Improving our understanding of the social determinants of mental health: A data linkage study of mental health records and the 2011 UK census.

AUTHORS: Cybulski L[1]., Chilman N[1,4]., Jewell A[2]., Dewey M E[1]., Hildersley R[1]., Morgan C[1, 4]., Huck R[3]., Hotopf M[1]., Stewart R[1,2]., Pritchard M[5]., Wuerth M[1]., Das-Munshi J[1,2,4].

Corresponding author: Dr Jayati Das-Munshi, Department of Psychological Medicine, Institute of Psychiatry, Psychology & Neurosciences (IoPPN), De Crespigny Park, London SE5 8AF, United Kingdom. Email: jayati.das-munshi@kcl.ac.uk

Key words: Psychiatric epidemiology; Electronic healthcare records; Data linkage; mental health

AFFILIATIONS:

1. Institute of Psychiatry, Psychology & Neurosciences, King's College London
2. South London & Maudsley NHS Foundation Trust
3. The Data Linkage Hub, Office for National Statistics
4. ESRC Centre for Society and Mental Health, King's College London
5. Norwich Medical School, University of East Anglia, Norwich

Word count: 3911

Abstract

Objectives To address the lack of individual-level socioeconomic information in electronic health care records, we linked the 2011 census of England and Wales to patient records from a large mental healthcare provider. This paper describes the linkage process and methods for mitigating bias due to non-matching.

Setting South London and Maudsley NHS Foundation Trust (SLaM), a mental health care provider in southeast London.

Design Clinical records from SLaM were supplied to the Office of National Statistics (ONS) for linkage to the census through a deterministic matching algorithm. We examined clinical (ICD-10 diagnosis, history of hospitalisation, frequency of service contact) and sociodemographic (age, gender, ethnicity, deprivation) information recorded in CRIS as predictors of linkage success with the 2011 Census. To assess and adjust for potential biases caused by non-matching, we evaluated inverse probability weighting for mortality associations.

Participants Individuals of all ages in contact with SLaM up until December 2019 (N=459,374).

Outcome measures: Likelihood of mental health records' linkage to census.

Results 220,864 (50.4%) records from CRIS linked to the 2011 census. Young adults (Prevalence ratio (PR) 0.80, 95% CI 0.80-0.81), individuals living in more deprived areas (PR 0.78,0.78-0.79), and minority ethnic groups (e.g., Black African, PR 0.67, 0.66-0.68) were less likely to match to census. After implementing inverse probability weighting, we observed little change in the strength of association between clinical/demographic characteristics and mortality (e.g., presence of any psychiatric disorder: unweighted PR 2.66, 95% CI 2.52, 2.80; weighted PR 2.70, 95% CI 2.56, 2.84)

Conclusions Lower response rates to the 2011 census amongst people with psychiatric disorders may have contributed to lower match rates, a potential concern as the census informs service planning and allocation of resources. Due to its size and unique characteristics, the linked dataset will enable novel investigations into the relationship between socioeconomic factors and psychiatric disorders.

Article summary

Strengths and limitations of this study

- This is the first time mental healthcare electronic records have been linked to ONS census at the individual-level in England. Due to its scale, ethnic diversity and demographic characteristics, and abundance of detailed information on a variety of socioeconomic and demographic indicators acquired through the linkage to census records, this dataset will enable novel investigations into the causes, trajectories and outcomes of psychiatric disorders.
- A significant strength of the study is that we could assess and adjust for potential biases caused by non-matching related to age, gender and deprivation.
- Whilst we observed differences between individuals that matched to census, and those that did not, our weighted analyses were able to show that these differences did not substantially alter associations with mortality outcomes.
- Due to the nature of the deterministic linkage algorithm, we could not determine the causes of non-linkage.

Introduction

The growing size and depth of routinely collected administrative data available for research is transforming the study of mental disorders. Traditional epidemiological methods, such as prospective cohort or case-control studies, can present considerable methodological, logistical, and financial challenges due to a high degree of attrition (1), the inherent difficulties in selecting controls (2), and the costs associated with data collection. Electronic health records (EHRs) and other administrative data from public services are therefore increasingly being utilised in epidemiological investigations because they partially address the issue of data loss by collecting information from all individuals who interact with services (3). They also provide a convenient mechanism for sampling controls and eliminate the need for data collection. However, despite their strengths, EHRs typically contain limited information on socioeconomic characteristics at the individual level. Data on occupational classification, long-term unemployment, ethnicity, housing tenure, education, migration, and other relevant socioeconomic measures are often either missing, inaccurate, or collected infrequently, hindering efforts to better understand relationships between mental health and socioeconomic and sociodemographic factors. In prior EHR research, the influence of social determinants has largely been assessed through area-level measures of deprivation, which may not accurately correspond to an individual's socioeconomic circumstances, potentially biasing observed associations and obfuscating inferences that can be made.

To address these issues, we linked clinical records from the South London and Maudsley (SLaM) Mental Health Trust accessed through its Clinical Record Interactive Search (CRIS) platform, to administrative records from the 2011 population census for England and Wales. The modern census of England and Wales, organised and conducted by the Office for National Statistics (ONS) (4), is a rich source of information on a multitude of socioeconomic indicators such as ethnicity, religion, education, employment, housing, migration, and citizenship and also includes self-rated measures of health and functioning. Because of the size and the considerable ethnic diversity of the mental health services' catchment area from which CRIS records are derived, we anticipated that this linkage would facilitate the assessment of several pressing questions on the social determinants of onset, course, and outcomes of severe mental health conditions that have thus far only been examined in case-control and prospective cohort studies limited by small sample sizes and significant attrition.

The purpose of this paper is to describe the creation of this data resource and to outline the methodology employed in linking individual records from the two sources We also sought to describe the cohort's characteristics and to assess how these were associated with successful matches to census records. Finally, to evaluate the potential influence of records not matching on study outcomes, we compared unweighted and inverse probability weighted mortality estimates.

