Office for National Statistics

Feasibility research: A rules-based approach to estimate disability prevalence using linked administrative data in England and Wales

This working paper explores the development and assessment of a rules-based approach to estimate disability prevalence in England and Wales using a large linked administrative dataset.

Contact:
National Statistician's Analysis Unit
Health.Data@ons.gov.uk
+44 1329 444110

Release date:
20 March 2026

Next release:
To be announced

Notice

24 March 2026

In the initial publication, Table 1 had two column headings which were incorrectly swapped. The column displaying percentages of disabled flags was labelled as “Percentages of non-disabled flags”, and the column displaying percentages of non-disabled flags was labelled as “Percentages of disabled flags”.

Table of contents

1 . Overview

This article outlines the development and assessment of a rules-based approach to estimate disability prevalence in England and Wales. Using a large, de-identified, linked administrative dataset and applying proxy disability indicators to each data source, we estimated disability prevalence, which we compared against self-reported disability from Census 2021.

At an England and Wales level, the proportion of individuals identified as "likely disabled" through the rules-based method aligned closely with Census 2021. The percentage of the population identified as "likely disabled" using the rules-based approach was 18.0%, compared with a prevalence of 17.7% based on Census 2021. However, accuracy varied across demographic groups, characteristics, and geographic areas, reflecting differences in data coverage, the nature of disability indicators within administrative sources, and the inherent limitations of proxybased classification.

This article summarises the strengths and limitations of this approach. While the method shows early promise, further work will be conducted. This will explore alternative approaches, including predictive modelling techniques, to enhance accuracy and reduce the proportion of records for which disability status cannot be derived.

2 . Background

Within the UK and internationally there is an evidence gap regarding the measurement of disability using administrative data. This paper examines whether linked administrative datasets can be used to generate proxy measures of disability using a rulesbased approach. The estimates of disability prevalence using this rules-based approach are benchmarked to prevalence estimated using self-reported disability from Census 2021.

Defining and measuring disability is complex because of the different conceptual models that exist. The medical model of disability suggests that people are disabled because of the impairments or conditions they have. In contrast, the social model suggests that disability is created by society through barriers, preventing people with certain medical conditions in fully taking part in society in the same way as those people without certain medical conditions.

Both models offer valuable insights into disability. The Equality Act (2010) for Great Britain and the Disability Discrimination Act (1995) for Northern Ireland - which underpin antidiscrimination legislation - align with the social model. Conversely, the NHS often operates under the medical model, providing necessary healthcare services.

The Government Statistical Service (GSS) harmonised standard for disability determines if an individual would be identified as disabled. This standard establishes how to collect and report disability statistics to ensure comparability across different data collections. This is either according to the Equality Act (2010) for Great Britain or the Disability Discrimination Act (1995) for Northern Ireland; and as such, the measurement reflects the social model of disability.

The disability harmonised standard is under continual development following the Inclusive Data Taskforce recommendations to review and update the harmonised standards every five years. However, most administrative datasets do not capture disability according to this harmonised definition. Instead, they record information relating to health conditions, service use, or educational support needs. As such, administrative indicators tend to align more closely with the medical model than the social model, with implications for coherence, comparability, and user acceptability of derived disability measures.

Given these challenges, the present research explores the feasibility of using available administrative indicators as proxies for disability, while acknowledging their limitations. The aim is not to replicate the harmonised definition directly, but rather to assess the extent to which administrative data can support the production of meaningful and methodologically robust estimates of disability prevalence when validated against Census 2021.

3 . Methods

Data sources

The rules-based approach was developed using 13 datasets. These datasets capture information across education, geography, health, benefits, and demographics, and vary in temporal coverage and population scope. These datasets are:

- Statistical Population Dataset version 4.3: 2021 (30 June 2021)
- Census 2021: 21 March 2021
- Demographic Index version 4.0.1 and version 4.2
- Demographic Index – Census 2021 lookup
- Early Years Census: 2001 to 2002, to 2020 to 2021 academic year
- English School Census: 2010 to 2011, to 2020 to 2021 academic year
- Welsh School Census: 2010 to 2011, to 2020 to 2021 academic year
- Individualised Learner Record: 2015 to 2016, to 2020 to 2021 academic year
- Lifelong Learning Wales Record: January 2000 to August 2021
- Higher Education Statistics Agency: 2010 to 2011, to 2020 to 2021 academic year
- Hospital Episode Statistics: March 2020 to March 2021
- NHS Talking Therapies: financial year ending 2013 to financial year ending 2021
- Benefits and Income Dataset: financial year ending 2011 to December 2021

All data sources were de-identified before commencing work on the rules-based approach.

Population spine and linkage

The population spine used in this research is the Statistical Population Dataset 2021 (version 4.3), comprising 57.4 million records. Nonhealth administrative data sources were linked to this population spine using lookup information from the Demographic Index, specifically versions 4.0.1 or 4.2, depending on the source dataset. For more information, see our Understanding quality of the Statistical Population Dataset in England and Wales using the 2021 Census - Demographic Index linkage article. Healthrelated administrative data already contained a Census 2021 identifier and were therefore linked to the population spine through preexisting Census 2021 – Demographic Index linkages.

This section reports:

- the size of the population spine
- the number of individuals successfully linked to each administrative data source
- the total number of individuals present within each source

It is important to note that administrative data may include individuals who died or emigrated before the creation of the Statistical Population Dataset 2021. Furthermore, the administrative data sources differ in their reference periods, which affects the extent to which each is influenced by deaths and emigration. These temporal differences should be considered when interpreting linkage rates and coverage patterns across the datasets.

Population spine

The number of records in the Statistical Population Dataset 2021 with valid Census 2021 disability status is 51,435,680.

The number of records in the Statistical Population Dataset 2021 without valid Census 2021 disability status is 6,007,100.

Therefore, the total number of records in the dataset with and without valid Census 2021 disability status is 57,442,780.

Count of administrative data records linked to population spine:

- Benefits and Income Dataset: 23,052,190
- Hospital Episode Statistics: 18,914,695
- English School Census: 11,178,690
- Early Years Census: 5,356,115
- Individualised Learner Record: 4,348,975
- Higher Education Statistics Agency: 4,267,000
- NHS Talking Therapies: 936,140
- Welsh School Census: 634,300
- Lifelong Learner Record Wales: 178,375

Count of administrative data records pre-linkage:

- Benefits and Income Dataset: 34,543,700
- Hospital Episode Statistics: 19,601,630
- English School Census: 14,870,025
- Early Years Census: 6,116,950
- Individualised Learner Record: 9,671,960
- Higher Education Statistics Agency: 12,875,410
- NHS Talking Therapies: 1,013,215
- Welsh School Census: 873,735
- Lifelong Learner Record Wales: 1,503,435

Creation of disability flags

Administrative data sources contain limited direct indicators of disability. Consequently, a rulesbased approach was developed to derive an indicator of likely disability status for each individual across the administrative data sources. For each source, individuals were assigned to one of three categories: likely disabled, likely nondisabled, or unknown disability status.

