Completeness and representativeness of small area socioeconomic data linked with the UK Clinical Practice Research Datalink (CPRD)

Preveina Mahadevan, Mia Harley, Stuart Fordyce, Susan Hodgson, Rebecca Ghosh, Puja Myles, Helen Booth, Eleanor Axson

ABSTRACT

Background The Clinical Practice Research Datalink (CPRD) holds primary care electronic healthcare records for 25% of the UK population. CPRD data can be linked via practice postcode in England, to area-level socioeconomic status (SES) data including the Index of Multiple Deprivation (IMD), the Carstairs Index and the Townsend Deprivation Index; as well as rural–urban classification (RUC). This study aims to describe the completeness and representativeness of CPRD-linked SES and RUC data.

Methods Patients currently registered at general practices contributing data to the May 2021 snapshots of CPRD GOLD (n=445 587) and CPRD Aurum (n=13 278 825) were used to assess the completeness and representativeness of CPRD-linked SES and RUC data against the UK general population.

Results All currently registered patients had complete SES and RUC data at practice level across the UK. Most English patients in CPRD GOLD (78%), CPRD Aurum (94%) and combined (93%) had SES and RUC data at patient level. Patient-level SES data in CPRD for England were comparable to England’s general population (average IMD decile in CPRD 5.52±0.00 vs 5.50±0.02). CPRD UK practices were on average in more deprived areas than the UK general population (6.06±0.07 vs 5.50±0.02). A slightly higher proportion of CPRD patients and practices were from urban areas (85%) as compared with the UK general population (82%).

Conclusion Completeness of CPRD-linked SES and RUC data is high. The CPRD populations were broadly representative of the general populations in the UK in terms of SES and RUC.

INTRODUCTION

The Clinical Practice Research Datalink (CPRD) databases comprise a sample of primary care electronic healthcare records (EHRs) in the UK for over 62 million historical patients, of whom 16 million are currently registered patients. CPRD collects anonymised patient data from a network of general practices (GPs) across the UK. There are two CPRD primary care databases. CPRD GOLD consists of data sourced from GPs using Vision EHR software, whereas CPRD Aurum consists of data sourced from GPs using EMIS EHR software. CPRD data have been used for public health research for over 30 years, and more than 2900 peer-reviewed publications have used these data. Primary care data from CPRD are linked to a range of other datasets, including small area socioeconomic status (SES) and rural–urban classification (RUC) data to provide a fuller picture of health in the UK.

WHAT IS ALREADY KNOWN ON THIS TOPIC

Clinical Practice Research Datalink (CPRD)-linked area-level socioeconomic status (SES) measures estimate relative deprivation of a patient or general practice based on a variety of socioeconomic measures associated with their small geographical area. It is an essential health determinant indicator and useful for healthcare research. This study assesses the completeness and representation of CPRD-linked area-level SES measures compared with the UK general population.

WHAT THIS STUDY ADDS

Overall, this study confirms that the completeness of the CPRD-linked area-level SES is high, CPRD practice-level SES and rural–urban classification (RUC) data is 100% complete across the UK and most English patients in CPRD GOLD (78%), CPRD Aurum (94%), and combined (93%) had patient-level SES and RUC data. CPRD patient population is broadly representative of the general population in England for patient-level SES, but CPRD practices are from slightly more deprived areas in the UK. The study supports researchers to make appropriate choices on use of small area data for public health research and benefit.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

This study provides advice for researchers using deprivation measures in combination with CPRD. Studies that use patient-level SES as a main exposure, researchers are advised to focus on patients in England from the combined CPRD databases, rather than using UK-wide data and substituting practice SES for patient data where this is absent. Studies wishing to use patient-level SES as a descriptor, stratifying variable and/or covariate should consider whether the practice-level measure is a sufficient proxy for their study design and note this in the methodology and/or limitations of any resulting output.
CPRD-linked area-level SES measures estimate relative deprivation of a patient or GP based on a variety of socioeconomic measures associated with their small geographical area. Different SES measures consider combinations of income, education, employment, access to public/health services, indoor/outdoor environments and proxies of these, such as owning a car. Studies have shown that SES measures are just as important as tobacco use, unhealthy diet, physical inactivity and harmful use of alcohol in predicting health outcomes.1–3 Area-level SES is not just a health determinant but is useful in healthcare research for applications such as planning and targeting of health and social care services and addressing and lowering health inequalities.

