Office for National Statistics

Article

Comparing self-reported morbidity with electronic health records, England: 2021

Validating Census 2021 responses and electronic health records against each other, and against administrative health data and self-reported employment characteristics.

Contact:
Piotr Pawelek, Daniel Ayoubkhani, Vahe Nafilyan, Charlotte Bermingham
health.data@ons.gov.uk
+44 1633 455825

Release date:
23 June 2023

Next release:
To be announced

Table of contents

1. Main points

2. Overview of morbidity data sources

3. Methods used to assess coherence

4. Coherence of measures of morbidity derived from electronic health records (EHRs) with those collected from Census 2021

5. Benchmarking measures of morbidity derived from electronic health records (EHRs) and Census 2021 against hospital admission, death and self-reported employment characteristics

6. Glossary

7. Data sources and quality

8. Related links

9. Cite this research article

1 . Main points

- Measures of morbidity derived from self-reported Census 2021 data and routinely collected health records are positively correlated with one another, with people in poor or very poor general health and disabled people more likely to have a history of chronic health conditions than people in good or very good general health and non-disabled people.
- Self-reported data and routinely collected health records are similarly good at predicting administrative measures of morbidity and mortality (admission to hospital for any reason, death due to any cause, and death related to coronavirus (COVID-19) in the year following Census 2021).
- However, economic inactivity (defined as not working and not actively seeking work or being available to start) because of sickness at the time of Census 2021 was more accurately predicted by self-reported data than routinely collected health records.
- This may be because subjective measures of health and disability are better able to reflect the underlying severity of health conditions in terms of the impact they have on individuals' day-to-day activities, because the attribution of inactivity to sickness is subjective and, therefore, may be more accurately predicted by subjective measures of health; another cause may be that self-reported health and disability status and economic inactivity status were measured concurrently on Census 2021.

2 . Overview of morbidity data sources

The 2021 census of England and Wales included three questions related to self-reported health status and disability.

How is your health in general?

- Very good.
- Good.
- Fair.
- Bad.
- Very bad.

See our General health variable: Census 2021 release for more information.

Do you have any physical or mental health conditions or illnesses lasting or expected to last 12 months or more?

- Yes.
- No.

People who chose "Yes" to this question were then asked:

Do any of your conditions or illnesses reduce your ability to carry out day-to-day activities?

- Yes, a lot.
- Yes, a little.
- Not at all.

See our Disability variable: Census 2021 release for more information on both of these questions.

These questions aim to quantify and characterise aspects of morbidity and disability. Alternatively, measures of morbidity may be derived from electronic health records (EHRs), such as those routinely collected in primary care and hospital settings through the course of diagnosing, treating, and managing illness. For this analysis, the measures we derived from EHRs were:

- binary flags for each, and any, chronic health condition of interest
- number of nights spent in hospital for each, and any, chronic health condition of interest, using primary diagnoses recorded in the Hospital Episode Statistics Admitted Patient Care (HES APC) dataset only
- presence of frailty
- binary flag for hospital admission for any reason, using the HES APC dataset only
- number of hospital admissions for any reason, using the HES APC dataset only
- number of nights spent in hospital for any reason, using the HES APC dataset only
- number of hospital admissions for any reason involving critical care, using the HES APC dataset only
- number of nights spent in hospital for any reason involving critical care, using the HES APC dataset only
- number of Accident and Emergency (A&E) attendances, using the A&E and Emergency Care Data Set (ECDS) datasets

There are relative advantages and disadvantages to each of these approaches. For example, collecting questionnaire data places a burden on respondents, which is not the case with routinely collected EHRs. Furthermore, Census 2021 did not collect data on specific health conditions, and self-reported measures are subjective in nature.

On the other hand, the comprehensiveness of EHR data is limited by the range of diagnostic codes and patients available in the chosen database. Moreover, the extent to which health conditions are recorded will be influenced by healthcare-seeking behaviours, which may differ between socio-demographic groups. Self-reported measures are also able to reflect the impact of health conditions on individuals' day-to-day functioning (that is, disability), which is not possible for measures derived from EHRs.