An individual's disability flag was derived using information from their most recent available record. This approach was adopted to minimise the risk of assigning a likely disabled status based on a historic health condition or disabilityrelated indicator that no longer applied.

The criteria used to assign likely disabled or likely nondisabled status varied by data source, reflecting differences in the structure of the underlying information. Where important information was missing or invalid, the disability status was set to unknown. The rules used to identify individuals as likely disabled were as follows:

- Early Years Census: in receipt of special educational needs provision or disability-related funding
- English School Census: in receipt of special educational needs provision and with valid code for primary special educational needs type, or in receipt of disability-related funding
- Welsh School Census: valid code for special educational needs type
- Individualised Learner Record: has a learning difficulty, disability or health problem and valid code for the nature of the learning difficulty, disability or health problem
- Lifelong Learning Record Wales: has a disability and/or learning difficulty and valid code for the primary type of disability or learning difficulty
- Higher Education Statistics Agency: valid code for disability type
- NHS Talking Therapies: valid code for disability type
- Hospital Episode Statistics: medical diagnosis or treatment defined by blocks of the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10), which had a relatively strong association with disability (60% or more of people with the ICD-10 block code were disabled on Census 2021)
- Benefits and Income Dataset: in receipt of one or more of Personal Independence Payment, Disability Living Allowance, Attendance Allowance, Employment and Support Allowance, Severe Disablement Allowance, Incapacity Benefit or Passported Incapacity Benefit

An individual was classified as likely disabled if they had been assigned a likely disabled flag in at least one administrative data source. If an individual had not been flagged as likely disabled in any source, they were classified as likely nondisabled provided they had been flagged as likely nondisabled in at least one source. Individuals who did not meet either condition were assigned an unknown disability status. Where there were conflicting disability statuses for an individual by data source the most recent record for an individual was taken.

This approach was designed to avoid prioritising any single administrative dataset. It recognises that sources vary in their coverage, the types of disabilityrelated information they collect, and the age groups they predominantly represent.

Table 1: Proportion of records by derived disability status flags for each administrative data source from the population spine

Administrative data source	No flag (%)	Likely non-disabled (%)	Likely disabled (%)
Early Years Census	90.7	9.0	0.3
English School Census	80.5	16.4	3.1
Welsh School Census	98.9	0.9	0.2
Individualised Learner Record	92.7	5.8	1.5
Lifelong Learner Record Wales	99.9	0.0	0.1
Higher Education Statistics Agency	92.8	6.0	1.2
NHS Talking Therapies	98.5	0.7	0.8
Hospital Episode Statistics	67.1	26.9	6.1
Benefits and Income Dataset	59.9	30.3	9.8

Source: Bespoke data linkage across multiple administrative datasets from the Office for National Statistics

4 . Results

Overall disability prevalence

Across England and Wales, 18.0% of individuals were classified as likely disabled using the rulesbased method, representing a difference of 0.5 percentage points above the Census 2021 estimate, see Table 2. The discrepancy between the estimated and Censusderived proportions is notably larger for the nondisabled population (25.1 percentage points). However, if the majority of individuals with an unknown disability status are in fact nondisabled, and were classified accordingly, this difference would be substantially reduced. The divergence between estimated and Census 2021 rates was also smaller for England than for Wales. This could be because of the availability of health administrative data sources for Wales when compared with England.

Table 2: Overall counts and proportions of disability produced from the rules-based approach and corresponding counts and proportions of disability from Census 2021

Likely disability status	Rules-based (Count)	Rules-based (%)	Census (Count)	Census (%)
England and Wales: Disabled	10,312,640	18.0	10,444,775	17.5
England and Wales: Non-disabled	32,982,555	57.4	49,152,765	82.5
England and Wales: Unknown	14,147,580	24.6	0	0.0
England: Disabled	9,774,415	17.9	9,774,510	17.3
England: Non-disabled	31,488,905	57.7	46,715,540	82.7
England: Unknown	13,330,470	24.4	0	0.0
Wales: Disabled	538,225	18.9	670,265	21.6
Wales: Non-disabled	1,493,655	52.4	2,437,230	78.4
Wales: Unknown	817,110	28.7	0	0.0

Source: Bespoke data linkage across multiple administrative datasets from the Office for National Statistics

While results at the overall level show similarities to census data (Table 2), it is important to consider accuracy at the individual level and by socio-demographic breakdowns.

Confusion matrix for overall population

Table 3 summarises the results of the confusion matrix, as well as the number of individuals for whom a rulesbased disability status or a Census 2021 disability status could not be obtained.

Table 3: Confusion matrix terms for rules-based approach disability estimates, disabled as positive class, non-disabled as negative class

Measure	Count
True positives (likely disabled on both administrative data and Census 2021)	5,425,370
False positives (likely disabled on administrative data, non-disabled on Census 2021)	4,067,520
True negatives (likely non-disabled on both administrative data and Census 2021)	27,265,405
False negatives (likely non-disabled on administrative data, disabled on Census 2021)	3,297,695
Likely disabled on administrative data, unknown Census 2021 status	819,755
Likely non-disabled on administrative data, unknown Census 2021 status	2,419,460
Unknown status on administrative data, disabled on Census 2021	572,980
Unknown status on administrative data, non-disabled on Census 2021	10,806,715
Unknown status on both administrative data and Census 2021	2,767,885

Source: Bespoke data linkage across multiple administrative datasets from the Office for National Statistics

Sensitivity was calculated as the proportion of disabled individuals correctly identified as disabled by the rulesbased approach, defined as the number of true positives divided by the sum of true positives and false negatives.

Specificity was calculated as the proportion of nondisabled individuals correctly identified as nondisabled, defined as the number of true negatives divided by the sum of true negatives and false positives.