Methods

Data sources used for creating the cohort

CRIS

SLaM provides mental health care to approximately 1.3 million residents in an urban, ethnically diverse, and relatively deprived catchment area comprised of four south London boroughs: Croydon, Lambeth, Lewisham, and Southwark. It is one of Europe's largest mental health care providers and covers all mental health services provided by the National Health Service (NHS), including the Improving Access to Psychological Therapies (IAPT) service, child and adolescent mental health services (CAMHS) and adult mental health, as well as general hospital liaison and various embedded specialist services (e.g., the eating disorders outpatient service). Since 2007, clinical records for all SLaM services have been electronic-only, provided by its electronic Patient Journey System (ePJS) in the form of tick boxes, drop-down lists, free text, and document attachments (3, 5). The CRIS application was developed to enable these records to be used for research within a robust data security and governance framework requiring a combination of data processing pipelines, including de-identification, supplemented by natural language processing (NLP) techniques to provide text-derived metadata (3). Thus, CRIS provides the entirety of a patient's mental health record, including information from structured data fields (e.g., age, sex, diagnosis), but also de-identified free-text information, such as clinical correspondence letters, documents outlining care plans and detentions under the mental health act, and routine clinical notes. Diagnostic data is captured through codes from the 10th edition of the International Classification of Disease (ICD), which may appear in both structured and unstructured data fields.

2011 census data

We utilised the results from the 2011 census of England and Wales as they were the most recent at the time that we initiated this data linkage project. The 2011 census was sent out to every household in England and Wales, and additional measures were taken to ensure the representation of individuals living in communal establishments, such as care homes, prisons, and student halls, and of individuals without a fixed address, such as travellers or rough sleepers (6). The person response rate for the 2011 census was 94%, making it the most comprehensive and representative source of socioeconomic and demographic data in England and Wales (7). Census variables are categorised as 'standard' or 'derived', depending on whether the information they pertain to was explicitly referred to in census questions or derived from respondents responses to other questions (8). For example, 'standard' variables relate to information such as accommodation type, employment status, long-term health problems and disability, caring responsibilities, and religious affiliation, whilst 'derived' variables relate to occupational social class, household deprivation, tenure of household (i.e., rented or owned), degree of educational qualifications, economic activity (i.e., employed, retired, job seeking, etc.,), and family com-

position, and many others. For more information about the census, please see https://www.ons.gov.uk/census/2011census.

Linked dataset creation

We sought access to identifiable information for all individuals who had interacted with SLaM mental health services, including IAPT, up until 31 December 2018. This was done through the Health Research Authority (HRA) by obtaining approval from the Confidential Advisory Group (CAG) to identify patients under Section 251 (9). The reason for seeking access was to enable the linkage of records from CRIS and the 2011 census, which do not have a common identifier (e.g. NHS number) and therefore must be linked through the use of identifiable information, such as name, date of birth, and address. Records from CRIS were then supplied to the ONS, who acted as the trusted linkage function on behalf of the Administrative Data Research Centre for England (ADRCE) and conducted the linkage to the 2011 census. Once records had been matched, identifiable information was removed, and each of the records were given an identifier. The de-identified matched file was then hosted in the ONS secure environment, and accessible only to accredited researchers with project-specific approvals to access the data.

For the present analyses, we report associations between the clinical dataset (CRIS) and the census match 'flag' generated following linkage. We removed observations if they contained erroneous birthdates (e.g., year of birth was 1900), or if individuals had died before the census (23 March 2011) or were born afterwards (Figure 1). Research Ethics Committee (REC) approvals for the establishment of the linked research database were also obtained, which was an approved in addition to the existing REC approvals for CRIS (see Ethical Approvals section below).

Linkage methodology

Records were linked deterministically through a series of matchkeys comprised of information common to both datasets to create unique identifiers. Because a single matchkey might be unable to resolve inconsistencies between data sources, multiple matchkeys were employed. Table 1 summarises each matchkey, the degree to which they uniquely identified records in each dataset, the proportion of CRIS to census matches, and the specific discrepancy they intended to address. For instance, matchkey 2 did not include postcode, thereby allowing records to match on name and date of birth, even if the individual's residence had changed. Matchkeys were by the proportion of unique observations that they identified and required exact matches on all the selected variables. To reduce the risk of false positives, records only linked on a matchkey if it was unique on both datasets. That is, when a record in one dataset matched multiple records in the other dataset, no matches were made, and a new match was instead attempted with the next matchkey in the hierarchy. Once records matched, they were removed from the pool of records eligible to be selected for matching; another match with these records

could therefore no longer be attempted. This means that there was no way to review and unlink matches made earlier in the hierarchy on the basis that the true match was identified at later stages of the matching procedure. Matchkeys 1-11 constitute a set of standard matchkeys that are routinely employed when data owned by the ONS is linked to another dataset (10). We also investigated whether the number of linked records could be increased by attempting further linkage with a set of experimental matchkeys on a randomly selected sample of CRIS data. This additional analysis resulted in matchkey 12.

Measures

We examined an array of routinely recorded sociodemographic and clinical variables in the health record as predictors for successful matching (successful matching denoted through a 'match flag' as described above), including age, sex, ethnicity, marital status, referral date, history of admission to psychiatric hospital, clinical diagnosis by ICD-10 chapter, and frequency of service contact. This information was primarily sourced from structured data fields in the health record (e.g., a drop-down list). Diagnostic information was supplemented by meta-data derived from a bespoke validated NLP algorithm applied to text fields (e.g., clinical correspondence) (3, 11). We classified psychiatric disorder diagnoses according to ICD-10 F chapter headings, with an additional "other diagnoses" category (e.g., "Unspecified mental disorder"). We categorised ethnicity following the '18+1'ONS standard (12), although we merged some categories due to low cell counts. including an aggregation of all mixed ethnicity groups. Similarly, we placed individuals who were married or in a civil union in the same category. Age was calculated by subtracting the date of patients' first recorded contact with services from their birthdates and arranged into 7 age bands (less than 25 years old, 25-34, 35-44, 45-54, 55-64, 65 years or older). We also extracted information on incident inpatient admission. Clinical records in CRIS also store information on death, which is obtained on a monthly basis from the NHS' "Service User Death Report" (13). We used this information to examine mortality as a secondary outcome in order to assess and adjust for potential biases introduced by non-matching. We also explored if outcomes varied by deprivation with the Index of Multiple Deprivation (IMD), an area-level composite measure of deprivation based on income, employment, crime, barriers to housing, health and disability, living environment, and skills and training (14). IMD scores are provided for small geographical areas that correspond to approximately 1,500 individuals, known as a Lower-layer Super Output Area (LSOA). Scores are assigned according to a patient's postcode that was on record closest to the Census date, and placed in quartiles, with a higher score indicating higher levels of deprivation.