Throughout the coronavirus (COVID-19) pandemic, both census and EHR-derived measures of morbidity have been used in Office for National Statistics (ONS) analysis of socio-demographic risk factors for COVID-19 death. These include:

- ethnicity, detailed in our Ethnic and religious contrasts in deaths involving COVID-19 articles
- religion, set out in our Deaths involving COVID-19 by religious group article
- disability status
- occupation, reported by BMJ Journals in its Occupation and COVID-19 mortality in England article

Inclusion of these measures has been vital to account for health-related factors that may vary across socio-demographic groups and partly determine outcomes from COVID-19 infection. If left unaccounted for, such factors could seriously bias analytical outputs and lead to incorrect interpretations.

However, there is a knowledge gap regarding which of these data sources, EHRs or census, leads to better measures of underlying ill-health and morbidity. This information could inform future analyses relating to COVID-19 and health outcomes more generally.

Therefore, this analysis aims to compare the coherence of health measures derived from Census 2021 and EHRs. In turn, it will try to validate both sets of measures against administrative data on hospital admission or mortality and self-reported data on employment characteristics, which may be affected by ill-health and disability.

3 . Methods used to assess coherence

Coherence of measures of morbidity derived from electronic health records and Census 2021

We computed correlation coefficients for each pair of measures derived from electronic health record (EHRs) and Census 2021, with the precise coefficient being determined by the variable types. For example, when:

- the EHR-derived measure was continuous or ordinal and the census measure was ordinal, we calculated Kendall's tau coefficient
- the EHR-derived measure was binary and the census measure was ordinal, we calculated the rank-biserial coefficient

For more information, see the 'Journal of Diagnostic Medical Sonography's' Measures of Association article.

Validating measures of morbidity derived from EHRs and Census 2021 against administrative measures of health and self-reported employment characteristics

We validated each EHR-derived and Census 2021 measure of morbidity against benchmark administrative measures of morbidity and employment characteristics (see Section 7: Data sources and quality).

By validating both EHR-derived and census measures of health against external, administrative data on morbidity or mortality, we recognise that neither set of measures are the ideal "gold standard". Rather, each is a measurable instrument aiming to proxy an unobserved concept.

We transformed continuous and ordinal EHR-derived and census measures to binary variables. We then fitted logistic regression models with each benchmark measure as the dependent variable and chose a probability cut-off by maximising the value of the F1 score. Individuals with a probability greater than, or equal to, the cut-off were then assigned the value 1, while those with a probability below the cut-off were assigned the value 0. Binary EHR-derived and census measures were kept in the raw state (0 for not present, 1 for present), as were the benchmark measures of health and employment characteristics (which were all binary to begin with).

Each pairwise validation involved producing a two-by-two matrix of person counts. The four cells in each matrix were defined as:

- $C_{0,0}$: the number of individuals for whom the binary EHR-derived or census-based morbidity measure and the benchmark measure were both 0 (that is, true-negative classifications)
- $C_{0,1}$: the number of individuals for whom the binary EHR-derived or census-based morbidity measure was 0 but the benchmark measure was 1 (that is, false-negative classifications)
- $C_{1,0}$: the number of individuals for whom the binary EHR-derived or census-based morbidity measure was 1 but the benchmark measure was 0 (that is, false-positive classifications)
- $C_{1,1}$: the number of individuals for whom the binary EHR-derived or census-based morbidity measure and the benchmark measure were both 1 (that is, true-positive classifications)

Using the count definitions above, we computed a range of metrics from the two-by-two matrix for each validation, including:

- sensitivity, or the true positive rate, as $C_{1,1}$ divided by ($C_{0,1}$ plus $C_{1,1}$)
- specificity, or the true negative rate, as $C_{0,0}$ divided by ($C_{0,0}$ plus $C_{1,0}$)
- positive predictive value (PPV), or the precision, as $C_{1,1}$ divided by ($C_{1,0}$ plus $C_{1,1}$)
- negative predictive value (NPV), the complement of the false-omission rate, as $C_{0,0}$ divided by ($C_{0,0}$ plus $C_{0,1}$)
- F1 score, a combined measure (harmonic mean) of sensitivity and PPV, as 2 multiplied by ((PPV multiplied by Sensitivity) divided by (PPV plus sensitivity))
- phi coefficient, a measure of correlation between the actual and predicted counts