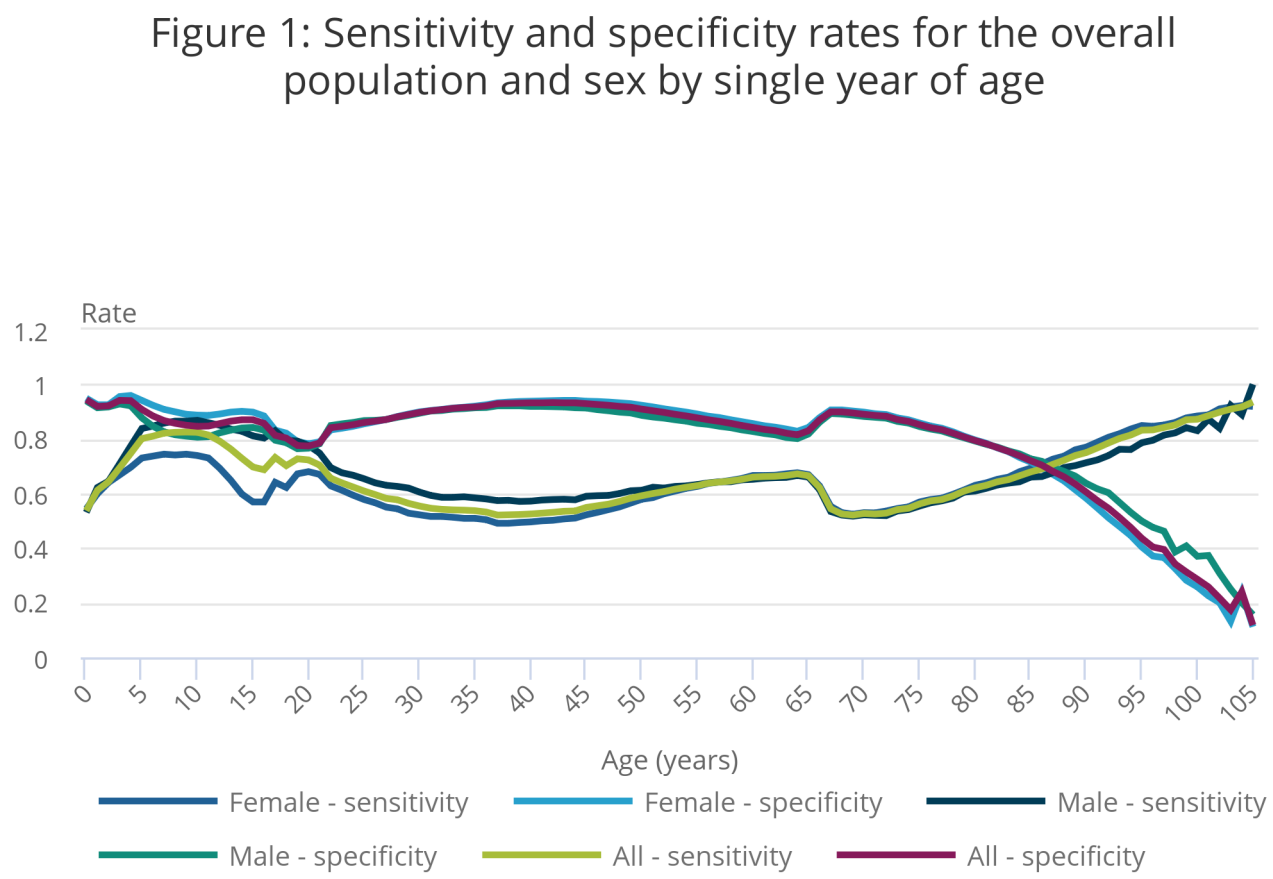
In this analysis, disability was treated as the positive class and nondisability as the negative class. Sensitivity and specificity therefore reflect the extent to which the rulesbased approach accurately classifies individuals according to their true disability status as recorded in Census 2021. Only records with valid disability information in both administrative and census data were included in these calculations. The overall sensitivity rate was 0.62, and the overall specificity rate was 0.87.

Sensitivity and specificity rates by sex and age

Overall, males exhibited higher sensitivity rates (0.64) than females (0.61), whereas females demonstrated higher specificity rates (0.88) compared with males (0.85).

Across the age profile, both sensitivity and specificity displayed notable variation, with pronounced peaks and troughs at ages associated with key life-course events, for example, school attendance and retirement (Figure 1). At the oldest ages, sensitivity reached its highest levels, while specificity was comparatively low. In contrast, the youngest ages were characterised by some of the lowest sensitivity and highest specificity values.

Figure 1: Sensitivity and specificity rates for the overall population and sex by single year of age



Source: Bespoke data linkage across multiple administrative datasets from the Office for National Statistics

Notes:

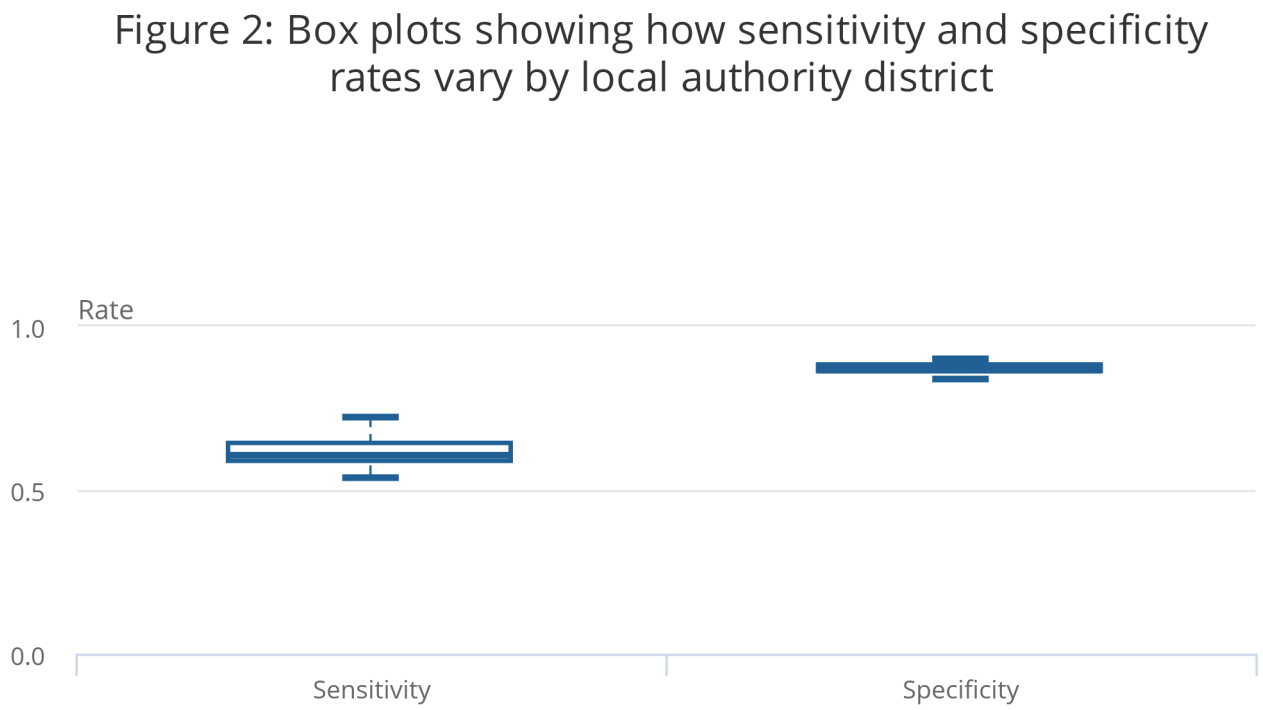
1. A sensitivity rate is the proportion of disabled people who are correctly identified as disabled through our flagging method. A specificity rate is the proportion of non-disabled people who are correctly identified as non-disabled through our flagging method.

