from least deprived (Q1) to most deprived (Q5). Quintiles were calculated based
52
53 172 on all available LS members (not just cancer patients), separately for each sex.
54
55
56
57
58

59 173 **Patient and public involvement**
60

1
2
3 174 Due to data protection, we do not have access to individual identifying data from the ONS-LS and so
4
5 175 it was not possible to directly involve these participants in the analyses and discussion for this study.
6
7 176 Our aim is to share these results with patients and public through publication, in order to address
8
9
10 177 public health issues surrounding health inequalities. In addition, we included cancer patient
11
12 178 representatives at each stage of the design, implementation and analysis of this study, as part of the
13
14 179 research team.

18 180 **Data analysis**

19
20
21
22 181 Males and females were analysed separately, for all cancer types combined and for individual
23
24 182 cancers. We tested the degree of concordance between each pairwise combination of the six
25
26 183 deprivation variables: individual-level income quintile, education and occupation groups; and LSOA-
27
28 184 level quintiles for income, education and occupation. Concordance was quantified using Cramer's V
29
30 185 statistic, a measure of the concordance between pairs of categorical variables derived from a chi-
31
32 186 squared statistic, with 95% confidence intervals also approximated from the chi-squared distribution
33
34
35 187 [35]. The measure has the big advantage of not assuming that categories are ordinal. Cramer's
36
37 188 $V < 0.10$ are generally interpreted as low concordance and $V > 0.30$ high, although the values depend
38
39
40 189 in part on the number of categories in the variable with the lowest number of groups (V can be
41
42 190 slightly higher where group numbers are fewer [35]). In most comparisons here, this is the same
43
44 191 (five groups), except for comparisons involving individual-level occupation (three groups).

45
46
47
48 192 For each type of deprivation metric (i.e. education, income or occupation) we assessed the extent to
49
50 193 which the area-level value accurately predicted the 'true' individual-level value. Individuals were
51
52 194 considered 'deprived' if their individual-level value was either *no qualifications* or *1-4 GCSEs*
53
54 195 (education), *technical, routine and manual* (occupation), or below the 40th centile of income
55
56 196 (*quintiles 4 and 5*). A binary classification was applied to the corresponding area-level deprivation
57
58 197 variable, which was repeated using each ventile of the area-level variable as the binary threshold.
59
60

1
2
3 198 For ventile 1 as threshold, individuals in ventiles 2-20 were categorised as deprived; for ventile 2 as
4
5 199 threshold, individuals in ventiles 3-20 were categorised as deprived; and so on. Three aspects of
6
7 200 predictive ability were then measured: (1) accuracy, the total proportion of individuals correctly
8
9 201 classified; (2) sensitivity, the proportion of 'deprived' individuals correctly classified by the area-level
10
11 202 measure; and (3) specificity, the proportion of 'not deprived' individuals correctly classified by the
12
13 203 area-level measure. Using these measures, we generated ROC curves [36] for each type of
14
15 204 deprivation measure and calculated the area-under-curve (AUC) to summarize the ability of the
16
17 205 area-based measure to predict individual-level deprivation.
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

208

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

1
2
3 232 some evidence of higher concordance between variables at the individual level for women than for
4
5 233 men.
6
7
8

9 234 The patterns observed in the overall cancer patient cohort were also observed for each cancer when
10
11 235 examined separately (Tables S4-S9). There was suggestive evidence of higher concordance between
12
13 236 deprivation variables for bladder cancer patients than for other cancer types, but small sample size
14
15
16 237 and wide confidence intervals around the estimates make these results hard to interpret.
17
18

19
20 238
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

1
2
3 239 **DISCUSSION**
4
5
6

7 240 The main aim of this study was to assess the concordance between individual and ecological
8
9 241 deprivation measures. Area-level income displayed particularly low concordance with individual-
10
11 242 level income status; whereas area-level occupation, and, to a lesser extent education, appear to
12
13 243 have slightly higher concordance with individual-level measures. Additionally, the results showed
14
15 244 that aggregated area-level deprivation metrics are weak predictors of individual-level deprivation
16
17 245 status in the cancer patient cohort analysed here. These results have important and wide-ranging
18
19 246 implications for the interpretation of studies that examine the impact of deprivation on health
20
21 247 outcomes, particularly those that form the basis of policies aimed at addressing inequalities. If
22
23 248 aggregated area-level deprivation metrics do not fully represent socio-economic variation between
24
25 249 individuals, then policies based on these measures risk misunderstanding the relationship between
26
27 250 health and deprivation.
28
29
30
31
32