Statistical methods

Using the census match flag, we compared linked and unlinked records to better understand which factors were associated with successful linkage between CRIS and Census records. Because odds ratios fail to approximate relative risks when outcomes are common, we estimated prevalence ratios directly through a modified Poisson model with a robust variance estimator following methods outlined by Zou (15). We opted for this method over a log-binomial modelling approach as it addresses the potential issue of model non-convergence (15). We estimated crude prevalence ratios (PR) indicating the association between demographic (e.g., gender, age, ethnicity, neighbourhood deprivation) and clinical characteristics (e.g., psychiatric diagnosis, history of admission) recorded in CRIS and the probability of matching to census records.

Weighted analyses

A potential issue with linking datasets is that not all records will match, and that this might introduce bias if some parameters (e.g., gender) are related to both matching status and outcomes of interest (16). One way of mitigating the influence of biases due to non-matching is through inverse probability weighting (IPW). IPW weights each observation inversely to its probability of being matched so that those which are less likely to be matched receive higher weight (17). Because we had near complete data in CRIS on gender, age, and area-level deprivation, irrespective of matching status, we could assess and adjust for non-matching related to these characteristics by weighting the matched sample. We calculated the probability of matching through a logistic regression by entering match status as the outcome variable (i.e., 1 = matched; 0 = did not match), with age group, gender and deprivation quartile as covariates. These probabilities were then converted into weights using the following formula, with P indicating the probability of matching of the j^{th} observation: $1 - P_j$. We then estimated weighted and unweighted prevalence ratios to measure the association between demographic (e.g., marital status, ethnicity) and clinical variables (i.e., diagnosis of a mental disorder, history of admission, frequency of contact with services etc.,) and all-cause mortality. The weighted and unweighted estimates were adjusted by age, gender, and deprivation quartile.

Results

Cohort characteristics

We identified 459,374 records in CRIS, of which 231,387 (50.4%) matched the 2011 census through matchkeys 1-12 (Table 1). We then applied further exclusion criteria, reducing our matched cohort to 220,864 cases (Figure 1), which is the denominator for all proportions reported below. Just over half of total cohort members were women (54.6%) and the largest ethnic group was White British

(52.9%), followed by Black Caribbean (13.8%) and Black African (4.8%). Nearly two-thirds (65.7%) of cohort members were single and/or separated. The average age of the cohort was 37 (standard deviation: 20).

Predictors of non-linkage

We observed differences within all demographic and clinical categories that we examined as predictors for matching success (Table 2). For sex, men were less likely to match compared with women (PR 0.92, 95% CI 0.91-0.92). Relative to the youngest age group, those aged between 25 and 44 matched less frequently, but conversely, individuals 44 years or older were more likely to match, with the oldest age group (65+) having the highest probability of matching (PR 1.31, 1.29-1.34). Widowed (PR 1.27, 1.25-1.28) and married (PR 1.24, 1.23-1.25) individuals matched more often than those whose who were unmarried. The probability of matching was lower for all minority ethnic groups compared with the White British group, with individuals identifying as White Other or Black African ethnicity the least likely to match. We observed a monotonic relationship between deprivation and matching success, with matching probability decreasing as deprivation increased. Matching success also appeared to vary by referral year, with the highest proportion (59.1%) seen in individuals referred in 2011 (the year of the census), with the next highest in the year after (2012; 57.9%) and before (2010; 55.9%) (Figure 2). Matching success varied by ICD-10 diagnosis (Table 2), with relatively lower rates in individuals diagnosed with mental and behavioural disorders due to psychoactive substance use (F10-F19) or schizophrenia, schizotypal and delusional disorders (F20-F29) (PRs 0.86, 0.85-0.87, and 0.91, 0.89-0.92, respectively), and higher rates in those with Organic mental disorders (F00-F09) (PR 1.38, 1.36-1.40). Similarly, frequent contact with services was associated with a higher probability of matching (1-10 contacts: PR 1.04, 1.04-1.05) compared with individuals without repeated contacts.

Weighted vs. unweighted mortality estimates

Weighted prevalence ratios estimating risk of death tended to be higher for most categories examined compared with unweighted estimates (Table 3); however, the differences were generally very small.

After adjusting for age, gender and deprivation quartile, individuals who were widowed were at the highest risk of death (Table 3). Relative to other minority ethnic groups, the White British ethnic category was associated with the highest risk of death, as indicated by the lower prevalence ratios in all other ethnic groups. However, weighted estimates for the association between ethnicity and all-cause mortality did not vary greatly, compared with unweighted estimates. As can be seen in Table 3, all psychiatric disorders were associated with an increased risk of death, except for behavioural and emotional disorders with onset usually occurring in childhood and adolescence.

Discussion

Summary of results

To our knowledge, this is the first time in which large-scale routine electronic health records from a major secondary mental healthcare provider have been successfully linked to individual-level sociodemographic data from census in England. The resultant dataset draws from an urban and ethnically diverse catchment area from which 220,864 secondary mental healthcare records were linked deterministically to detailed sociodemographic data from the 2011 census of England and Wales**.** Overall, half (50.4%) of records in the secondary mental healthcare dataset linked to 2011 census, and our analyses revealed differences between matched and non-matched records with respect to several sociodemographic and clinical characteristics. We observed the lowest match rates among young adults, individuals living in more deprived areas, and among members of ethnic minority groups. We applied weights to assess how non-matching influenced mortality estimates and observed negligible differences between unweighted and weighted estimates, suggesting that non-linkage to census did not significantly bias associations.