By sex, females showed higher sensitivity at age 0 years and from age 58 years onwards, while males had higher sensitivity across most ages between 1 and 57 years. For specificity, females had higher rates for ages 0 to 21 years and 27 to 81 years, with males exhibiting higher specificity at the remaining ages. The largest sex difference in sensitivity was observed between ages 5 and 20 years, during which males consistently recorded higher sensitivity. In contrast, the greatest differences in specificity were observed from age 92 years and over, where males displayed higher specificity than females.

Sensitivity and specificity rates by geography

Sensitivity rates for local authorities range from 0.47 for City of London to 0.73 for Knowsley. Specificity rates for local authorities range from 0.83 for Blackpool to 0.91 for Vale of Glamorgan (Figure 2).

Figure 2: Box plots showing how sensitivity and specificity rates vary by local authority district



Source: Bespoke data linkage across multiple administrative datasets from the Office for National Statistics

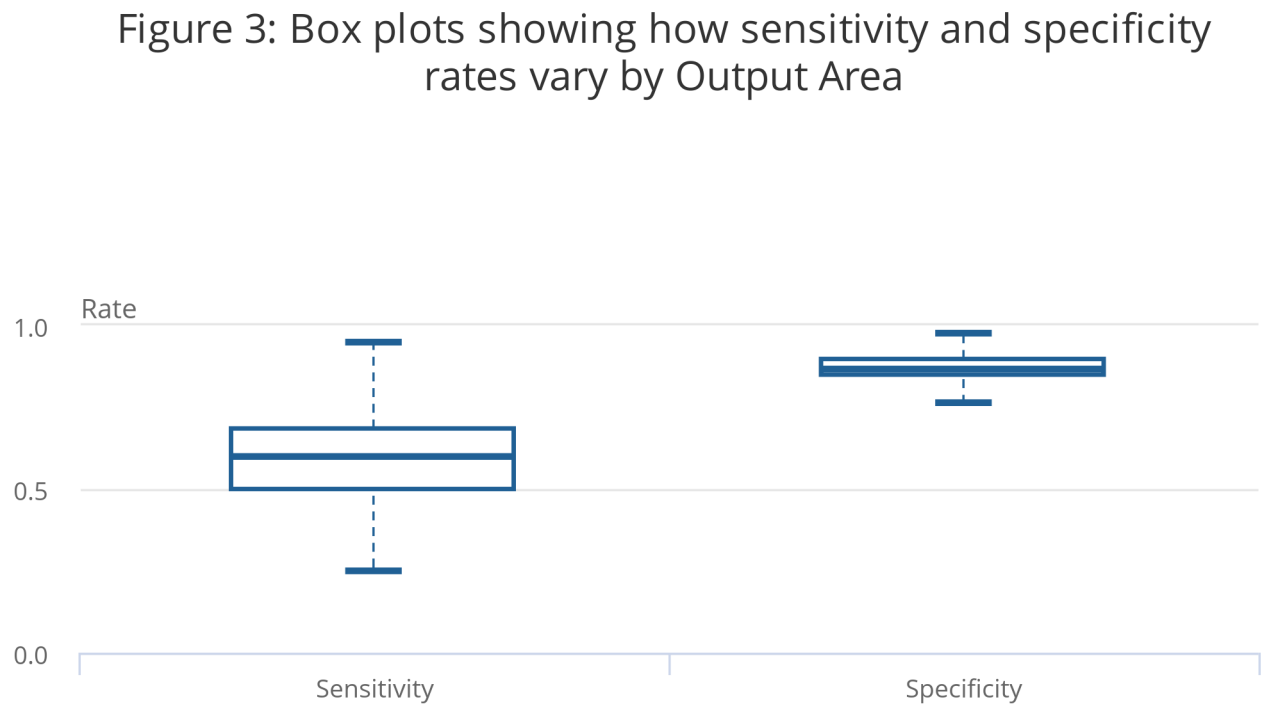
Notes:

1. A sensitivity rate is the proportion of disabled people who are correctly identified as disabled through our flagging method. A specificity rate is the proportion of non-disabled people who are correctly identified as non-disabled through our flagging method.

Substantial variation in sensitivity and specificity rates is observed at the Output Area level (Figure 3). At this geographical granularity, both metrics span the full possible range, with maximum values of 1 for both sensitivity and specificity (these are outliers and are therefore not visable on the figure). In contrast, the minimum sensitivity and specificity rates fall to 0.14 and 0.25, respectively. The larger number of outliers in Figure 3 relative to Figure 2 is likely attributable to the smaller population sizes within many Output Areas. This issue is further compounded when records must be excluded from the calculation of sensitivity and specificity because of either an invalid Census 2021 disability status or an unknown disability status derived from administrative data.

A consistent pattern across all geographical levels is that the variation in specificity rates is narrower than the variation in sensitivity rates.

Figure 3: Box plots showing how sensitivity and specificity rates vary by Output Area



Source: Bespoke data linkage across multiple administrative datasets from the Office for National Statistics

Notes:

1. A sensitivity rate is the proportion of disabled people who are correctly identified as disabled through our flagging method. A specificity rate is the proportion of non-disabled people who are correctly identified as non-disabled through our flagging method.

Comparison of estimated and observed disability prevalence by socio-demographic characteristics and geography

In this section, observed disability prevalence refers to the Census 2021 disability status recorded for individuals within the Statistical Population Dataset 2021. It is important to note that the observed disability prevalence presented here may differ from that reported in Table 2, as the Statistical Population Dataset does not include all individuals enumerated in Census 2021.

The figures in this section have been produced without adjustment for the 14.1 million individuals for whom a disability status could not be derived using the rulesbased approach (see Table 2). Should additional data become available that enable disability status to be estimated for these individuals, the patterns depicted in the charts may differ substantially.