33 251 The calculation of the IMD income domain is based on the proportion of individuals in an area
34
35 252 eligible for low-income tax credits or benefits. It is therefore principally an estimator of the
36
37 253 distribution of very low incomes, and provides relatively little information about the distribution of
38
39 254 mid- to high-incomes. On the other hand, the individual-level income estimation method we used
40
41 255 generates a continuous scale of income, the quintiles of which separate individuals with higher
42
43 256 incomes from middle and lower incomes more effectively. An additional consideration is the
44
45 257 calculation of an individual's income, which is not directly collected as part of census data in the UK
46
47 258 and we therefore had to use an estimation method [33]. While this method is validated on UK data,
48
49 259 it is nonetheless likely to introduce a degree of error, and perhaps especially so for those individuals
50
51 260 managing periods of insecure employment or unemployment, whose occupations will be the least
52
53 261 well-documented in the census. As such, ecological and individual metrics quantify income variation
54
55 262 in different ways and might not be expected to closely match with one another. Income deprivation
56
57 263 carries a major weight in the calculation of the IMD for area-level statistics, but our analyses show
58
59
60

1
2
3 264 that it is not straightforward to translate this to individual circumstances. Differentially targeting
4
5 265 healthcare funding towards the poorest communities, based on area-level income metrics, is a
6
7 266 sensible policy with important potential benefits in terms of reducing inequalities, but it is
8
9
10 267 nonetheless also important to recognise that this could overlook some individuals, and perhaps
11
12 268 especially those with low income but not in the lowest income bracket.
13
14
15

16 269 For occupation, the area-level IMD domain is based on the proportion of unemployment in an area.
17
18 270 In our individual-level data, unemployed individuals were treated as missing data [31] and would
19
20 271 therefore have been categorised by proxy (wherever possible) using the occupational category of
21
22 272 another adult in the same household. This approach makes an imperfect assumption that the type of
23
24 273 occupation of an unemployed individual can be approximated by the occupation of another adult in
25
26 274 the same household (usually a spouse or partner). However, the relatively good predictive accuracy
27
28 275 of area- and individual-level occupation variables in our results suggests that there is a fair degree of
29
30 276 geographic clustering of levels of unemployment and occupation types. Interestingly, concordance
31
32 277 between individual and ecological occupation measures was not affected by a sensitivity analysis we
33
34 278 carried out with unemployed individuals included in the analysis as part of the *technical, routine and*
35
36 279 *manual* group, which could be explained by levels of unemployment being highest in these types of
37
38 280 jobs [37].
39
40
41
42
43

44 281 Our results showed that the ability of area-level education to predict individual status was similar to
45
46 282 occupation, although slightly lower. In the case of education, the area-level IMD domain represents
47
48 283 the proportion of people in an area with no qualifications, which was one of the individual-level
49
50 284 categories we included for education, and this data was directly available from the census. As such,
51
52 285 we might have expected close concordance between the two education variables. Although
53
54 286 concordance is higher than for the respective income metrics, concordance is low overall and the
55
56 287 predictive ability is consistent with the full picture presented by our results that area-level measures
57
58
59
60

1
2
3 288 only capture some of the variation in deprivation, and do not fully represent individual deprivation
4
5 289 status.
6
7
8

9 290 Our results suggest that, at least for cancer patients diagnosed in England and Wales, area-level
10
11 291 statistics are not a very good proxy for individual-level deprivation status, indeed for income
12
13 292 deprivation they are only a small improvement upon the toss of a coin. This is somewhat consistent
14
15 293 with a recent study of a French population by Bryere *et al* [38], although we generally found slightly
16
17 294 lower predictive power for area-level variables to predict individual-level deprivation. A major
18
19 295 difference between the two analyses is that where Bryere *et al* used data that was a random sample
20
21 296 of the population, we focussed on a cancer patient cohort. In particular, the cohort focussed on
22
23 297 cancer types with wide socio-economic inequalities in survival [5], and survival inequalities were of
24
25 298 interest as survival differences can be readily interpreted in terms of healthcare provision and
26
27 299 performance. However, it may be interesting for further research to validate these results on the
28
29 300 overall population cohort in the ONS-LS.
30
31
32
33
34

35 301 Data availability has undoubtedly been a limiting factor in the ability of previous research to consider
36
37 302 both area- and individual-level effects of deprivation. Aggregated data is typically more easily
38
39 303 accessible and therefore predominantly features in inequalities research. Our results have
40
41 304 implications for the interpretation of studies that rely solely on area-level measures of deprivation
42
43 305 such as the IMD. These are useful tools for summarising geographic trends, but our results suggest
44
45 306 that caution is needed in terms of extending the interpretation to individual deprivation
46
47 307 circumstances. We are not suggesting that aggregated deprivation statistics should not be used, or
48
49 308 that the use of aggregated data produces unreliable results for the effect of ecological deprivation.
50
51 309 On the contrary, our results show that area- and individual-level health inequalities should be
52
53 310 viewed as independent phenomenon, both of interest, and that their separate effects as well as
54
55 311 their interaction are likely to be important for understanding and reducing socio-economic
56
57 312 differences. For example, further research could address the extent to which inequalities in cancer
58
59
60