Analysis of records not matching

There are multiple reasons why non-linkage might occur. Firstly, the match rate in our study will have been inherently constrained by the proportion of cases in the CRIS cohort that responded to the 2011 census in the first place. The average response rate within the four London boroughs that comprise the SLaM catchment was lower (88%) compared with the national average (94%) (7). Among younger individuals (25-34 year-olds), who constituted a large proportion of our sample, the response rate was even lower in this region (84%). More mobile populations, which may include migrant and other groups temporarily moving into an area for work alongside people with severe mental illnesses (18), may have been less likely to have taken part in the census. Individuals who moved into the SLaM catchment area and accessed services after 2011 would by default be unable to match. In addition, a growing body of evidence shows that racially minoritised groups, migrants, and other socioeconomically marginalised groups are more likely to face discrimination in their interaction with governmental institutions in the UK, such as the police and the criminal justice system (19, 20), and the NHS (21). It is conceivable that such experiences might coalesce into a general sense of institutional distrust among some members of these communities that is manifested in lower rates of participation. Whatever the cause may be, it would nevertheless seem improbable that our match rate would exceed the average census response rate specific to the SLaM region or the various demographic groups that were prevalent in our sample. It is also well established that unit non-response can be considerable among individuals with a history of mental health disorders, who because of their illnesses might find it challenging to participate (22) or may be more mobile (18). Individuals with mental disorders are also more likely to experience objective social isolation (e.g., have fewer measurable contacts with other individuals) (23) and might consequently be less likely to be captured through proxy responses

(i.e., family members responding in their stead). Indeed, surveys conducted annually since 2004 by the Quality Care Commission (CQC), the independent regulator of healthcare in the UK, have never observed response rates of above 41% in community mental health samples (24).

Another factor that merits consideration is the underlying methodology employed in the matching itself. In our study, records were matched deterministically through matchkeys comprised of administrative information collected in both datasets. Inaccuracies or differences (e.g., wrong post-code, incorrect date of birth, name changes due to marriage, or alternative or erroneous spelling of names) in how these data were recorded might therefore have prevented some records from successfully matching. For example, previous linkage of health records to the census in Scotland highlighted a higher chance of clerical error with respect to the spelling of names for minority ethnic groups, leading to lower match rates (25). As individuals from these groups were preponderant in our cohort, it is possible that clerical error accounted for a degree of non-matching in our study. Moreover, because most matchkeys required postcode information to match and because the match rate peaked among individuals who were referred the year the census was taken, it is possible that the deterministic matching methodology that we employed also missed some individuals who had a different address at the time they interacted with SLaM services and responded to the census. This is supported by higher observed levels of matching (60%) for those with an address recorded in the mental health records at the time of census, in 2011, and is consistent with the interpretation that a high proportion of the sample in this study were potentially more mobile. Comparisons to previous efforts of linking the 2011 Census to other administrative data could help disentangle the relative effects of sample-specific non-participation (e.g., cohort member mobility or non-participation due to mental illness) and issues related to the methodology itself (e.g., sensitivity of matchkeys). However, data linkage methods and the measurement of the linkage quality are continuously evolving within the ONS following the adaptation of new working environments and data sharing agreements, which preclude a fair comparison to other data linkage efforts involving the 2011 Census. Our weighted analyses nevertheless indicated that missingness had a negligible influence on relevant study outcomes, such as associations of clinical/sociodemographic characteristics with all-cause mortality.

Finally, together with existing evidence from cohort studies of substantial attrition among participants diagnosed with mental illnesses, and of non-participation in community surveys, our findings point to non-response being a significant contributor to the low match-rate that we observed. Since the Census informs the planning, funding, and commissioning of local services, such as schools and health services, the potential underrepresentation of individuals with mental illnesses is concerning and merits further investigation.

Strength and weaknesses

We believe that this is the first study to link census data in England to clinical records from a population in contact with secondary mental health care services. Because of the cohort's size, unique sociodemographic composition, and abundant individual-level data on a multitude of important sociodemographic indicators provided by the linkage, we expect this dataset to facilitate novel investigations into health inequalities among people living with mental disorders. The overall size of the cohort is several magnitudes larger than previous UK based mental health cohorts (26), particularly with respect to ethnic minority groups and specific clinical sub-populations (e.g., individuals with severe mental illnesses). The degree of non-linkage that we observed is a potential source of bias. However, we had comprehensive data on many relevant characteristics for the fully enumerated cohort, irrespective of matching status, and could therefore determine through non-response weighting the relative influence that missingness related to these characteristics had, on all-cause mortality estimates. We intend to incorporate these weights in all future analyses to minimise sources of bias. Although the area is ethnically diverse with a good overall representation of Black Caribbean and Black African people, other prevalent ethnic minority groups in England, such as Indian, Pakistani and Bangladeshi populations, are less well represented. Although the highly urban nature of the south London catchment area may be generalisable to other urbanised locations in England, inferences relating to more rural areas may not be possible. There is some evidence that matching of administrative records can be improved through the use of probabilistic techniques (27), but these were not utilised by the ONS for this linkage. It is possible that we could have obtained a higher match rate had record matching been supplemented with probabilistic methods. One of the challenges with the linkage methods employed here is that we could not conclusively determine the exact causes of non-linkage. For instance, we could not quantify the relative degree to which non-linkage was caused by unit non-response or clerical errors in how data was recorded.

Ethical approvals

CRIS has Research Ethics Committee approval as a source of anonymised data for secondary analysis (Oxford REC C, reference 18/SC/0372). The current CRIS-Census linkage was supported through: REC reference for CRIS-Census Linkage: 18/SC/0003. Additional approvals from the Confidential Advisory Group to access patient information without consent, for the purposes of linkage, were obtained (CAG S251 reference: 17/CAG/0204). Approvals were also sought and obtained from the National Statistician's Data Ethics Advisory Committee (NSDEC) for approvals to use linked CRIS-census data for specified projects.

Patient and public involvement

Patient involvement was supported through consultation with the SLaM Clinical Data Linkage Service (CDLS) Data Linkage Service User and Carer Advisory Group, an advisory group of carers and individuals with lived experience of mental illnesses and mental health service use (28), who were consulted at key points during the project. In addition, a CRIS oversight committee which is chaired by a service user, approves all projects proposing to use CRIS-linked data.

Acknowledgements

We are grateful to Hitesh Shetty (SLAM-BRC CRIS team) for his support with data management.