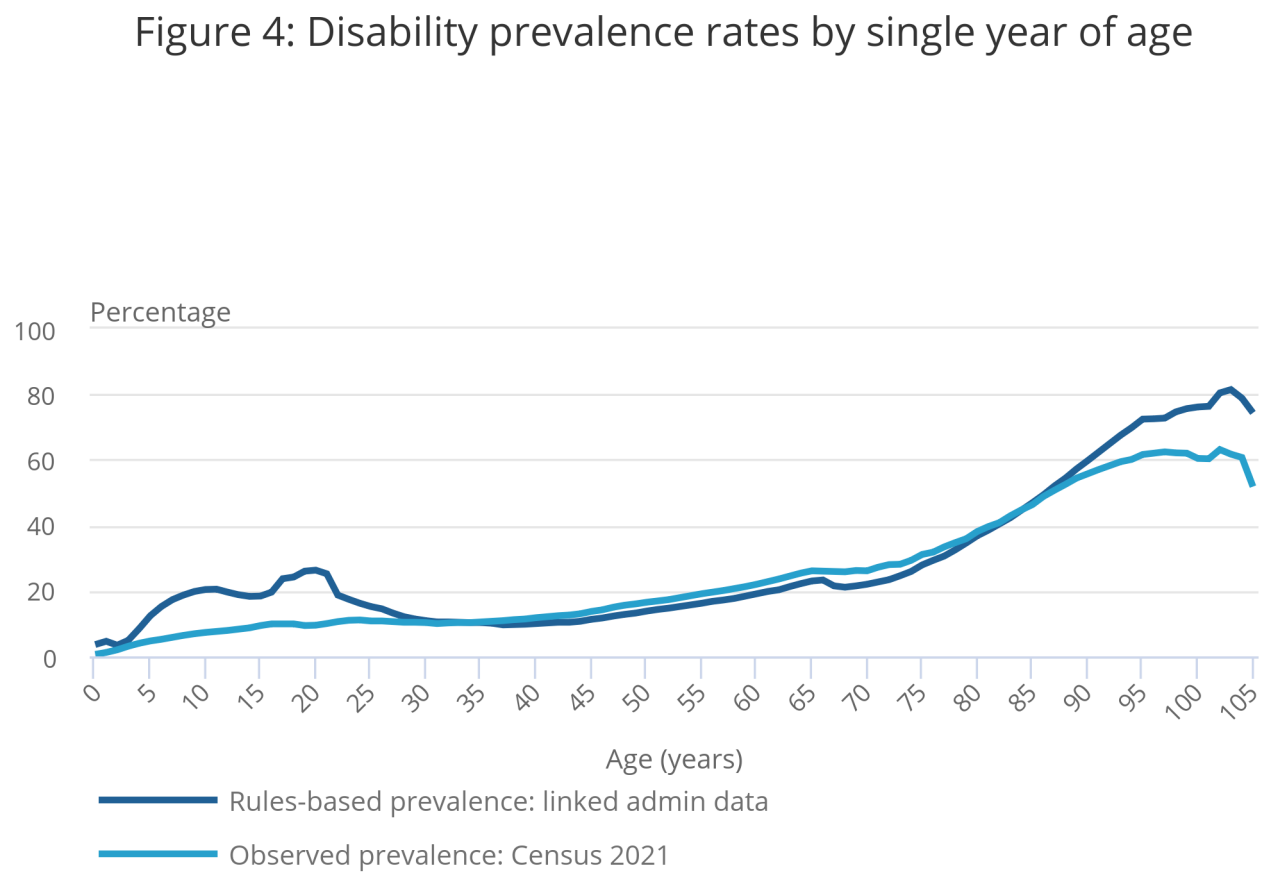
Overall and sex

The rulesbased approach produced an overestimate of disability prevalence of 1.8 percentage points when compared with Census 2021 at the aggregate level. This overestimation was more pronounced for males (2.9 percentage points) than for females (0.7 percentage points).

Age

Figure 4 shows the estimated and observed disability prevalence by single year of age. Marked discrepancies are evident for ages below 30 years and above 85 years, where the estimated prevalence exceeds the observed prevalence. Between these age ranges, the estimated and observed prevalence trends align more closely.

Figure 4: Disability prevalence rates by single year of age



Source: Bespoke data linkage across multiple administrative datasets from the Office for National Statistics