4 . Coherence of measures of morbidity derived from electronic health records (EHRs) with those collected from Census 2021

Prevalence of chronic health conditions recorded in EHRs by Census 2021 health and disability statuses

Of the overall linked population in this analysis (see Section 7: Data sources and quality):

- 82.2% self-reported being in either very good general health (48.3%) or good general health (33.9%)
- 17.8% self-reported being in either fair general health (12.7%), bad general health (4.0%), or very bad general health (1.1%)
- 75.8% self-reported not having any physical or mental health conditions or illnesses lasting or expected to last 12 months or more
- 24.2% self-reported having any physical or mental health conditions or illnesses lasting or expected to last 12 months or more, with 7.0% not having their ability to carry out day-to-day activities reduced at all, 10.1% having their ability to carry out day-to-day activities reduced a little, and 7.1% having their ability to carry out day-to-day reduced activities a lot
- 47.4% had any of the 15 EHR-derived chronic conditions(see Section 7: Data sources and quality for the full list)

The difference between prevalence of any of the 15 EHR-derived chronic conditions and self-reported bad general health and disability is likely because of the objective and subjective nature of the former and the latter, respectively.

The vast majority (88.2%) of people who self-identified as being in very bad health had a history of at least one of the 15 chronic health conditions we analysed, compared with 26.1% of those who were in very good health. For each of the 15 chronic health conditions considered in the analysis, the percentage of people with a history of the condition (according to EHRs over the 10 years prior to Census 2021) increased as the self-identified health status (as reported in Census 2021) worsened (Table 1).

The highest percentages of people with a history of chronic conditions were consistently within those who self-identified as being in very bad health, with hypertension being the individual condition that had the highest percentage within this group (51.3%, compared with 4.3% among people in very good health). A considerable percentage of people in very bad health also had a history of obesity (49.2% versus 9.9% among people in very good health), frailty (43.7% versus 3.2%), severe mental illness (40.5% versus 5.6%) and depression (38.5% versus 5.5%).

Table 1: Percentages of people with a history of chronic health conditions by self-reported general health status, England, 2021

Chronic Condition	Very Good Health	Good Health	Fair Health	Bad Health	Very Bad Health	All People
Any of the Chronic Conditions	26.13	54.94	78.71	86.56	88.22	47.41
Hypertension	4.31	18.61	39.75	47.32	51.29	15.89
Obesity	9.93	26.52	41.40	48.32	49.21	21.51
Frailty	3.22	7.55	21.49	33.81	43.74	8.68
Severe Mental Illness	5.57	12.92	22.64	36.43	40.46	11.85
Depression	5.45	12.55	21.60	34.79	38.48	11.44
Diabetes	1.02	6.63	18.38	24.75	28.36	6.37
Asthma	6.26	11.32	16.63	22.49	25.30	10.15
Hospital Admission For Any Condition	1.32	4.86	12.50	18.85	24.61	5.00
Coronary Heart Disease	0.54	3.64	12.56	19.27	23.90	4.12
Chronic Obstructive Pulmonary Disease	0.21	1.58	7.94	17.29	23.25	2.59
Chronic Kidney Disease	0.51	2.73	8.85	12.43	15.07	2.96
Cancer	0.81	3.17	7.44	10.05	14.19	2.97
Hospital Admission Involving Critical Care	0.43	1.47	4.64	7.98	11.20	1.74
Osteoporosis	0.40	1.67	4.99	8.08	10.73	1.84
Dementia	0.03	0.22	1.77	4.76	8.23	0.59
Chronic Liver Disease	0.23	1.01	3.06	5.87	8.00	1.17
Rheumatoid Arthritis	0.13	0.80	3.00	5.67	6.78	1.01

Source: From the Office for National Statistics

Notes:

1. Chronic health conditions were derived from electronic health records over the 10 years prior to the day of Census 2021. General health status was self-reported on Census 2021 (see our General health variable: Census 2021 article).