This study explored the following objectives: (1) the completeness of the CPRD-linked SES and RUC data; (2) the agreement and correlation between the different patient-level and practice-level CPRD-linked SES and RUC data in England and (3) the representativeness of CPRD-linked SES and RUC data against the general population in the UK, Great Britain (GB), and the individual nations.

The results will enable researchers to further understand the CPRD-linked SES data offerings by describing their completeness, usability, limitations and representativeness. This will help to inform choice of data sources, interpretation of results, and translation of those results into public health improvements for patients.

METHODS

The study population included all currently registered, acceptable patients through May 2021 builds of CPRD GOLD5 and CPRD Aurum.3 These were patients permanently registered at actively contributing practices, excluding transferred out and deceased patients. Over 80% of permanent registrations were deemed to be acceptable for use in research6 (online supplemental figure 1).

Currently, CPRD data are linked to the following area-level SES data7–10: Indices of Multiple Deprivation (IMD) composite and domains, Carstairs Index, Townsend Deprivation Index; as well as RUC. The availabilities of the different CPRD-linked SES and RUC data are outlined in figure 1. As standard, researchers are provided only one patient level and/or one practice level CPRD-linked SES data for a single study, in order to prevent the possibility of deductive disclosure of the location of practices/patients.

CPRD-linked SES and RUC classifications are determined at a small area level associated with either a patient or a GP. All small areas in each UK nation are ranked according to their level of deprivation relative to that of other areas within that nation. The small area levels are: the lower layer super output area (SOA) geographical level in England and Wales (average of 1600 residents and a minimum of 300 households); the SOA level in Northern Ireland; and the data zone level in Scotland.11 The small areas are designed to be similar in population size and social characteristics, such as tenure of household and dwelling type. For approved studies, CPRD can provide linked SES or RUC measures for the small area associated with a patient (patient level) in England or with a GP (practice level) across the UK.

The IMD index is created by the UK Ministry of Housing, Communities and Local Government and uses mainly administrative data, such as benefit records, as well as census data. It provides both social domain-specific and composite scores based on weighted ranks for different social domains: income, education, employment, housing, environment, crime and access to services.12–15

The Townsend Deprivation Index covers all nations of the UK and is derived from weighted measures of unemployment, car ownership, house ownership and household overcrowding from the UK Census.16 The Carstairs Index covers all nations of GB and is also based on measures from the census.17 Unlike IMD, Townsend and Carstairs are comparable across nations of the UK (figure 1).

The RUC is a measure of urbanisation, based on resident population only, and does not reflect the land use, policy or financial characteristics of an area.18–21 The RUC is produced for England and Wales combined, but separate data sources and methodologies are used for Scotland and Northern Ireland making these non-comparable across nations.

Patient and practice small areas are assigned using the most recent postcode recorded for the patient or the practice. Patient postcodes are linked to their small area by National Health Service (NHS) Digital (NHSD), CPRD’s trusted third party, as CPRD does not hold identifiable patient information. Postcodes and small areas are linked using the NHS Postcode Directory (NHSPD).22 SES and RUC are then assigned to patients and practices on weighted ranks for different social domains: income, education, employment, housing, environment, crime and access to services.

The completeness of CPRD-linked SES and RUC data were assessed by calculating the count and proportion of patients...
with a given measure across the CPRD databases individually and combined. Agreement and correlation between different measures at patient level, and between patient and practice levels in England were assessed by calculating the count, Spearman correlation coefficients and proportion of patients with matching SES quintile rank at patient and practice levels. Quintiles were used as these are the most commonly requested quantile breakdown of SES ranks.

Representativeness of CPRD-linked SES and RUC at patient and practice level were compared with the most recently reported general population measures for the relevant UK nations. The mean±SD decile rank for each SES measure in each national population was defined as 5.50±0.02 because the definition of deciles requires that 50% of the national population place above the average and 50% place below the average. Deciles were used for a more detailed picture of representation.