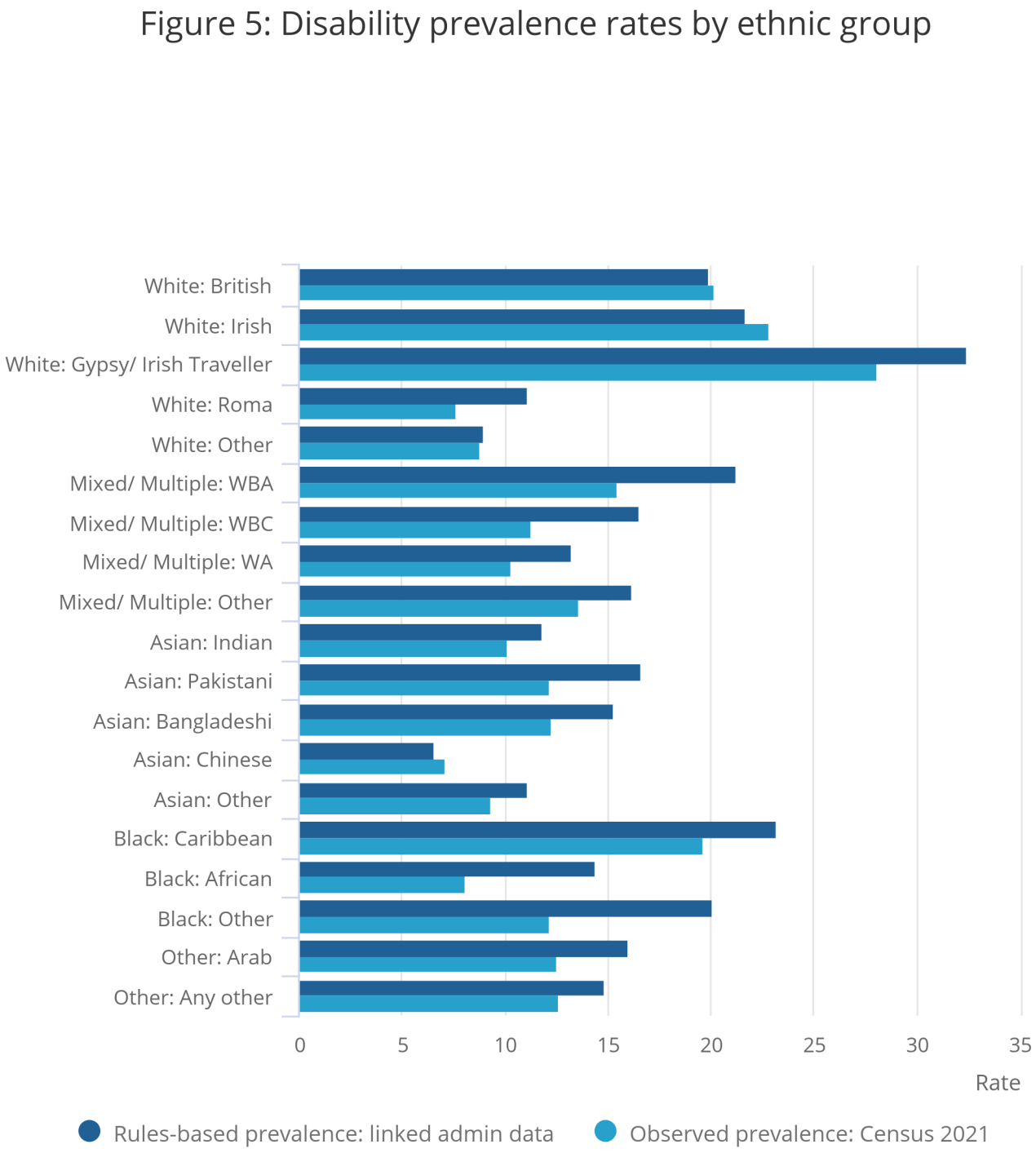
Geography

The rules-based approach shows variability within local authorities, with the City of London having a disability prevalence rate of 8.1% compared with Blackpool having a prevalence rate of 26.2%. The local authority with the biggest difference in prevalence rate using the rules-based approach compared with the observed value being Liverpool, where the rules-based approach overestimates the disability prevalence rate by 5.3 percentage points. When looking at Middle layer Super Output Areas, variance increased with Kensington and Chelsea 012 having a disability prevalence rate of 4.3% compared with Wirral 016 with an estimated rate of 35.0%. For Middle layer Super Output Area, the rules-based approach overestimates the disability prevalence of Nottingham by 13.8 percentage points when comparing the rate with the observed value, the largest difference at this geographical level.

Ethnicity

The rules-based approach overestimated disability prevalence for all but three ethnic groups, these were White: English, Welsh, Scottish, Northern Irish or British, White: Irish and Asian, Asian British or Asian Welsh: Chinese (Figure 5). This overestimation may be the result of age differences for the different ethnic groups, For more information see our Ethnic group by age and sex, England and Wales: Census 2021 article. The degree of overestimation ranged from 0.2 to 7.9 percentage points. Although the approach did not precisely reproduce disability prevalence levels for each ethnic group, it aligned well with the overall pattern of disability prevalence observed across groups.

Figure 5: Disability prevalence rates by ethnic group



Source: Bespoke data linkage across multiple administrative datasets from the Office for National Statistics

Notes:

1. Full labels for ethnic group codes are as follows: White: English, Welsh, Scottish, Northern Irish or British, White: Irish, White: Gypsy or Irish Traveller, White: Roma, White: Other, Mixed or Multiple ethnic groups: White and Black Caribbean (WBC), Mixed or Multiple ethnic groups: White and Black African (WBA), Mixed or Multiple ethnic groups: White and Asian (WA), Mixed or Multiple ethnic groups: Other, Asian, Asian British or Asian Welsh: Indian, Asian, Asian British or Asian Welsh: Pakistani, Asian, Asian British or Asian Welsh: Bangladeshi, Asian, Asian British or Asian Welsh: Chinese, Asian, Asian British or Asian Welsh: Other, Black, Black British, Black Welsh, Caribbean or African: Caribbean, Black, Black British, Black Welsh, Caribbean or African: African, Black, Black British, Black Welsh, Caribbean or African: Other, Other ethnic group: Arab, Other ethnic group: Any other ethnic group.

General health

The rulesbased approach overestimated disability prevalence for very good or good health, by 5.9 and 0.9 percentage points respectively. Conversely, it underestimated disability prevalence among those reporting fair, bad or very bad health, with underestimations of 15.6, 15.1 and 5.2 percentage points, respectively. Despite these discrepancies in magnitude, the estimated data nonetheless replicate the expected pattern of increasing disability prevalence as selfreported general health worsens.

Highest level of qualification

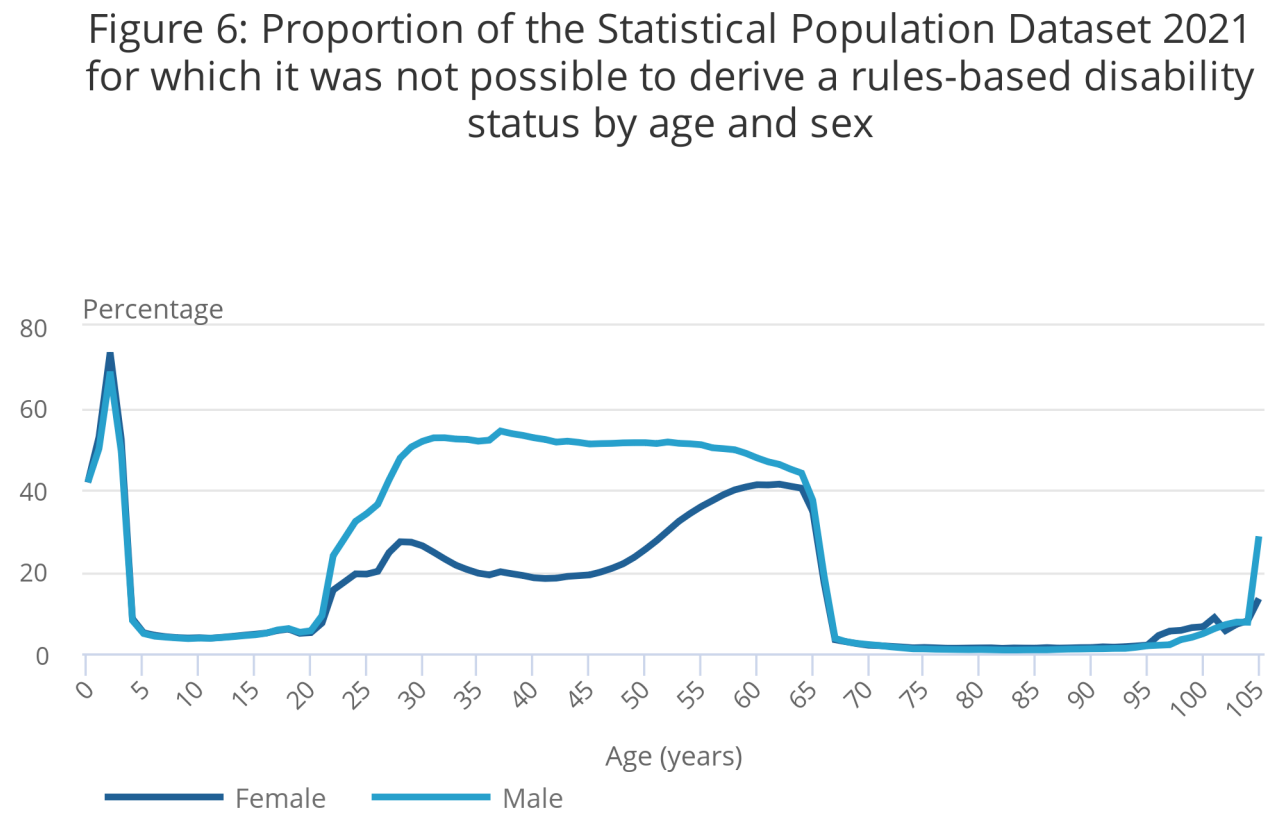
The rulesbased approach underestimated disability prevalence across all qualification levels with the exception of those with no qualifications, for whom disability prevalence was overestimated by 0.3 percentage points. The greatest underestimation was observed for individuals holding "Other" qualifications as their highest level, at 2.6 percentage points. As with other sociodemographic characteristics, although the estimated values do not match observed levels precisely, the overall pattern of disability prevalence by qualification level is maintained.