1
2
3 313 outcomes are related to area-level factors such as the availability of health care services and
4
5 314 resources, in comparison to individual-level factors such as symptom awareness and individual
6
7 315 means to access appointments and treatment. Further, establishing whether or not, for instance,
8
9 316 more deprived cancer patients experience better outcomes when living in an affluent area
10
11 317 compared to living in a more deprived area, due to increased availability of health care services and
12
13 318 resources, is integral to fully understanding these differentials and thus the way in which resources
14
15 319 should be deployed to address them.
16
17
18
19

20 320 Our data suggest, in fact, that where interventions such as cancer symptom awareness campaigns or
21
22 321 screening have been directed at ecologically deprived areas, a significant minority of deprived
23
24 322 patients will have missed out. The policies to reduce health inequalities set out in the NHS long-term
25
26 323 plan [20] are based on research using aggregate measures of deprivation. If the mechanism by which
27
28 324 deprivation affects cancer survival principally functions at an individual level, it follows that such
29
30 325 campaigns may have had limited efficiency. Conversely, if ecological factors are the predominant
31
32 326 driver of inequalities this approach will have had greater traction. The fact that inequalities are not
33
34 327 significantly reducing, even in the context of policy change [13], suggests the latter is, even if only
35
36 328 partially, at work.
37
38
39
40
41

42 329 In conclusion, we have shown that individual and contextual deprivation are not highly concordant
43
44 330 with each other in a cancer patient cohort, and we argue that this shows the potential for individual
45
46 331 and contextual factors to have independent effects on health inequalities. Further research will be
47
48 332 important to disentangle these factors and enable more targeted policy recommendations,
49
50 333 especially in terms of individual-level deprivation effects, which have not received much research
51
52 334 attention to date. An improved understanding of how individual deprivation affects health outcomes
53
54 335 has potential to inform more effective policies to reduce health inequalities.
55
56
57
58
59
60

1
2
3 337 **Funding:** This study was supported by a grant from the Economic and Social Research Council
4
5 338 (ES/S001808/1) and partially by a programme grant from Cancer Research UK (grant number
6
7 339 C7923/A18525).

10
11 340 **Ethics approval:** London School of Hygiene and Tropical Medicine Research Ethics Committee: online
12
13 341 application 14600; approved 01/02/2018.

16
17 342 **Competing interests:** The authors have no conflicts of interest to declare.

18
19
20
21 343 **Acknowledgements:** This work makes use of data from the National Cancer Data Repository
22
23 344 prepared by the National Cancer Intelligence Network in association with the National Disease
24
25 345 Registration Service and the cancer registries of England and Wales. The permission of the Office for
26
27 346 National Statistics to use the Longitudinal Study is gratefully acknowledged, as is the help provided
28
29 347 by staff of the Centre for Longitudinal Study Information & User Support (CeLSIUS). CeLSIUS is
30
31 348 supported by the ESRC Census of Population Programme under project ES/R00823X/1. The authors
32
33 349 alone are responsible for the interpretation of the data in this paper. This work contains statistical
34
35 350 data from ONS which is Crown Copyright. The use of the ONS statistical data in this work does not
36
37 351 imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data.
38
39 352 This work uses research datasets which may not exactly reproduce National Statistics aggregates.
40
41 353 We gratefully acknowledge the participation of all members of the ONS Longitudinal Study.

42
43
44
45
46
47 354 **Contributorship statement:** All authors (FCI, AB, IMA, MB, LEB and LMW) contributed to the study
48
49 355 design. FCI, AB, IMA and LMW analysed the data. All authors (FCI, AB, IMA, MB, LEB and LMW)
50
51 356 contributed to the interpretation of the results. FCI, AB, IMA and LMW prepared the manuscript. All
52
53 357 authors (FCI, AB, IMA, MB, LEB and LMW) commented on and approved the final manuscript.

54
55
56
57 358 **Data sharing statement:** Data are not publicly available but can be accessed via appropriate
58
59 359 application to the ONS Longitudinal Study.

360 **References**

- 361 1. Butler DC, Petterson S, Bazemore A, Douglas KA. Use of measures of socioeconomic deprivation in
362 planning primary health care workforce and defining health care need in Australia. *Aus J Rural*
363 *Health* 2010, **18**, 199-204
- 364
- 365 2. Rey G, Jouglu E, Fouillet A, Hemon D. Ecological association between a deprivation index and
366 mortality in France over the period 1997-2001: variations with spatial scale, degree of urbanicity,
367 age, gender and cause of death. *BMC Public Health* 2009, **9**, 33
- 368
- 369 3. Coleman MP, Babb P, Sloggett A, Quinn MJ, De Stavola BL. Socio-economic inequalities in cancer
370 survival in England and Wales. *Cancer* 2001, **91**, 208-216
- 371
- 372 4. Coleman MP, Rachet B, Woods LM, Mitry E, Riga M, Cooper N *et al.* Trends and socio-economic
373 inequalities in cancer survival in England and Wales up to 2001. *Br J Cancer* 2004, **90**, 1367-1373
- 374
- 375 5. Rachet B, Ellis L, Maringe C, Chu T, Nur U, Quaresma M *et al.* Socio-economic inequalities in cancer
376 survival in England after the NHS cancer plan. *Br J Cancer* 2010, **103**, 446-453
- 377
- 378 6. Singh GK, Williams SD, Siahpush M, Mulhollen A. Socio-economic, rural-urban, and racial
379 inequalities in US cancer survival. *J Cancer Epidemiol* 2011
- 380
- 381 7. Hastert TA, Beresford SA, Sheppard L, White E. Disparities in cancer incidence and mortality by
382 area-level socio-economic status: a multi-level analysis. *J Epidemiol Community Health* 2015, **69**,
383 168-176
- 384