Funding

This paper represents independent research part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. LC and MW are supported by a grant from the ESRC (ES/S002715/1). RH is currently funded by a doctoral studentship granted by the UKRI ESRC LISS-DTP managed by King's College London. JD and CM are part supported by the ESRC Centre for Society and Mental Health at King's College London (ESRC Reference: ES/S012567/1) and by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London and the National Institute for Health Research (NIHR) Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust. MH is a NIHR Senior Investigator. RS is part-funded by: i) the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre at the South London and Maudsley NHS Foundation Trust and King's College London; ii) the NIHR Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust; iii) UKRI – Medical Research Council through the DATAMIND HDR UK Mental Health Data Hub (MRC reference: MR/W014386); iv) the UK Prevention Research Partnership (Violence, Health and

Society; MR-VO49879/1), an initiative funded by UK Research and Innovation Councils, the Department of Health and Social Care (England) and the UK devolved administrations, and leading health research charities. The views expressed are those of the authors and not necessarily those of the ESRC, NHS, the NIHR or the Department of Health and Social Care or King's College London.

Author Contributions

JD conceived the study, designed the work and led acquisition of the linked dataset and interpretation of findings. LC led design, analysis and interpretation of findings. JD and LC led drafting of the manuscript. AJ supported the design and acquisition of the linked dataset, with MP. NC conducted the initial analysis of findings. MED advised on statistical analyses and interpretation. CM, MH, RHu, RHi, MW and RS contributed to the interpretation of findings. All authors were involved in drafting the work or revising it critically prior to submission and all authors approved the final version to be published and agree to be accountable for all aspects of the work.

Conflicts of interest

MH is principal investigator of the RADAR-CNS, a pre-competitive public-private collaboration on mobile health funded by the Innovative Medicine Initiative with cash and in-kind contributions paid to the university from Janssen, Lundbeck, UCB, MSD and Biogen. RS declares research support in the last 3 years from Janssen, GSK and Takeda. All other authors have no conflicts of interest to declare.

Data sharing

Data from SLaM are owned by a 3rd party SlaM BRC CRIS tool which provides access to anonymised data derived from SlaM electronic medical records. These data can only be accessed by permitted individuals from within a secure firewall (i.e., remote access is not possible and the data cannot be sent elsewhere) in the same manner as the authors. Our team is interested in supporting collaboration with interested researchers, subject to appropriate approvals and accreditation status. Requests to access data can be directed to jayati.das-munshi@kcl.ac.uk

Figure legends:

Figure 1. Flow chart illustrating the sample selection process for the census matched/not matched dataset

Figure 2. Proportion of electronic patient records identified via the Clinical Research Interactive Search (CRIS) matched to census by referral year

Table 1. Matchkey composition, uniqueness by dataset, and discrepancy addressed

	Matchkey	Uniqueness by dataset %			Issue addressed by matchkey
		Census	CRIS‡	CRIS to Census match rate (N = 231,387 (%))	
1	Forename, Surname, DOB, Sex, Post-code	100	98.7	87,780 (39.0)	None – exact agreement
2	Forename, Surname, DOB, Sex	99.6	96.3	30,019 (13.0)	Moving out of area
3	Forename Initial, Surname Initial, DOB, Sex, Postcode District	99.9	97.2	43,587 (18.8)	Forename, surname and post-code discrepancy
4	Forename Initial, DOB, Sex, Postcode	99.97	98.3	9,545 (4.1)	Surname discrepancy
5	Surname Initial, DOB, Sex, Postcode	99.9	97.8	5,241 (2.3)	Forename discrepancy
6	Forename, Surname, Sex, Postcode	99.9	98.3	23,635 (10.2)	Date of birth discrepancy
7	Forename bi-gram†, Surname bi-gram, DOB, Sex, Postcode Area	99.8	97.1	17,016 (7.4)	Name discrepancy and moving within area
8	Forename, Surname, Year of Birth, Sex, Postcode District	99.8	97.7	3,073 (1.3)	Date of birth and moving within area
9	First Middle Name, Surname, DOB, Sex, Postcode	99.96	98.2	48 (0.0)	Forename and middle name transpositions
10	Second Middle Name, Surname, DOB, Sex, Postcode	99.96	98.1	12 (0.0)	Forename and second middle name transposition
11	Forename, Surname, DOB, Postcode	100	98.7	902 (0.4)	Sex discrepancy
12	Forename bi-gram, Surname bi-gram, Postcode	93.6	95.8	10,529 (4.6)	Name, sex and date of birth dis-crepancy