Table 2 shows that, for the majority of the 15 chronic health conditions considered in the analysis, the percentage of people with a history of the condition (according to EHRs over the 10 years prior to Census 2021) increased with the self-reported severity of the impact of disability (as reported in Census 2021). See Section 7: Data sources and quality for the full list of chronic health conditions.

Among disabled people whose ability to carry out day-to-day activities was limited a lot, 80.5% had a history of any of the 15 chronic health conditions, while 15.9% had been hospitalized for any of them at some point. However, 36.3% of non-disabled people also had a history of any of these 15 conditions, while 2.8% had been hospitalized for them at some point.

The highest percentages of people with a history of chronic conditions were consistently within disabled people whose ability to carry out day-to-day activities was limited a lot, with obesity being the individual condition that had the highest percentage within this group (43.4% compared with 16.6% among non-disabled people). A considerable percentage of disabled people whose ability to carry out day-to-day activities was limited a lot also had a history of hypertension (41.6% versus 10.7% among non-disabled people), severe mental illness (34.5% versus 7.1%), frailty (33.4% versus 5.2%) and depression (32.5% versus 7.0%).

Table 2: Percentages of people with a history of chronic health conditions by self-reported disability status, England, 2021

Chronic Condition	No	Yes - Not Reduced At All	Yes - Reduced A Little	Yes - Reduced A Lot	All People
Any of the Chronic Conditions	36.28	70.99	74.03	80.50	47.41
Obesity	16.63	31.59	35.83	43.36	21.51
Hypertension	10.70	24.95	30.49	41.57	15.89
Severe Mental Illness	7.09	19.26	26.52	34.52	11.85
Frailty	5.22	8.81	17.24	33.40	8.68
Depression	6.97	18.54	25.37	32.46	11.44
Diabetes	3.23	15.41	13.95	20.26	6.37
Asthma	7.36	19.05	17.89	20.19	10.15
Coronary Heart Disease	2.11	5.61	9.88	15.93	4.12
Hospital Admission For Any Condition	2.83	7.45	10.93	15.91	5.00
Chronic Obstructive Pulmonary Disease	0.97	2.64	6.92	13.78	2.59
Chronic Kidney Disease	1.62	3.67	6.68	11.22	2.96
Cancer	1.88	4.79	6.32	8.00	2.97
Osteoporosis	0.94	2.09	4.12	7.86	1.84
Hospital Admission Involving Critical Care	0.91	2.08	3.85	7.21	1.74
Dementia	0.09	0.24	1.25	5.40	0.59
Rheumatoid Arthritis	0.33	1.28	3.12	5.07	1.01
Chronic Liver Disease	0.60	1.60	2.59	4.82	1.17

Source: From the Office for National Statistics

Notes:

1. Chronic health conditions were derived from electronic health records over the 10 years prior to the day of Census 2021. Disability status was self-reported on Census 2021 (see our Disability variable: Census 2021 article).

This analysis demonstrates the relative advantages and disadvantages of using each data source to define morbidity. While health and disability status were self-reported by Census 2021 respondents, the EHR-derived chronic health conditions are based on clinical diagnoses and thus provide objective markers of the presence of disease. Furthermore, the census does not capture information on specific health conditions or impairments which may be informative for some types of analysis.

However, the 15 health conditions considered in this analysis do not cover all possible conditions that could lead to individuals describing themselves as being in poor health or being disabled.

Some individuals with underlying health conditions may choose not to interact with the National Health Service (NHS), so their conditions would not be recorded in EHRs from the NHS, even if they describe themselves as being in poor health or being disabled.

The chronic health conditions were derived from EHRs over the look-back period 22 March 2011 to 21 March 2021 (that is, during the 10 years prior to Census Day). Some individuals could have recovered completely or experienced a significant improvement in their health, or changed their disability status, by the time they responded to Census 2021.

Correlations between measures of morbidity derived from electronic health records with those collected on Census 2021

Figure 1 presents correlation coefficients between measures of morbidity derived from EHRs and responses to Census 2021 (for more information on correlation coefficients, see Section 3: Methods used to assess coherence). Correlation coefficients range from negative one to positive one and represent the strength and direction of a bivariate association between two variables. A value of negative one indicates a perfect negative association. A value of zero indicates no association. A value of positive one indicates a perfect positive association.