- 1
2
3 385 8. Hagedoorn P, Vandenheede H, Vanthomme K, Willaert D, Gadeyne S. A cohort study into head
4
5 386 and neck cancer mortality in Belgium (2001-11). *Oral Oncol* 2016, **61**, 76-82
6
7 387
8
9
10 388 9. Ellis L, Coleman MP, Rachet B. How many deaths would be avoidable if socio-economic
11
12 389 inequalities in cancer survival in England were eliminated ? *Eur J Cancer* 2012, **48**, 270-278
13
14 390
15
16 391 10. Woods LM, Rachet B, O'Connell D, Lawrence G, Tracey E, Willmore A *et al*. Large differences in
17
18 392 patterns of breast cancer survival between Australia and England : a comparative study using cancer
19
20 393 registry data. *Int J Cancer* 2009, **124**, 2391-2399
21
22 394
23
24 395 11. Rutherford MJ, Ironmonger L, Ormiston-Smith N, Abel GA, Greenberg DC, Lyratzopoulos G *et al*.
25
26 396 Estimating the potential survival gains by eliminating socio-economic and sex inequalities in stage of
27
28 397 diagnosis of melanoma. *Br J Cancer* 2015, **112**, S116-123
29
30 398
31
32 399 12. Herbert A, Abel GA, Winters S, McPhail S, Elliss-Brookes L, Lyratzopoulos G. Are inequalities in
33
34 400 cancer diagnosis through emergency presentation narrowing, widening, or remaining unchanged?
35
36 401 Longitudinal analysis of English population-based data 2006-2013. *J Epidemiol Community Health*
37
38 402 2018, **73**, 3-10
39
40 403
41
42 404 13. Exarchakou A, Rachet B, Belot A, Maringe C, Coleman MP. Impact of national cancer policies on
43
44 405 cancer survival trends and socio-economic inequalities in England, 1996-2013. *BMJ* 2018, **360**, k764
45
46 406
47
48 407 14. Lokar K, Zagar T, Zadnik V. Estimation of the ecological fallacy in the geographical analysis of the
49
50 408 association of socio-economic deprivation and cancer incidence. *Int J Environ Res Public Health* 2019,
51
52 409 **16**, 296
53
54
55
56
57
58
59 410
60

- 1
2
3 411 15. Webster TF, Hoffman K, Weinberg J, Vieira V, Aschengrau A. Community- and individual-level
4
5 412 socio-economic status and breast cancer risk: multi-level modelling on Cape Cod, Massachusetts.
6
7 413 *Environ Health Perspect* 2008, **116**, 1125-1129
8
9 414
10
11
12 415 16. Bryere J, Menvielle G, Dejardin O, Launay L, Molinie F, Stucker I *et al.* Neighborhood deprivation
13
14 416 and risk of head and neck cancer: a multilevel analysis from France. *Oral Oncol* 2017, **71**, 144-149
15
16 417
17
18 418 17. Lamont EB, Zaslavsky AM, Subramanian SV, Meilleur AE, He Y, Landrum MB. Elderly breast and
19
20 419 colorectal cancer patients' clinical course: patient and contextual influences. *Med Care* 2014, **52**, 809-
21
22 420 817
23
24 421
25
26 422 18. Finke I, Behrens G, Weisser L, Brenner H, Jansen L. Socio-economic differences and lung cancer
27
28 423 survival – systematic review and meta-analysis. *Frontiers in Oncology* 2018, **8**, 1-20
29
30 424
31
32 425 19. Sloggett A, Young H, Grundy E. The association of cancer survival with four socio-economic
33
34 426 indicators. *BMC Cancer* 2007, **7**, 20
35
36 427
37
38 428 20. NHS. 2019. The NHS Long Term Plan. <https://www.longtermplan.nhs.uk/> Accessed: March 2020
39
40 429
41
42 430 21. Shelton N, Marshall CE, Stuchbury R, Grundy E, Dennet A, Tomlinson A *et al.* Cohort profile: the
43
44 431 Office for National Statistics Longitudinal Study. *Int J Epidemiol* 2019, **48**, 383-384
45
46 432
47
48 433 22. Hattersley L, Creeser R. 1995. Longitudinal Study 1971-1991: History, organisation and quality of
49
50 434 data. HMSO, London.
51
52 435
53
54
55
56
57
58
59
60