† Bi-gram refers to the first two letters of the name

‡ CRIS = Clinical Research Interactive Search

Table 2. CRIS cohort characteristics and their association with census matching

Cohort characteristics N = 420,387	**Matched** N = 220,387 (%)	**Non-matched** N = 199,523 (%)	**Prevalence Ratio** † (95% CI)
Gender			
Female	125,014 (56.6)	104,008 (52.3)	Reference
Male	95,669 (43.3)	95,015 (47.7)	0.92 (0.91, 0.92)
Other	16 (0.1)	26 (0.1)	0.70 (0.47, 1.03)
Marital status ‡			
Single/Separated	86,472 (62.1)	82,129 (70.0)	Reference
Cohabiting	9,519 (6.8)	9,628 (8.2)	0.97 (0.95, 0.98)
Divorced	5,227 (3.8)	4,228 (3.6)	1.07 (1.05, 1.09)
Married/Civil union	30,249 (21.7)	17,139 (14.6)	1.24 (1.23, 1.25)
Widowed	7,862 (5.6)	4197 (3.6)	1.27 (1.25, 1.28)
Age group			
24 and under	76,826 (34.8)	70,351 (35.4)	Reference
25-34	38,248 (17.3)	52,513 (26.4)	0.81 (0.80, 0.81)
35-44	35,197 (16.0)	34,898 (17.5)	0.96 (0.95, 0.97)
45-54	29,481 (13.4)	21,115 (10.6)	1.12 (1.11, 1.13)
55-64	15,837 (7.2)	8,440 (4.2)	1.25 (1.24, 1.26)
65+	25,081 (11.4)	11,685 (5.9)	1.31 (1.30, 1.32)
Ethnicity			
White British	105,578 (60.5)	68,008 (44.2)	Reference
Irish	3,086 (1.8)	3,435 (2.2)	0.78 (0.76, 0.80)
Black Caribbean	22,348 (12.8)	23,023 (15.0)	0.81 (0.80, 0.82)
Black African	8,420 (4.8)	12,141 (7.9)	0.67 (0.66, 0.68)
Indian	3,653 (2.1)	2,906 (1.9)	0.92 (0.90, 0.94)
Pakistani	1,150 (0.7)	1,340 (0.9)	0.76 (0.73, 0.79)
Bangladeshi	721 (0.4)	680 (0.4)	0.85 (0.80, 0.89)
Chinese	801 (0.5)	1,076 (0.7)	0.70 (0.67, 0.74)
Other Asian	3,192 (1.8)	4,024 (2.6)	0.73 (0.71, 0.75)
Other Ethnic	9,546 (5.5)	12,002 (7.8)	0.73 (0.72, 0.74)
Other White	11,488 (6.6)	20,046 (13.0)	0.60 (0.59, 0.61)
Mixed, including other mixed	4,653 (2.7)	5,065 (3.3)	0.79 (0.77, 0.80)
Deprivation quartile §			
1 (least deprived)	62,673 (29.4)	36,748 (20.1)	Reference
2	51,957 (24.3)	46,958 (25.7)	0.83 (0.83, 0.84)
3	50,214 (23.5)	49,178 (26.9)	0.80 (0.80, 0.81)
4 (most deprived)	48,634 (22.8)	49,978 (27.3)	0.78 (0.78, 0.79)
Any psychiatric diagnosis			
No	57964 (26.2)	61632 (30.9)	Reference
Yes	162900 (73.8)	137891 (69.1)	1.12 (1.11, 1.12)
Psychiatric diagnosis by ICD-10 chapter			
No record of diagnosis	57964 (26.2)	61632 (30.9)	Reference
F00-F09 Organic, including symptomatic, mental disorders	13133 (5.9)	6514 (3.3)	1.38 (1.36, 1.40)

F10-F19 Mental and behavioural disorders due to psychoactive substance use	10442 (4.7)	14575 (7.3)	0.86 (0.85, 0.87)
F20-F9 Schizophrenia, schizotypal and delusional disorders	8363 (3.8)	10625 (5.3)	0.91 (0.89, 0.92)
F30-F36 Mood [affective] disorders	44161 (20.0)	36959 (18.5)	1.12 (1.11, 1.13)
F40-F48 Neurotic, stress-related and somatoform disorders	25854 (11.7)	20579 (10.3)	1.15 (1.14, 1.16)
F50-F59 Behavioural syndromes associated with physiological disturbances and physical factors	4965 (2.2)	2989 (1.5)	1.29 (1.26, 1.31)
F60-F69 Disorders of adult personality and behaviour	1312 (0.6)	1505 (0.8)	0.96 (0.92, 1.00)
F70-F79 Mental retardation	640 (0.3)	674 (0.3)	1.00 (0.95, 1.06)
F80-F89 Disorders of psychological development	4545 (2.1)	2298 (1.2)	1.37 (1.35, 1.40)
F90-F98 Behavioural and emotional disorders with onset usually occurring in childhood and adolescence	7060 (3.2)	5092 (2.6)	1.20 (1.18, 1.22)
F99 Unspecified mental disorder	17611 (8.0)	14518 (7.3)	1.13 (1.12, 1.14)
Other diagnoses	24814 (11.2)	21563 (10.8)	1.10 (1.09, 1.12)
History of admission			
No	210,526 (95.3)	187,743 (94.1)	Reference
Yes	10,338 (4.7)	11,780 (5.9)	0.88 (0.87, 0.90)
Face to face contacts			
No contacts	115,430 (52.3)	110,632 (55.4)	Reference
1-10 contacts	67,802 (30.7)	59,442 (29.8)	1.04 (1.04, 1.05)
11+ contacts	37,632 (17.0)	29,449 (14.8)	1.10 (1.09, 1.11)

† Prevalence ratios were unadjusted

‡ The divorced and widowed categories also included civil unions that had ended, whether due to death or legal dissolution of the civil union

§ Deprivation was measured through the Index of Multiple Deprivation

Table 3. Characteristics of census matched CRIS cases and unweighted and weighted prevalence ratios for all-cause mortality

Cohort characteristics (N = 220,864)	**Deceased** (N = 18,363)	**Alive** (N = 202,501)	**Prevalence ratio (95% CI)** † Unweighted	Weighted
Marital status ‡				
Single/Separated	5,078 (32.8)	81,394 (65.7)	Reference	Reference
Cohabiting	147 (1.0)	9,372 (7.6)	0.44 (0.38,0.51)	0.41 (0.35, 0.48)
Divorced	891 (5.8)	4,336 (3.5)	0.83 (0.78, 0.88)	0.82 (0.77, 0.87)
Married	5,140 (33.2)	25,109 (20.3)	0.85 (0.82, 0.88)	0.83 (0.81, 0.87)
Widowed	4,203 (27.2)	3,659 (3.0)	1.15 (1.12, 1.19)	1.14 (1.11, 1.18
Ethnicity				
White British	12,033 (73.3)	93,545 (59.1)	1	Reference
Irish	626 (3.8)	2,460 (1.6)	0.99 (0.93, 1.05)	0.99 (0.93, 1.06)
Black Caribbean	1,322 (8.1)	21,026 (13.3)	0.72 (0.69, 0.75)	0.72 (0.68, 0.75)
Black African	316 (1.9)	8,104 (5.1)	0.62 (0.56, 0.69)	0.63 (0.57, 0.70)
Indian	360 (2.2)	3,293 (2.1)	0.81 (0.74, 0.88)	0.80 (0.74, 0.87)
Pakistani	76 (0.5)	1,074 (0.7)	0.78 (0.64, 0.94)	0.80 (0.65, 0.98)
Bangladeshi	23 (0.1)	698 (0.4)	0.56 (0.39, 0.81)	0.57 (0.39, 0.83)
Chinese	46 (0.3)	755 (0.5)	0.71 (0.56, 0.89)	0.71 (0.56, 0.90)
Other Asian	225 (1.4)	2,967 (1.9)	0.74 (0.66, 0.83)	0.75 (0.67, 0.84)
Other Ethnic	551 (3.4)	8,995 (5.7)	0.83 (0.77, 0.90)	0.82 (0.76, 0.89)
Other White	801 (4.9)	10,687 (6.8)	0.80 (0.76, 0.85)	0.80 (0.75, 0.85)
Mixed, including other mixed	43 (0.3)	4,610 (2.9)	0.32 (0.24, 0.42)	0.30 (0.22, 0.40)
Any psychiatric diagnosis				
No	2980 (9.8)	116616 (29.9)	1	1
Yes	27407 (90.2)	273384 (70.1)	2.66 (2.52, 2.80)	2.70 (2.56, 2.84)
Psychiatric diagnosis by ICD-10 Chapter				
No record of diagnosis	2980 (9.8)	116616 (29.9)	1	1
F00-F09 Organic, including symptomatic, mental disorders	10924 (35.9)	8723 (2.2)	3.25 (3.08, 3.43)	3.32 (3.14, 3.51)