For Kendall's tau coefficient:

- absolute values of 0 to 0.09 can be interpreted as very weak association
- absolute values of 0.10 to 0.19 can be interpreted as weak association
- absolute values of 0.20 to 0.29 can be interpreted as moderate association
- absolute values of 0.30 or above can be interpreted as strong association

For more information, see 'Educational and Psychological Measurement's' Assessing Magnitude of Effect from Rank-Order Correlation Coefficients.

For the Glass rank biserial coefcient:

- absolute values of 0.11 to 0.28 can be interpreted as weak association
- absolute values of 0.28 to 0.43 can be interpreted as moderate association
- absolute values of 0.43 or above can be interpreted as strong association

For more information, see 'Educational and Psychological Measurement's' Critical Values of the Rank-Biserial Correlation Coefficient.

Figure 1: There was a positive association between chronic health conditions and self-reported morbidity measures

Correlation coefficients between histories of chronic health conditions derived from electronic health records and self-reported health and disability measures collected on Census 2021, stratified by age group, England, 2021

Notes:

1. Chronic health conditions were derived from electronic health records over the 10 years prior to the day of Census 2021. Health and disability statuses were self-reported on Census 2021 (see our General health variable: Census 2021 and Disability variable: Census 2021 articles).

Download the data

.xlsx

Figure 1 demonstrates that all chronic health conditions derived from EHRs were positively associated with the severity of bad health and reduction in daily activity because of disability, as measured on Census 2021, but to varying degrees of strength. Across all ages (Figure 1, columns one and two), the strongest association between an individual health condition and Census 2021 measures was for dementia, at 0.78 for general health and 0.75 for disability. Other conditions that were moderately associated with general health and disability were chronic obstructive pulmonary disease (COPD) (correlation coefficients of 0.71 and 0.54, respectively), rheumatoid arthritis (0.63 and 0.56) and coronary heart disease (0.62 and 0.42).

However, the strength of the positive associations between EHR-derived conditions and Census 2021 health and disability measures are likely to be affected by age. Older people are more likely than younger people to have a health condition and also to be in worse health or to be disabled. Prevalence measures for each and any of the 15 EHR-derived chronic conditions and Census 2021 self-reported general health and disability statuses are available in the accompanying data tables.

When stratifying the results by broad age group, the correlation coefficients were generally higher for people aged under 65 years (Figure 1, columns three and four) than for those aged 65 years or older (Figure 1, columns five and six). One possible reason for this is that age alone (separately from the presence of health conditions) explains a relatively large proportion of the variability in self-reported health and disability statuses among older people. Conversely, age alone is unlikely to substantially contribute to poor health or disability among younger people, and the presence of underlying health conditions is likely to be a larger contributing factor.

Additional results showing the degree of correlation between Census 2021 measures of health and disability and the number of nights spent in hospital during the ten-year look-back period for each of the chronic health condition are available in the accompanying data tables. Correlation coefficients are consistently positive but generally weak, with the highest correlation coefficient being 0.19 for disability and 0.21 for general health for any of the 15 health conditions, and the highest correlation coefficient for any individual condition being 0.13 for coronary heart disease for general health.

5 . Benchmarking measures of morbidity derived from electronic health records (EHRs) and Census 2021 against hospital admission, death and self-reported employment characteristics

Predictive accuracy of morbidity measures

Figure 2 presents F1 scores as a measure of the accuracy of morbidity measures derived from EHRs and Census 2021 in predicting hospital admission, death, inactivity because of sickness, and working 15 hours or less. The F1 score ranges from zero to one, with higher scores indicating better predictive performance in terms of precision and sensitivity. Individual scores for precision, sensitivity and all other metrics specified in Section 3: Methods used to assess coherence are available in the accompanying data tables.