- 1
2
3 436 23. Winkleby MA, Jatulis DE, Frank E, Fortmann SP. Socioeconomic status and health: how
4
5 437 education, income, and occupation contribute to risk factors for cardiovascular disease. *Am J Public*
6
7 438 *Health* 1992, **82**, 816-20
8
9
10 439
11
12 440 24. Office for National Statistics. 2004. The English Indices of Deprivation Statistical Release.
13
14 441 [https://webarchive.nationalarchives.gov.uk/20100407164233/http://www.communities.gov.uk/arc](https://webarchive.nationalarchives.gov.uk/20100407164233/http://www.communities.gov.uk/archived/general-content/communities/indicesofdeprivation/216309/)
15
16 442 [hived/general-content/communities/indicesofdeprivation/216309/](https://webarchive.nationalarchives.gov.uk/20100407164233/http://www.communities.gov.uk/archived/general-content/communities/indicesofdeprivation/216309/) Accessed: January 2020
17
18
19 443
20
21 444 25. StatsWales. 2005. Welsh Indices of Multiple Deprivation 2005.
22
23 445 [https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-](https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-Multiple-Deprivation/Archive/WIMD-2005)
24
25 446 [Multiple-Deprivation/Archive/WIMD-2005](https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-Multiple-Deprivation/Archive/WIMD-2005) Accessed: January 2020
26
27
28 447
29
30 448 26. Office for National Statistics. 2015. The English Indices of Deprivation Statistical Release.
31
32 449 <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015> Accessed: January
33
34 450 2020
35
36
37 451
38
39 452 27. StatsWales. 2014. Welsh Indices of Multiple Deprivation 2014.
40
41 453 [https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-](https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-Multiple-Deprivation/Archive/WIMD-2014)
42
43 454 [Multiple-Deprivation/Archive/WIMD-2014](https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-Multiple-Deprivation/Archive/WIMD-2014) Accessed: January 2020
44
45
46 455
47
48 456 28. Nakosteen RA, Zimmer MA. Spouse selection and earnings: evidence of marital sorting. *Economic*
49
50 457 *Inquiry* 2001, **39**, 201-213
51
52
53 458
54
55 459 29. Mansour H, McKinnish T. Same-occupation spouses: preferences or search costs? *J Population*
56
57 460 *Economics* 2018, **31**, 1005-1033
58
59 461
60

- 1
2
3 462 30. Schwartz CR, Mare RD. Trends in educational assortative marriage from 1940 to 2003.
4
5 463 *Demography* 2005, **42**, 621-646
6
7 464
8
9
10 465 31. Office for National Statistics. 2010. The National Statistics Socio-economic Classification: User
11
12 466 Manual. Palgrave Macmillan, London.
13
14 467 <https://www.ons.gov.uk/methodology/classificationsandstandards/otherclassifications/thenationals>
15
16 468 [taticssocioeconomicclassificationnssecrebasedonsoc2010](https://www.ons.gov.uk/methodology/classificationsandstandards/otherclassifications/thenationals) Accessed: January 2020
17
18
19 469
20
21 470 32. Centre for Longitudinal Study Information and User Support. 2020. Data Dictionary
22
23 471 [https://www.ucl.ac.uk/infostudies/silva-php-](https://www.ucl.ac.uk/infostudies/silva-php-resources/researchProjects/celsius/standalone/index.php)
24
25 472 [resources/researchProjects/celsius/standalone/index.php](https://www.ucl.ac.uk/infostudies/silva-php-resources/researchProjects/celsius/standalone/index.php) Accessed: January 2020
26
27
28 473
29
30 474 33. Clemens T, Dibben C. A method for estimating wage, using standardised occupational
31
32 475 classifications, for use in medical research in the place of self-reported income. *BMC Medical*
33
34 476 *Research Methodology* 2014, **14**, 59
35
36 477
37
38
39 478 34. Banks J, Nazroo J, Steptoe A. *The Dynamics of Ageing: Evidence from the English Longitudinal*
40
41 479 *Study of Ageing*. The Institute for Fiscal Studies: London, UK, 2014.
42
43 480
44
45 481 35. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Lawrence Erlbaum
46
47 482 Associates: New Jersey, USA, 1988.
48
49 483
50
51
52 484 36. Janssens ACJW, Martens FK. Reflection on modern methods: revisiting the area under the ROC
53
54 485 curve. *Intl J Epidemiol* 2020, doi: 10.1093/ije/dyz274
55
56 486
57
58
59
60