Missingness

Table 2 indicates that the rulesbased approach was unable to assign a disability status for 14.1 million individuals, representing 24.6% of the population spine. Missingness was more prevalent among males, for whom 31.0% lacked a derived disability status, compared with 18.4% of females.

Figure 6 shows that missingness peaks among preschool children, older adults, and those of working age. While rates were broadly similar for males and females across most age groups, a notably higher proportion of workingage males lacked a derived disability status.

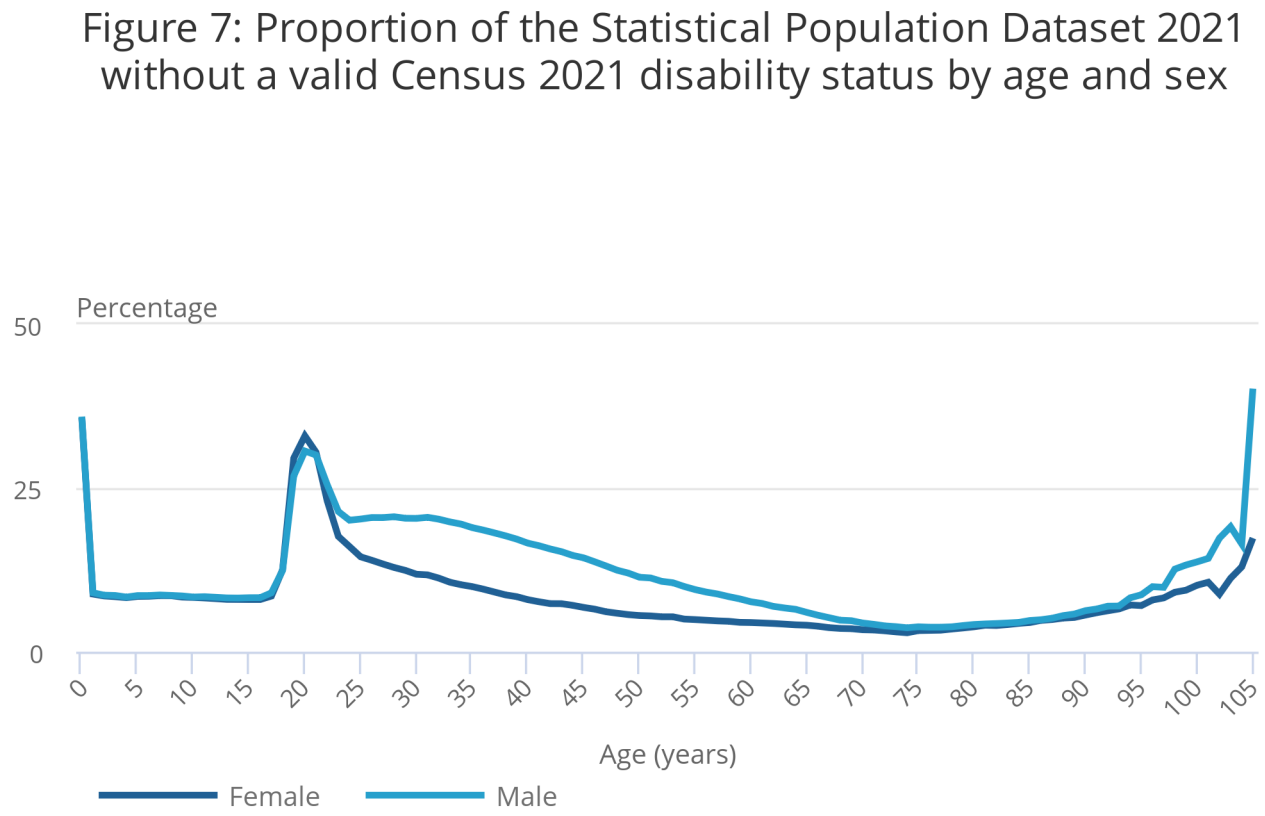
Figure 6: Proportion of the Statistical Population Dataset 2021 for which it was not possible to derive a rules-based disability status by age and sex



Source: Bespoke data linkage across multiple administrative datasets from the Office for National Statistics

Table 3 indicates that more than 6 million individuals either did not link to Census 2021 or lacked a valid Census 2021 disability status, representing 10.5% of the population spine. Of these individuals, the rulesbased approach was able to derive a disability status for 3.2 million; however, accuracy could not be evaluated because of the absence of corresponding observed disability information. Males were more likely than females to be missing a Census 2021 disability status (12.4% compared with 8.6%). Figure 7 further shows that missingness varied substantially by age, with the highest proportions observed among the youngest and oldest age groups, as well as those aged 19 to 22 years.

Figure 7: Proportion of the Statistical Population Dataset 2021 without a valid Census 2021 disability status by age and sex



Source: Bespoke data linkage across multiple administrative datasets from the Office for National Statistics

5 . Discussion and limitations

Administrative data measures of disability

Across the administrative data sources used in this research, direct measures of disability are limited. Consequently, we needed to integrate a wide range of datasets and develop proxy indicators to identify individuals who are "at risk" of disability.