F10-F19 Mental and behavioural disorders due to psychoactive substance use	2784 (9.2)	22233 (5.7)	4.47 (4.16, 4.81)	4.77 (4.43, 5.13)
F20-F9 Schizophrenia, schizotypal and delusional disorders	1889 (6.2)	17099 (4.4)	2.88 (2.66, 3.11)	3.05 (2.81, 3.31)
F30-F36 Mood [affective] disorders	4607 (15.2)	76513 (19.6)	2.23 (2.10, 2.36)	2.23 (2.10, 2.37)
F40-F48 Neurotic, stress-related and somatoform disorders	1288 (4.2)	45145 (11.6)	1.58 (1.47, 1.70)	1.54 (1.43, 1.66)
F50-F59 Behavioural syndromes associated with physiological disturbances and physical factors	123 (0.4)	7831 (2.0)	1.51 (1.22, 1. 85)	1.54 (1.24, 1.90)
F60-F69 Disorders of adult personality and behaviour	149 (0.5)	2668 (0.7)	3.30 (2.66, 4.10)	3.57 (2.85, 4.48)
F70-F79 Mental retardation	137 (0.5)	1177 (0.3)	4.30 (3.35, 5.53)	4.53 (3.49, 5.87)
F80-F89 Disorders of psychological development	60 (0.2)	6783 (1.7)	1.40 (1.01, 1.95)	1.33 (0.95, 1.87)
F90-F98 Behavioural and emotional disorders with onset usually occurring in childhood and adolescence	53 (0.2)	12099 (3.1)	0.85 (0.59, 1.24)	0.88 (0.60, 1.29)
F99 Unspecified mental disorder	1869 (6.2)	30260 (7.8)	2.61 (2.44, 2.79)	2.66 (2.48, 2.85)
Other diagnoses	3524 (11.6)	42853 (11.0)	2.50 (2.35, 2.65)	2.55 (2.40, 2.72)
History of admission				
No	17,207 (93.7)	193,319 (95.5)	1	1
Yes	1,156 (6.3)	9,182 (4.5)	1.43 (1.36, 1.50)	1.49 (1.42, 1.57)
Face to face contacts				
No contacts	3,465 (18.9)	111,965 (55.3)	1	1
1-10 contacts	10,316 (56.2)	57,486 (28.4)	2.42 (2.34, 2.51)	2.52 (2.42, 2.62)
11+ contacts	4,582 (25.0)	33,050 (16.3)	2.56 (2.47, 2.67)	2.68 (2.57, 2.79)

† All models adjusted for age, gender, and deprivation quartile

‡ Civil unions were also included in the divorced, married, and widowed categories

References

1. Knudsen AK, Hotopf M, Skogen JC, Overland S, Mykletun A. The health status of nonparticipants in a population-based health study: the Hordaland Health Study. Am J Epidemiol. 2010;172(11):1306-14.
2. Wacholder S, McLaughlin J L., Silverman D. T., Mandel J. Selection of Controls in Case-Control Studies: I. Principles. Am J Epidemiol. 1992;135(9):1019-28.
3. Perera G, Broadbent M, Callard F, Chang CK, Downs J, Dutta R, et al. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. BMJ Open. 2016;6(3):e008721.
4. Office of National Statistics. The modern census 2016 [Available from: https://www.ons.gov.uk/census/2011census/howourcensusworks/aboutcensuses/censushistory/themoderncensus.
5. Stewart R, Soremekun M, Perera G, Broadbent M, Callard F, Denis M, et al. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. BMC Psychiatry. 2009;9:51.
6. Office of National Statistics. Questionnaires, delivery, completion and return 2016 [04/10/2022]. Available from: https://www.ons.gov.uk/census/2011census/howourcensusworks/howwetookthe2011census/howwecollectedtheinformation/questionnairesdeliverycompletionandreturn.
7. Office of National Statistics. Response and imputation rates 2016 [Available from: https://webarchive.nationalarchives.gov.uk/ukgwa/20160115211827/http:/www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/quality/quality-measures/response-and-imputation-rates/index.html.
8. Office of National Statistics. Variables and classifications 2014 [Available from: https://www.ons.gov.uk/census/2011census/2011censusdata/2011censususerguide/variablesandclassifications.
9. Health Research Authority. Confidentiality Advisory Group 2022 [Available from: https://www.hra.nhs.uk/approvals-amendments/what-approvals-do-i-need/confidentiality-advisory-group/.
10. Office of National Statistics. Beyond 2011: Matching Anonymous Data. 2013.
11. Das-Munshi J, Chang C-K, Dutta R, Morgan C, Nazroo J, Stewart R, et al. Ethnicity and excess mortality in severe mental illness: a cohort study. The Lancet Psychiatry. 2017;4(5):389-99.
12. Office of National Statistics. Ethnic group, national identity and religion 2021 [Available from: https://www.ons.gov.uk/methodology/classificationsandstandards/measuringequality/ethnicgroupnationalidentityandreligion.
13. Roberts E, Wessely S, Chalder T, Chang C-K, Hotopf M. Mortality of people with chronic fatigue syndrome: a retrospective cohort study in England and Wales from the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Clinical Record Interactive Search (CRIS) Register. The Lancet. 2016;387(10028):1638-43.
14. Smith T, Noble M, Noble S, Wright G, McLennan D, Plunkett E. The English Indices of Deprivation 2015. Department for Communities and Local Government; 2015.
15. Zou G. A modified poisson regression approach to prospective studies with binary data. Am J Epidemiol. 2004;159(7):702-6.
16. Hofler M, Pfister H, Lieb R, Wittchen HU. The use of weights to account for non-response and drop-out. Soc Psychiatry Psychiatr Epidemiol. 2005;40(4):291-9.
17. Chesnaye NC, Stel VS, Tripepi G, Dekker FW, Fu EL, Zoccali C, et al. An introduction to inverse probability of treatment weighting in observational research. Clin Kidney J. 2022;15(1):14-20.
18. Lix LM, Hinds A, DeVerteuil G, Robinson JR, Walker J, Roos LL. Residential mobility and severe mental illness: a population-based analysis. Adm Policy Ment Health. 2006;33(2):160-71.