- 1
2
3 487 37. Office for National Statistics. 2018. Unemployment by previous occupation.
4
5 488 <https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/datasets/>
6
7 489 [unemploymentbypreviousoccupationunem02](https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/datasets/unemploymentbypreviousoccupationunem02) Accessed: March 2020
8
9
10 490
11
12 491 38. Bryere J, Pornet C, Copin N, Launay L, Gusto G, Grosclaude P *et al.* Assessment of the ecological
13
14 492 bias of seven aggregate social deprivation indices. *BMC Public Health* 2017, **17**, 86
15
16
17 493

For peer review only

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%	-	-

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

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

499

For peer review only

1
2
3 500 **Figure legends**
4
5

6 501 **Figure 1.** Consort diagram describing the dataset linkage and variables used in the analysis, as well as
7
8 502 the flow of LS members through the data processing steps: overall numbers, cancer patient sub-
9
10 503 population filtering, and missing data exclusions. Data source: ONS LS.
11
12

13 504
14
15

16 505 **Figure 2.** Cramer's $V \pm 95\%$ CI for all pairwise combinations of deprivation metrics. Strength of
17
18 506 concordance is indicated by darker shading for men in top half (green; N=4,826), and women in
19
20 507 bottom half (purple; N=4,721). Data source: ONS LS.
21
22

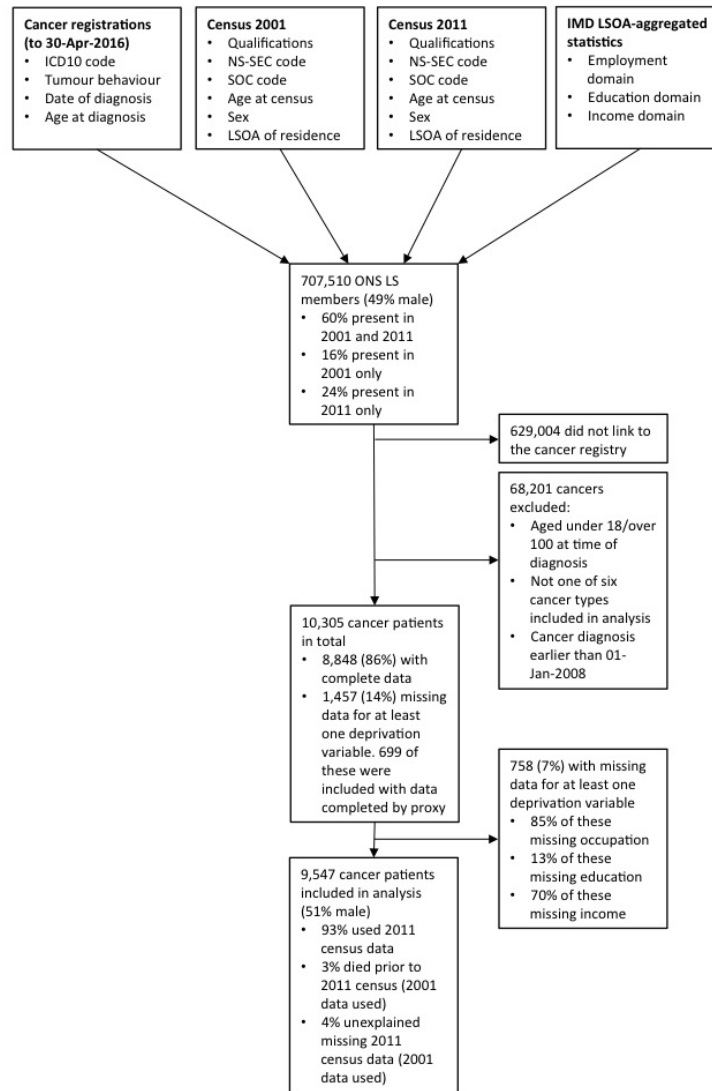
23 508
24
25

26 509 **Figure 3.** Stacked barplots showing proportions of men and women in each combination of
27
28 510 categories for (A) individual occupation vs. LSOA occupation quintiles; (B) individual education vs.
29
30 511 LSOA education quintiles; and (C) individual income vs. LSOA income quintiles. Data source: ONS LS.
31
32

33 512
34
35

36 513 **Figure 4.** Predictive accuracy of LSOA-level variables to predict deprived/not deprived individual
37
38 514 deprivation status (left); and ROC curves (right) plotted as sensitivity (true positive rate) against 1-
39
40 515 specificity (false positive rate). A/B: occupation; C/D: education; and E/F: income. Dashed lines
41
42 516 indicate LSOA ventile value with maximum predictive accuracy when used as the threshold value to
43
44 517 differentiate between deprived/not deprived, where deprived are those above this threshold. AUC
45
46 518 values are shown next to ROC curves. Data source: ONS LS.
47
48

49 519
50
51
52
53
54
55
56
57
58
59
60



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

48 190x275mm (96 x 96 DPI)