RESULTS
Completeness of SES and RUC data
There were no missing SES and RUC measures in both databases at the practice level for the UK. Across the combined CPRD GOLD and CPRD Aurum databases, 13 724 412 patients were currently registered in England and, among these, 12 805 555 (93%) had at least one patient-level measure (ie, IMD composite and domains, Carstairs, Townsend and/or RUC). A small percentage (6.7%, n=918 857) of currently registered acceptable patients had no patient-level measures available across the combined CPRD GOLD and CPRD Aurum data.

Agreement between CPRD patient-level and practice-level SES and RUC data in England
Among the 12 805 555 acceptable, currently registered patients in England with a patient-level measure in the combined databases, high agreement (r=0.80) between patient-level quintile rankings by each SES metric (IMD composite, Townsend and/or RUC). Agreement between patient-level IMD domains was highest between the income and health domains; the employment and education domains; the health and employment domains; and the housing and living environment domains (r≥0.75; online supplemental figure 2).

Overall, fewer than 45% of patients in England were assigned to the same SES rank quintile (based on their residence postcode) as their GP practice. This was consistently observed among the different SES measures. The highest levels of agreement between patient and practice ranks were observed between the most deprived ranks (~60%–70%) (figure 3). Patient-level and practice-level SES rankings were only moderately correlated, both within and between SES metrics (r=0.44–0.59; figure 2B). Patient-level and practice-level rankings were only slightly correlated between the different IMD domains (r<0.50) and moderately correlated within the same IMD domain (r=0.51–0.81; online supplemental figure 3).

RUC measures were the most highly correlated measure between patient- and practice-levels (r=0.73; figure 2B) with 98.4% of patients classified as urban were registered at a practice also classified as urban (figure 3).

Representativeness of the CPRD-linked SES and RUC data
The average decile deprivation ranks of CPRD practices in all geographies were more deprived compared with the average decile deprivation ranks of the general populations in each geography (figure 4A,B).

The average decile deprivation ranks of CPRD patients in England were comparable to the defined average ±SD decile deprivation rank of England’s general population (5.50±0.02) for all SES measures and in each database (figure 5A).

There was a higher percentage of urban CPRD practices in England, Northern Ireland and Wales as compared with the total percentage of urban practices in these nations; whereas, CPRD practices in Scotland were more rural compared with the total percentage for Scottish practices (online supplemental figure 4).

In England, CPRD patients lived in predominantly urban areas, broadly similar to the overall English profile (85% in urban areas CPRD vs 83% nationally) (figure 5B).

A) Correlation of SES rankings within patients

<table>
<thead>
<tr>
<th></th>
<th>IMD Composite Patient-Level</th>
<th>Carstairs Patient-Level</th>
<th>Townsend Patient-Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMD Composite Patient-Level</td>
<td>1.00</td>
<td>0.86</td>
<td>0.80</td>
</tr>
<tr>
<td>Carstairs Patient-Level</td>
<td>0.86</td>
<td>1.00</td>
<td>0.91</td>
</tr>
<tr>
<td>Townsend Patient-Level</td>
<td>0.80</td>
<td>0.91</td>
<td>1.00</td>
</tr>
</tbody>
</table>

B) Correlation of SES rankings and RUC between patient- and practice-levels

<table>
<thead>
<tr>
<th></th>
<th>IMD Composite Practice-Level</th>
<th>Carstairs Practice-Level</th>
<th>Townsend Practice-Level</th>
<th>RUC Practice-Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMD Composite Patient-Level</td>
<td>0.57</td>
<td>0.54</td>
<td>0.48</td>
<td>NA</td>
</tr>
<tr>
<td>Carstairs Patient-Level</td>
<td>0.53</td>
<td>0.59</td>
<td>0.57</td>
<td>NA</td>
</tr>
<tr>
<td>Townsend Patient-Level</td>
<td>0.44</td>
<td>0.53</td>
<td>0.58</td>
<td>NA</td>
</tr>
<tr>
<td>RUC Patient-Level</td>
<td>NA</td>
<td>0.73</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2 Correlation between socioeconomic status (SES) rankings (A) within patients and (B) between patients and their practice. Correlations between the Index of Multiple Deprivation (IMD) Composite, Carstairs Index and Townsend Deprivation Index both within patients and between patients and their practice. Correlation for rural–urban classification (RUC) between patients and their practice. Correlations in the Clinical Practice Research Datalink (CPRD) acceptable patient population from England only (CPRD GOLD and CPRD Aurum, combined) in May 2021. NA, not applicable.
DISCUSSION

This study provides further insights into the completeness and representativeness of CPRD-linked area-level SES and RUC data. Overall, SES data for patients and practices linked to CPRD GOLD and CPRD Aurum data was broadly similar to the UK population at the time of analysis.