These indicators do not fully align with the Government Statistical Service (GSS) harmonised standard for disability nor with the Equality Act (2010) and Disability Discrimination Act (1995) which reflect the social model of disability. Instead, the indicators available in administrative data align more closely with the medical model, which conceptualises disability as arising from an individual's impairments or conditions rather than the societal barriers that restrict participation. This misalignment has implications for coherence, comparability, and user acceptability, particularly given concerns that medical modelbased approaches may reinforce perceptions that focus on individual limitations rather than environmental or social constraints.

Coverage on the Statistical Population Dataset

Previous work on the Statistical Population Dataset (SPD) version 4.0 has highlighted that it is broadly consistent with Census 2021 at an aggregate level. However, there are challenges relating to overcoverage among younger workingage adults and undercoverage among older workingage adults.

Coverage patterns also vary geographically, with local authority differences reflecting the administrative sources currently available for linkage. At lower output area levels, specific population groups, such as university students, present additional challenges for correctly assigning individuals to addresses. These findings underline the importance of working collaboratively across the Office for National Statistics (ONS) to inform and refine appropriate coverage adjustments for this research. Relevant methodological work on Statistical Population Dataset quality and linkage practices is available in our Understanding the quality of the Statistical Population Dataset in England and Wales using the 2021 Census to Demographic linkage article.

Welsh health data

The Hospital Episode Statistics and NHS Talking Therapies administrative data used in the rules-based approach cover England only. Equivalent Welsh health data do exist but have not been indexed in such a way to allow linkage to the Statistical Population Dataset.

The lack of use of Welsh health data will possibly lead to less people in Wales being identified as likely disabled compared with how many there would have been had health data been used. This may be reflected in Table 2 with the estimated disability prevalence for Wales 2.7 percentage points lower than that measured in Census 2021 and the estimated unknown disability status prevalence for Wales 4.3 percentage points higher compared with England. For any future work, consideration will need to be given to expanding the coverage of health data to include Wales.

Arbitrary disability indicator from Hospital Episode Statistics

The construction of disability indicators from Hospital Episode Statistics (HES) required methodological decisions that were not informed by clinical expertise. Specifically, International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD10) diagnostic blocks were used as proxy indicators for disability.

Through linkage between HES and Census 2021, ICD10 blocks were grouped into three categories based on the proportion of individuals who identified as disabled under the Equality Act definition. Blocks with more than 60% disabled individuals were classified as "likely disabled", while those in the 40% to 60% and below 40% ranges were classified as "likely nondisabled". While pragmatic, this approach may be overly simplistic, as it makes assumptions about the relationship between particular health conditions and selfreported disability. Clinical input and further research on condition–disability relationships would strengthen future iterations of the indicator.

Time lag between administrative record and Census 2021

Administrative records reflect individuals' circumstances at varying points in time, and disability status may change because of medical intervention, social support, or changes in underlying health conditions. The rulesbased method relies on the most recent record available within each dataset, but differences in the timing of these records relative to Census 2021 may affect classification accuracy. This is particularly relevant for indicators that capture historical health conditions or disabilityrelated support that may no longer be current.

Missingness

The rules-based approach was unable to derive a likely disability status for nearly one in four (24.6%) individuals on the population spine. This level of missingness is unlikely to be acceptable for high quality estimates of disability at either the individual or population level. The age profile of individuals missing a derived disability status suggests that a large portion are working-age adults. These adults are not receiving disability benefits and are not found in hospital or psychological treatment datasets, which may suggest that the majority are non-disabled. However, given the large number of individuals unable to link to utilised administrative data, it would be hard to justify a global rule of applying a likely non-disabled status to these individuals.

Accessing further administrative data would likely help with reducing the number of individuals currently unlinked to any of the utilised administrative data. For the working-age population, HM Revenue and Customs Pay As You Earn Real Time Information may contain information on a large portion of those currently without a derived disability status. However, whether the information available is sufficient to allow for a rules-based disability status flag to be derived will need assessing.

Future work would likely benefit from assessing potential additional administrative data sources and their usefulness for a rules-based approach. While we have access to a range of administrative data sources, there are still plenty of data sources that we currently do not have access to. These include Blue Badge data and additional health data, such as prescriptions data, which may be worthwhile pursuing for the purposes of disability statistics.

Estimates at individual, population and subpopulation levels

At an aggregate level, the estimated proportion of individuals flagged as likely disabled (18.0%) closely aligns with the Census 2021 figure (17.5%). However, when restricted to individuals for whom both a rulesbased and Census 2021 disability status were available, only 16.2% were disabled in Census 2021 – 1.8 percentage points lower than the rulesbased estimate. Furthermore, at the individual level, the rulesbased approach correctly identified 62.0% of disabled individuals (sensitivity) and 87.0% of nondisabled individuals (specificity). These rates vary substantially across sociodemographic groups, including age, sex, ethnicity, health status, and geography.

An alternative to individual-level analysis for assessing the rules-based estimates is to compare the aggregated or population-level rules-based disability prevalence with the aggregated Census 2021 disability prevalence for individuals in the Statistical Population Dataset 2021 (see Figures 4 and 5). As with the sensitivity and specificity rates from individual-level analysis, the accuracy of aggregated disability prevalences can vary considerably across subpopulations. For example, the difference between the estimated and observed disability prevalence is greater for males compared with females and greater across Output Areas compared with local authorities.

6 . Future developments

Given the limitations identified in this study, for our next phase of work we will focus on developing a predictive modelling approach to derive disability status. One main challenge of the rulesbased method is the substantial proportion of individuals, 14.1 million, whose disability status could not be determined. Our forthcoming approach will use a predictive logistic regression model to reduce this level of missingness and to improve classification accuracy. We will apply standard modelling principles, including the use of separate training, testing, and validation datasets. As before, Census 2021 disability status will serve as the benchmark outcome, and the Statistical Population Dataset will continue to function as the population spine.

While progressing our model development, we will also continue further methodological refinement through engagement with subjectmatter experts, including specialists in disability statistics, administrative data linkage, and health data. Incorporating additional administrative data sources as they become available will also be a priority. The expansion of data coverage, particularly in areas such as health and social care, has the potential to address current gaps, improve the robustness of disability indicators, and reduce reliance on proxy measures.

7 . Related links

Improving disability data in the UK: 2019
Article | Released 2 December 2019
An introductory article looking at global drivers for improving how we look at disability, including a summary of new analysis on disabled people's lives, and proposals for addressing the gaps in evidence.

Inclusive Data Taskforce Implementation Plan
Web page | Last updated 4 February 2022
Broaden the range of methods that are routinely used and create new approaches to understandingexperiencesacross the population of the UK.

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

8 . Cite this methodology

Office for National Statistics (ONS), released 20 March 2026, ONS website, methodology, Feasibility research: A rules-based approach to estimate disability prevalence using linked administrative data in England and Wales