The likelihood of experiencing many of the chronic health conditions included in the analysis, as well as self-reported ill-health or disability, is likely to increase as individuals get older. Therefore, as with the coherence measures in Section 4: Coherence of measures of morbidity derived from electronic health records (EHRs) with those collected from Census 2021, the F1 scores are likely to be affected by age. The models used to produce the F1 scores in Figure 2 have thus been additionally adjusted for age to infer how much predictive power is added by the measures of morbidity, in addition to age. F1 scores for a model including only age are included to provide a benchmark for comparison. F1 scores from models not adjusted for age are available in the accompanying data tables.

Figure 2: Electronic health records and self-reported measures of morbidity perform similarly at predicting administrative measures of health

Age-adjusted F1 scores for morbidity measures derived from electronic health records and Census 2021 for predicting administrative measures of health and self-reported employment characteristics, England, 2021

Notes:

1. Chronic health conditions were derived from electronic health records over the 10 years prior to the day of Census 2021. Health and disability statuses were self-reported on Census 2021 (see our General health variable: Census 2021 and Disability variable: Census 2021 articles).

2. Admissions to hospital for any reason were identified from Hospital Episode Statistics for the year following the day of Census 2021.

3. Deaths due to any cause and death due to coronavirus (COVID-19) were identified from death registrations for deaths occurring during the year following the day of Census 2021.

4. Inactivity because of sickness and working 15 hours or less were self-reported on Census 2021 (see our Economic activity status and hours worked variable: Census 2021 release).

5. Inactivity in Census 2021 is defined as being without a job and having not actively sought work in the last four weeks, or not being available to start work in the next two weeks. See our Comparing Census 2021 and Labour Force Survey estimates of the labour market article.

Download the data

.xlsx

According to the age-adjusted F1 scores, EHR-derived measures of morbidity are similarly good at predicting administrative measures of health, namely hospital admission for any reason, death due to any cause and COVID-19 death, compared with Census 2021 measures of health and disability. Both EHR-derived and Census 2021 measures of morbidity add little predictive power in addition to age for predicting administrative measures of health and morbidity.

Binary indicators for having any of the 15 EHR-derived chronic conditions produced age-adjusted F1 scores of 0.25 and 0.17 for hospital admission and death due to any cause, respectively. Including binary indicators for all 15 of the chronic health conditions into a single logistic regression model produced F1 scores of 0.31 and 0.22 for hospital admission and death due to any cause, respectively. The number of hospital admissions for these 15 conditions produced age-adjusted F1 scores of 0.31 and 0.21 for hospital admission and death due to any cause, respectively, while the number of nights spent in hospital for these conditions produced F1 scores of 0.27 and 0.19, respectively.

The Census 2021 measure of general health produced F1 scores of 0.29 and 0.24 for hospital admission and death due to any cause, respectively. The Census 2021 measure of disability produced F1 scores of 0.28 and 0.21 for hospital admission and death due to any cause, respectively. In terms of the magnitude of the F1 score, the Census 2021 and EHR-derived measures of morbidity were similar for both hospitalization and death due to any cause.

Figure 2 shows that, according to the age-adjusted F1 score, Census 2021 measures of health and disability were better at predicting if someone was inactive because of sickness at the time of Census 2021 when compared with EHR-derived measures of morbidity.

When predicting inactivity because of sickness, self-reported disability and general health produced age-adjusted F1 scores of 0.65 and 0.57, respectively. A binary indicator for having any of the 15 EHR-derived chronic conditions, including binary indicators for all 15 of the chronic health conditions into a single logistic regression model, the number of hospital admissions, and number of nights spent in hospital, produced age-adjusted F1 scores of 0.18, 0.37, 0.28, and 0.24, respectively.

Figure 2 also shows that, according to the age-adjusted F1 score, EHR-derived measures of morbidity were similarly good at predicting if someone was working 15 hours or less at the time of Census 2021 compared with self-reported measures of health and disability.

When predicting working 15 hours or less, self-reported disability and general health produced age-adjusted F1 scores of 0.24 and 0.23, respectively. A binary indicator for having any of the 15 EHR-derived chronic conditions, including binary indicators for all 15 of the chronic health conditions into a single logistic regression model, the number of hospital admissions, and number of nights spent in hospital for these conditions, all produced age-adjusted F1 scores of 0.23.