19. Bowling B, Phillips C. Disproportionate and Discriminatory: Reviewing the Evidence on Police Stop and Search. Modern Law Review. 2007;70(6):936-61.
20. Phillips C, Bowling B. Ethnicities, racism, crime, and criminal justice. The Oxford Handbook of Criminology: Oxford University Press; 2012. p. 370-97.
21. Kapadia D, Zhang J, Salway S, Nazroo J, Booth A, Villarroel-Williams N, et al. Ethnic Inequalities in Healthcare: A Rapid Evidence Review.: NHS Race and Health Observatory (RHO); 2022.
22. Williams P, MacDonald A. The effect of non-response bias on the results of two-stage screening surveys of psychiatric disorder. Journal of Social Psychiatry. 1986;21:182-6.
23. Giacco D, Palumbo C, Strappelli N, Catapano F, Priebe S. Social contacts and loneliness in people with psychotic and mood disorders. Compr Psychiatry. 2016;66:59-66.
24. Care Quality Commission. Community mental health survey 2021 2022 [Available from: https://www.cqc.org.uk/publications/surveys/community-mental-health-survey-2021.
25. Bhopal R, Fischbacher C, Povey C, Chalmers J, Mueller G, Steiner M, et al. Cohort profile: Scottish health and ethnicity linkage study of 4.65 million people exploring ethnic variations in disease in Scotland. Int J Epidemiol. 2011;40(5):1168-75.
26. Morgan C, Dazzan P, Morgan K, Jones P, Harrison G, Leff J, et al. First episode psychosis and ethnicity: initial findings from the AESOP study. World Psychiatry. 2006;5(1):40-6.
27. Hagger-Johnson G, Harron K, Goldstein H, Aldridge R, Gilbert R. Probabilistic linkage to enhance deterministic algorithms and reduce data linkage errors in hospital administrative data. J Innov Health Inform. 2017;24(2):891.
28. National Institute of Health Research. Data Linkage Service User and Carer Advisory Group 2023 [Available from: https://www.maudsleybrc.nihr.ac.uk/patients-public/help-shape-our-research/.

Figure 1. Flow chart illustrating the sample selection process for the census matched/not matched dataset

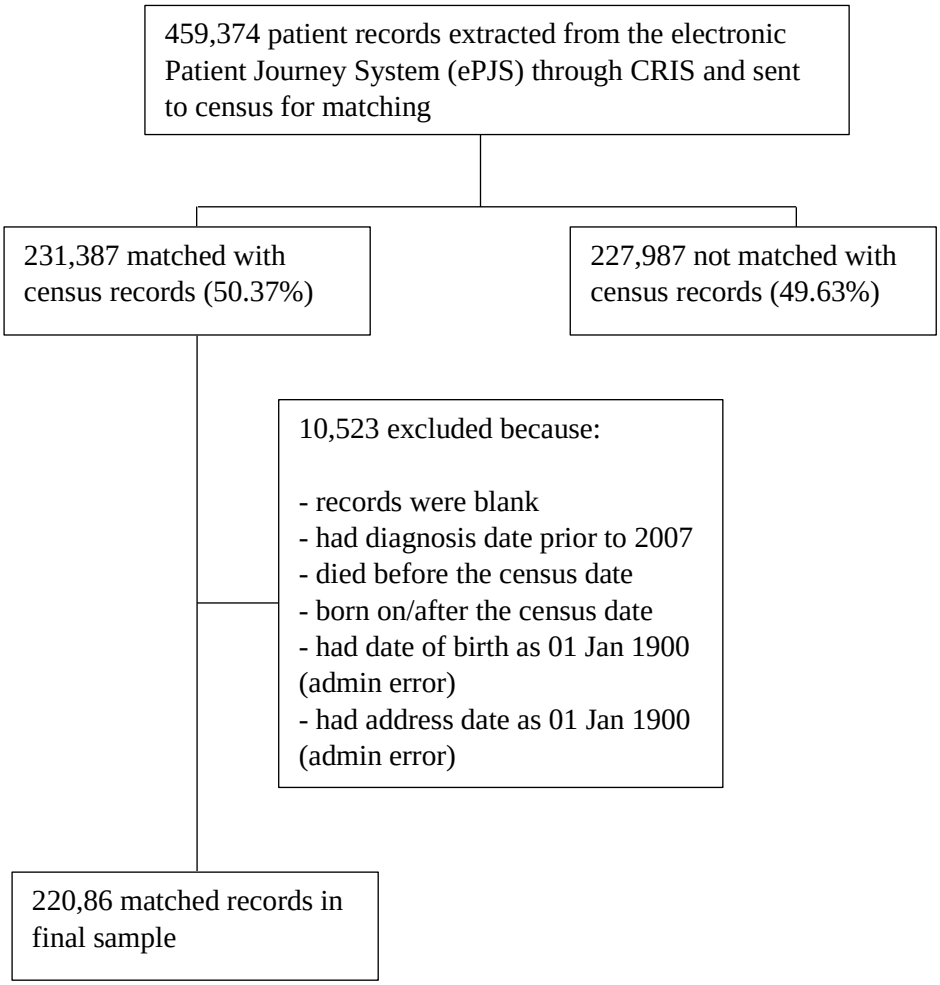

Figure 2. Proportion of electronic patient records identified via the Clinical Research Interactive Search (CRIS) matched to census by referral year