Completeness and correlation of CPRD-linked SES data

Area-level SES and RUC measures could be assigned to 100% of the practice postcodes for GPs registered with CPRD GOLD and CPRD Aurum and area-level SES and RUC measures could be assigned to approximately 93% of patients postcodes in England across the CPRD GOLD and CPRD Aurum databases combined. The availability of patient-level measures depends on the following factors: practice consent to participate in the CPRD patient-level linkage scheme; patients must not have a record indicating dissent for transmission of their personal confidential data to NHSD; and a full valid postcode of residence recorded for the patient in primary care.8 10 Missing data at the patient level is a result of these additional criteria not being met.

There was high correlation between the patient-level SES data between the IMD composite, Townsend, and Carstairs measures as well as between some patient-level IMD domains. As these measures are highly correlated and the provision of multiple

Figure 3 Proportion of patients with the same or different deprivation level as their practice per the (A) Index of Multiple Deprivation (IMD) Composite, (B) Carstairs Index and (C) Townsend Deprivation Index, or (D) rural–urban classification (RUC) in the Clinical Practice Research Datalink (CPRD) acceptable patient population from England only (CPRD GOLD and CPRD Aurum, combined) in May 2021. Colouring shows the level of agreement (%) between patient-level rank and practice-level rank.
measures increases the risk of reidentification of patients and practices, most studies would be served by including only one patient-level and/or one practice-level area-level measure to assess deprivation. When selecting a SES measure for a study, researchers should consider variables used to derive SES ranks for example, variables related to material deprivation and external validity, allowing results to be most comparable with other published work.

Poor agreement between SES data at patient level and practice level
The overall agreement and correlation between practice-level and patient-level SES was lower compared with the within patient-level measures. Studies in the past have used CPRD practice-level SES data as a proxy for missing patient-level data, however, this may introduce bias. Since January 2015, GP practices in England have been free to register new patients who live outside their practice boundary area. Therefore, the lower agreement may be due to patients registering at a practice near their place of work or education, rather than near their home. Another factor can be due to a GP having their main site postcode held by CPRD but have associated branch practices contribute data to CPRD using the same main site postcode. Where these branched practices are in different geographical areas, the SES of these branch practice patients may not correlate with the SES of the main site.

Historic patient postcodes are not maintained in GP records or held by CPRD, and therefore, all linkages to small areas conducted by NHSD using the NHSPD are based on the currently recorded postcode for a patient. It cannot be known if the postcode in the GP records reflects the patient’s current or historical residing address. Thus, some assumptions must be made when utilising patient area-level SES measures: (1) the postcode in their GP record is current and may not reflect historical exposure and (2) patient experiences the same deprivation as their small area average.

Researchers should consider whether a patient area-level SES measure is a suitable proxy for an individual-level SES measure in their studies. Researchers are asked to consider how these assumptions may affect their results and to note these in the methodology and/or limitations of any resulting output.

When using practice-level data to estimate a patient’s SES, one must assume that the patient experiences the same level of deprivation as the small area in which their GP’s main site is located. Studies where patient area-level SES is a main exposure may wish to limit their study population to patients in England with patient-level SES measures, using both databases combined to minimise missingness. Studies wishing to use patient area-level SES as a descriptor, stratifying variable and/or covariate may judge whether the practice area-level measure is a sufficient proxy for their purposes; however, it is recommended that this be noted in the methodology and/or limitations of any resulting output. To further investigate this, researchers may wish to conduct sensitivity analysis using patient-level and practice-level SES to see the impact on the resulting outputs.

In contrast to the SES measures, the correlation and agreement between the practice-level and patient-level RUC were high. In most situations, practice-level RUC could be used as a proxy for patient RUC, as judged by the investigators. Again, it is recommended that the use of practice-level RUC as a proxy for patient-level RUC is noted in the methodology and/or limitations of any resulting output.

CPRD practices are in more deprived areas of the UK
This study confirms that the average CPRD-linked patient-level SES in CPRD GOLD and CPRD Aurum databases, individually and combined, are similar to that of the general population in England.