	Occupation (individ)	Education (individ)	Income (individ)	Occupation (LSOA)	Education (LSOA)	Income (LSOA)
Occupation (individ)		0.40 (0.38-0.42)	0.38 (0.36-0.41)	0.18 (0.17-0.22)	0.20 (0.17-0.22)	0.19 (0.16-0.21)
Education (individ)	0.42 (0.39-0.44)		0.24 (0.22-0.27)	0.12 (0.09-0.15)	0.14 (0.11-0.16)	0.12 (0.09-0.15)
Income (individ)	0.55 (0.53-0.57)	0.30 (0.28-0.33)		0.08 (0.05-0.11)	0.09 (0.06-0.12)	0.09 (0.06-0.11)
Occupation (LSOA)	0.15 (0.12-0.18)	0.09 (0.07-0.12)	0.08 (0.05-0.10)		0.47 (0.45-0.49)	0.65 (0.63-0.66)
Education (LSOA)	0.16 (0.14-0.19)	0.12 (0.09-0.14)	0.08 (0.05-0.11)	0.48 (0.46-0.50)		0.49 (0.47-0.51)
Income (LSOA)	0.15 (0.12-0.18)	0.09 (0.07-0.12)	0.07 (0.05-0.10)	0.64 (0.62-0.66)	0.49 (0.47-0.51)	

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)

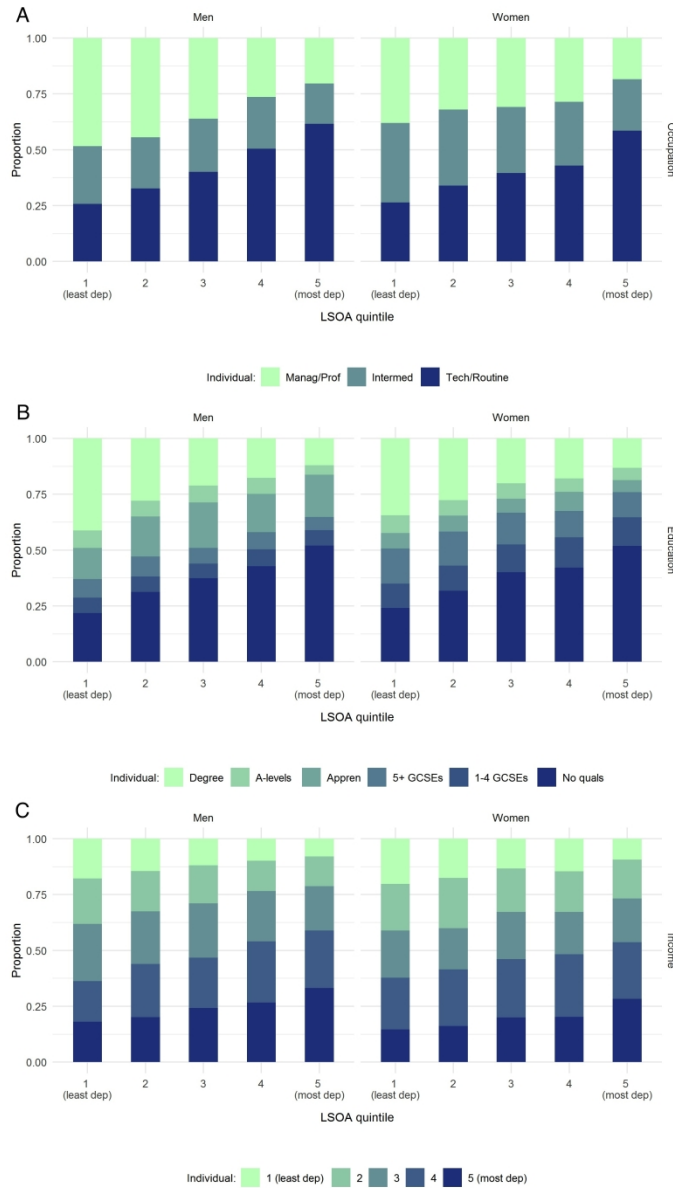


Figure 3. Stacked barplots showing proportions of men and women in each combination of categories for (A) individual occupation vs. LSOA occupation quintiles; (B) individual education vs. LSOA education quintiles; and (C) individual income vs. LSOA income quintiles. Data source: ONS LS.

635x1111mm (96 x 96 DPI)