Age on its own produced F1 scores of 0.24, 0.17, and 0.23 for hospital admission, death due to any cause, and working 15 hours or less, respectively. The F1 scores produced by age on its own for hospital admission, death due to any cause, and working 15 hours or less, were similar in magnitude to the univariate scores of both EHR-derived and Census 2021 measures of morbidity, which are available in the accompanying data tables.

Age on its own produced F1 scores of 0.13 and 0.23 for inactivity because of sickness and working 15 hours or less, respectively. This shows that, in addition to age, the EHR-derived measures of morbidity added little predictive power for predicting inactivity because of sickness, while self-reported health and disability status added more predictive power. However, for working 15 hours or less, both EHR-derived measures and self-reported measures of health and disability added little predictive power in addition to age.

F1 scores, PPV, NPV, sensitivity, specificity and phi coefficients between the number of nights spent in hospital for each individual chronic condition and administrative measures of health and self-reported employment characteristics are available in the accompanying data tables.

6 . Glossary

Coronavirus (COVID-19)

Coronaviruses are a family of viruses that cause disease in people and animals. They can cause the common cold or more severe diseases, such as COVID-19. COVID-19 is the name used to refer to the disease caused by the SARS-CoV-2 virus, which is a type of coronavirus. The Office for National Statistics (ONS) takes COVID-19 to mean the presence of SARS-CoV-2, with or without symptoms.

Logistic regression

Logistic regression is a statistical technique for modelling the relationship between two characteristics of interest (such as self-reported health status and the presence of chronic health conditions). The model can be used to understand the independent relationship between the two characteristics. This is while "adjusting" or "controlling" for other characteristics (such as age), which may be related to both of the two characteristics of interest.

Sensitivity

In the context of this analysis, sensitivity, also called the true positive rate, is the percentage of people predicted to have a particular characteristic who actually have it. A measure that is 100% sensitive means that all individuals who have the characteristic were correctly predicted to have it.

Specificity

In the context of this analysis, specificity, also called the true negative rate, is the percentage of people predicted not to have a particular characteristic who actually do not have it. A measure that is 100% specific means that all individuals who do not have the characteristic were correctly predicted to not have it.

Positive predictive value

In the context of this analysis, positive predictive value (PPV), also called precision, is the percentage of people predicted to have a particular characteristic who actually have it. A measure with a PPV of 100% means that all the people predicted to have the characteristic actually have it.

Negative predictive value

In the context of this analysis, negative predictive value (NPV) is the percentage of people predicted to not have a particular characteristic who actually do not have it. A measure with a NPV of 100% means that all the people predicted not to have the characteristic actually do not have it.

F1 Score

High PPV means that a high percentage of people predicted to have a particular characteristic actually have it. High sensitivity means that a high percentage of people who actually have a particular characteristic are predicted to have it. Ideally, a model should have both high sensitivity and high PPV but increasing PPV usually comes at a cost of lower sensitivity, and vice versa. Therefore, the F1 score is calculated as the harmonic mean of sensitivity and PPV to reflect the fact that both of these measures are important dimensions of a model's accuracy.

Phi coefficient

The phi correlation coefficient, also known as the Matthews coefficient, Yule phi or Mean Square Contingency coefficient, ranges from negative one to positive one, and represents the strength and direction of a bivariate association between two binary variables. A value of negative one indicates a perfect negative association. A value of zero indicates no association. A value of positive one indicates a perfect positive association. The phi coefficient is symmetrical, which means the two binary variables are interchangeable in terms of their ordering in the calculation.

7 . Data sources and quality

Data sources

This analysis used data from the following linked data sources:

- Census 2021, from the Office for National Statistics (ONS)
- Hospital Episode Statistics (HES) Admitted Patient Care (APC), Outpatient (OP), Accident and Emergency (A&E), and Emergency Care (ECDS) datasets from NHS digital
- General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR) from NHS digital

The study population was comprised of usual residents of England who responded to Census 2021 and could be linked to GDPPR via the NHS number, obtained from the 2019 Patient Demographic Service (PDS) dataset.