At practice level, the SES data linked via CPRD GOLD and CPRD Aurum databases, combined, are more deprived than their respective geographies: England, Northern Ireland, Scotland and Wales. According to Woll et al., the CPRD Aurum patient population was slightly less deprived in comparison to the English population. Since that publication in September 2018, CPRD has recruited more practices from more deprived areas in England into the CPRD Aurum database, with the average IMD.

Figure 5 (A) Representativeness of patient-level socioeconomic status (SES) measures—Indices of Multiple Deprivation (IMD) Composite, Carstairs Index and Townsend Deprivation Index—in the Clinical Practice Research Datalink (CPRD); CPRD GOLD and CPRD Aurum, individually and combined, in terms of average decile of deprivation ±SD for all acceptable, currently registered patients in England in May 2021. (B) Representativeness of patient-level rural–urban classification (RUC) in CPRD (CPRD GOLD and CPRD Aurum, individually and combined) in terms of proportion (%) rural and urban for all acceptable, currently registered patients in England in May 2021 compared with the national measure.


Original research
composite of these recruited practices being 6.51 (SD ±0.12). This has led to slightly more GP practices in deprived areas in CPRD Aurum compared with the national distribution of GPs. The completeness and the representativeness of CPRD data may continue to change over time as CPRD recruits new practices. Thus, it will be important for future public health research to repeat these analyses and compare findings.

Limitations
In this descriptive study, IMD and RUC were aggregated to the GB and UK geographies; however, it is important for researchers to note that the classifications for these measures for each nation are ranked independently within each nation and are not directly comparable between nations (figure 1). IMD and RUC should not be used for between nation comparisons in non-descriptive analyses without adjustment.28

While CPRD does maintain complete IMD and RUC data for all practices, due to technical processing some database builds may output with a small number of practices missing IMD or RUC data. This study used the latest CPRD-linked SES and RUC measures SES measures have been persistent over time and the updated metrics are highly correlated to historic metrics.12 For example, the Spearman’s rank correlation coefficient for quintiles, is 0.97, between the IMD composite classifications English IMD 2010 and 2015 and 0.98 between Scottish IMD 2009 and 2012 and 0.93 between NI 2010 and 2015 and 0.97 between Scottish IMD 2009 and 2011 and 0.98 between Welsh IMD 2011 and 2014 (data not shown).23

As discussed above, sometimes patients will not be linked to SES or RUC data for any number of reasons, resulting in missing data. Researchers may employ several methodologies to address such missing data, including complete-case analysis, multiple imputations, maximum likelihood-based formulations, full Bayesian models and weighting methods.29–32

This study investigated the completeness and representativeness of the IMD composite measure. For further understanding of the representation of this deprivation measure, researchers can investigate this at the more granular level of IMD domains. The IMD composite measure is derived from of a number of indicators covering different aspects ('domains') of material deprivation. CPRD can provide quintiles of a specific IMD domain at the level of the representation of this deprivation measure, researchers can use this information to assemble the data used in this study. For further information, please contact the authors in the first instance.

Competing interests
PMa, SF, SH, RG, PMy, HB and EA declare that this work was conducted during their current employment at the CPRD. MH declares that this work was conducted during their previous employment at the CPRD and that, outside of this work, they hold or have held in the last 36 months a doctoral studies stipend from the Medical Research Council to conduct research at the London School of Hygiene and Tropical Medicine.

Patient consent for publication
Not applicable.

Ethics approval
Not applicable.

Provenance and peer review
Not commissioned; externally peer reviewed.

Data availability statement
Data may be obtained from a third party and are not publicly available. This study is based in part on data from CPRD obtained under licence from the UK Medicines and Healthcare products Regulatory Agency (MHRA). The data are provided by patients and collected by the National Health Service (NHS) as part of their care and support. The interpretation and conclusions contained in this study are those of the authors alone. For further information, please contact the authors in the first instance.

Supplemental material
This content has been supplied by the author(s), it has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access
This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: http://creativecommons.org/licenses/by/4.0/.


Bhaskar A, Carstairs index 2011 for Lower-layer super output areas UK Data Service ReShare; 2014. https://reshare.ukdataservice.ac.uk/851497/