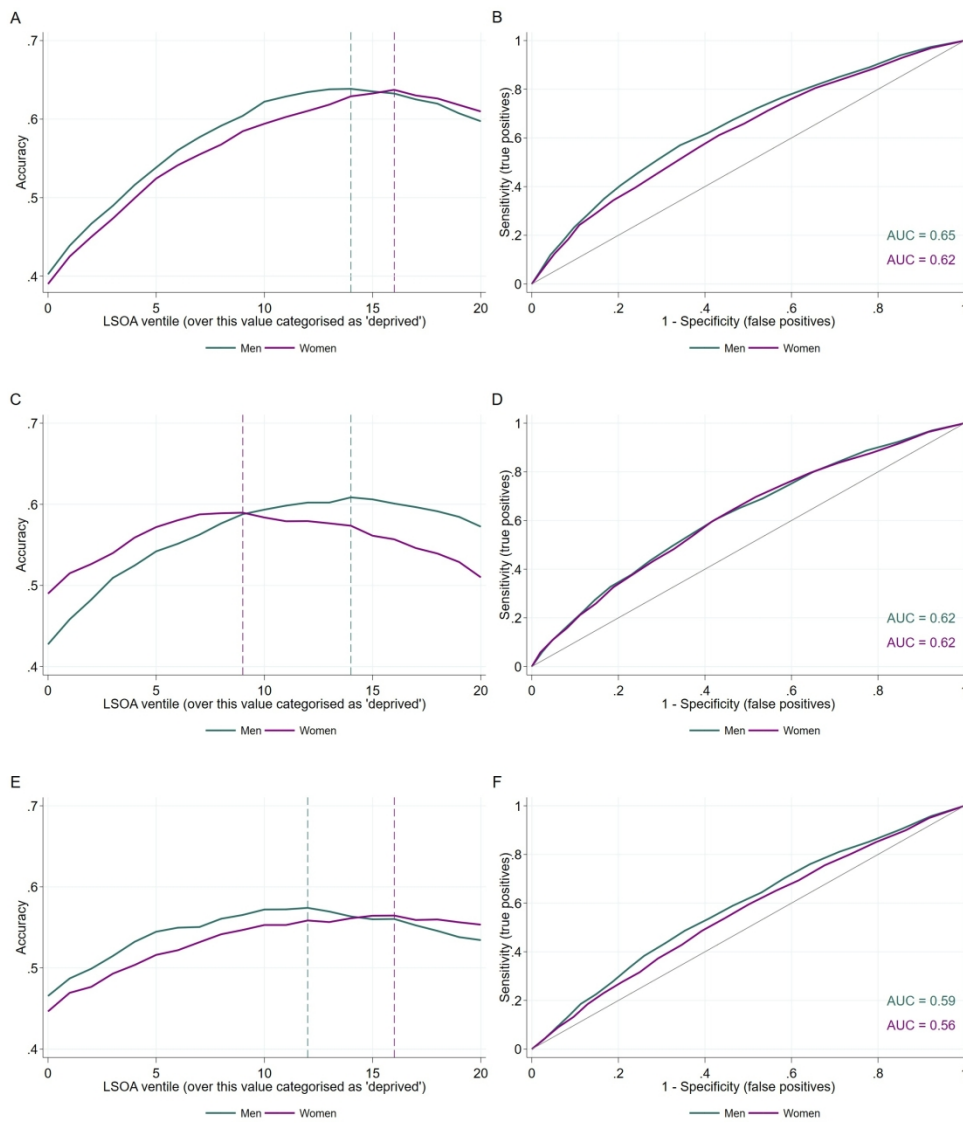


Figure 4. Predictive accuracy of LSOA-level variables to predict deprived/not deprived individual deprivation status (left); and ROC curves (right) plotted as sensitivity (true positive rate) against 1-specificity (false positive rate). A/B: occupation; C/D: education; and E/F: income. Dashed lines indicate 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)

1 **Supplementary Information**

2 **Table S1.** Cramer's $V \pm 95\%$ CI for all pairwise combinations of deprivation metrics – men in top half
 3 (shaded; N=4516), women in bottom half (unshaded; N=4332). These estimates were generated as a
 4 sensitivity analysis for the imputation used to complete missing deprivation data by proxy using
 5 other household adults, therefore these estimates exclude any individuals with imputed data. Data
 6 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%

22 **Table S4.** Breast cancer (C50) patients only: Cramer's $V \pm 95\%$ CI for all pairwise combinations of
 23 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)	

25 **Table S5.** Colon cancer (C18) patients only: Cramer's $V \pm 95\%$ CI for all pairwise combinations of
 26 deprivation metrics – men in top half (shaded; N=692), women in bottom half (unshaded; N=608).
 27 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)	

29 **Table S6.** Rectal cancer (C19-21) patients only: Cramer's $V \pm 95\%$ CI for all pairwise combinations of
 30 deprivation metrics – men in top half (shaded; N=521), women in bottom half (unshaded; N=349).
 31 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)	

33 **Table S7.** Prostate cancer (C61) patients only: Cramer's $V \pm 95\%$ CI for all pairwise combinations of
 34 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)						

36 **Table S8.** Bladder cancer (C67) patients only: Cramer's $V \pm 95\%$ CI for all pairwise combinations of
 37 deprivation metrics – men in top half (shaded; N=395), women in bottom half (unshaded; N=130).
 38 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)	

39
 40 **Table S9.** NHL cancer (C82-86) patients only: Cramer's $V \pm 95\%$ CI for all pairwise combinations of
 41 deprivation metrics – men in top half (shaded; N=378), women in bottom half (unshaded; N=304).
 42 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)	

43

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)