Chronic health conditions derived from electronic health records (EHRs)

We used EHRs to derive a set of measures of morbidity relating to the following chronic health conditions:

- hypertension
- obesity
- diabetes
- coronary heart disease
- chronic kidney disease
- chronic liver disease
- asthma
- chronic obstructive pulmonary disease (COPD)
- depression
- severe mental disorder (psychosis, schizophrenia, bipolar disorder)
- dementia
- rheumatoid arthritis
- osteoporosis
- cancer
- frailty

The conditions were inferred from a combination of the GDPPR and HES datasets, using records from the look-back period 22 March 2011 to 21 March 2021 (that is, during the 10 years prior to Census Day). In the case of GDPPR, we derived the conditions using code clusters defined by NHS England. In the case of HES, we used primary and secondary diagnoses recorded using International Classification of Diseases 10th Revision (ICD-10) codes in the APC and OP datasets.

Measures of morbidity derived from EHRs

The measures produced from EHRs relate to both the presence of health conditions and the severity of ill-health, including:

- binary flags for each, and any, chronic health condition of interest
- number of nights spent in hospital for each, and any, chronic health condition of interest, using primary diagnoses recorded in the Hospital Episode Statistics Admitted Patient Care (HES APC) dataset only
- presence of frailty
- binary flag for hospital admission for any reason, using the HES APC dataset only
- number of hospital admissions for any reason, using the HES APC dataset only
- number of nights spent in hospital for any reason, using the HES APC dataset only
- number of hospital admissions for any reason involving critical care, using the HES APC dataset only
- number of nights spent in hospital for any reason involving critical care, using the HES APC dataset only
- number of A&E attendances, using the A&E and ECDS datasets

Benchmark measures for validation of measures of morbidity derived from EHRs and census data

As benchmark measures of health, we derived binary flags indicating whether participants had been admitted to hospital for any reason or died of any cause between 22 March 2021 and 21 March 2022 (that is, during the one year following Census Day). We also derived a flag indicating death related to coronavirus (COVID-19), using U07.1 (COVID-19, virus identified) and U07.2 (COVID-19, virus not identified) ICD-10 codes.

The 2021 census of England and Wales included several questions related to the labour market, including questions about economic activity status and hours worked (see our Economic activity status and hours worked variable: Census 2021 release). We used responses to those questions to derive the following binary employment variables, which may be affected by ill-health, among people aged 16 to 64 years:

- not working because of long-term sickness
- working 0 to 15 hours (among those in work)

Strengths and limitations

This is a population-level study covering all usual residents of England who were enumerated at Census 2021 and could be linked to an NHS number. Therefore, the analysis is not subject to sampling error. However, not all people living in England in March 2021 were enumerated at Census 2021 (for example, because of non-response), and of those who were, not all could be linked to an NHS number via the PDS and onward to the GDPPR extract. In the linked population relative to the enumerated Census 2021 population:

- females are overrepresented and males are underrepresented
- people aged 20 to 29 years are the most underrepresented age group, and those aged 70 to 79 years are the most overrepresented age group
- white ethnic groups are overrepresented, while other ethnic groups are underrepresented

This linkage failure may be related to the likelihood of engaging with healthcare services, and thus introduce bias into our analysis.

The GDPPR extract includes approximately 40,000 medical codes out of approximately 1 million available for use by general practitioners. While many prevalent chronic health conditions are included within the scope of the extract, we were not able to identify people with some common conditions, such as generalised anxiety disorder.

Although we had access to Census 2021 data for individuals in both England and Wales, we only had access to EHRs for individuals in England. As a result, the study population does not include people in Wales.

8 . Related links

Disability, England and Wales: Census 2021
Bulletin | Released 19 January 2023
Information on disability in England and Wales, Census 2021 data.

General health, England and Wales: Census 2021
Bulletin | Released 19 January 2023
People's health across local authorities in England and Wales, Census 2021 data.

Economic activity status and hours worked variable: Census 2021
Web page | Released 16 May 2023
Definition of economic activity status and hours worked, categories, and changes since the 2011 Census for use with research and analysis using Census 2021 data.

9 . Cite this research article

Office for National Statistics (ONS), released 23 June 2023, ONS website, research article, Comparing self-reported morbidity with electronic health records, England: 2021