Producing disability estimates for England using predictive modelling and administrative data

Date: 30th January 2026

1. Executive summary

This paper presents feasibility research into producing regular disability prevalence estimates for England using linked administrative data and predictive modelling. Currently the census is the main source of disability estimates for the population within England.

The paper outlines the datasets used, the predictive modelling approach and the evaluation of model performance. We compared predicted estimates for 2021 with Census 2021 estimates for the whole population as well as for a range of sociodemographic breakdowns. We also assess the relative contribution of administrative data sources to the model. The findings demonstrate that administrative data can provide population level disability estimates that are broadly consistent with census measures, though performance varies across demographic groups and geographic areas.

We seek the Panel's feedback on the modelling approach, the interpretation of results, and priorities for the next phase of development

2. Introduction

Population estimates by disability status for England and Wales are currently derived from survey data (such as the Annual Population Survey) or from census data. With the 10-year gap between censuses and the challenges of reduced survey response rates (ONS 2025a), there is a growing user need for more regular, robust disability estimates.

There is also a widely recognised evidence gap around disability both in the UK and internationally, and more regular estimates are key to resolving that challenge and facilitating research and analysis into outcomes and inequalities for disabled people (Kuper et al 2025; NSIDAC 2024; IDTF 2021). During the COVID-19 pandemic, the ONS investigated differences in mortality rates by disability status, using Census data linked to mortality records (Bosworth et al 2021). The lack of more regular information on disability status meant that 2011 Census had to be used, which was out of date by the time of the pandemic.

As a result of these limitations, ONS has been undertaking research into the feasibility of using administrative data sources to identify disability status and produce population estimates by disability status. The first phase of work explored a 'rules-based' approach to producing estimates, for which we created a linked dataset with disability 'flags' for

each available administrative data source (due for publication 20 March 2026, ONS Working Paper series). While the approach performed well at a population level, it was hampered by the inability to derive a disability status for over 14 million people (mostly working age people who were not present in the administrative records utilised) and was inconsistent at lower geographies and across sub-populations.

The second phase of research, on which this paper reports, assessed the feasibility of using predictive modelling to produce estimates for 2021.

Measuring 'disability'

Defining and measuring disability is complex and there are several 'models' of disability (Zaks 2024). Census measures self-reported disability: 'people who assessed their day-to-day activities as limited by long-term physical or mental health conditions or illnesses' (ONS 2023a). Those with long-term physical or mental health conditions or illnesses could respond that their day-to-day activities were limited 'a lot', 'a little' or 'not at all' (ONS 2023a). This definition of a disabled person meets the current Government Statistical Service's harmonised standard for measuring disability (GSS 2019) and is in line with the Equality Act (2010).

The 'measure' discussed in this paper has some key differences. Unlike census, the model currently produces a binary measure: likely disabled / non-disabled (see 5. Next Steps for discussion of potential multinomial disability status outcome). Additionally, the administrative data sources used do not include a self-defined disability status and the indicators of disability available within them are often (though not always) related to health conditions or impairments. As a result, they are more closely aligned with a 'medical model' of disability (Haegele & Hodge 2016). It is also important to note that though most of the administrative datasets used align more closely to the medical model of disability, the types of disability information collected by each source differ.

3. Methods

Data sources

Table 3.1 summarises the administrative data sources used in this analysis. Because health records were not available for Wales, the analysis was only conducted for people living in England. Some Welsh education data sources were used to capture students living in England but whose place of study was based in Wales. All data sources were deidentified before analysis.

Table 3.1. *Data sources used in the predictive modelling*

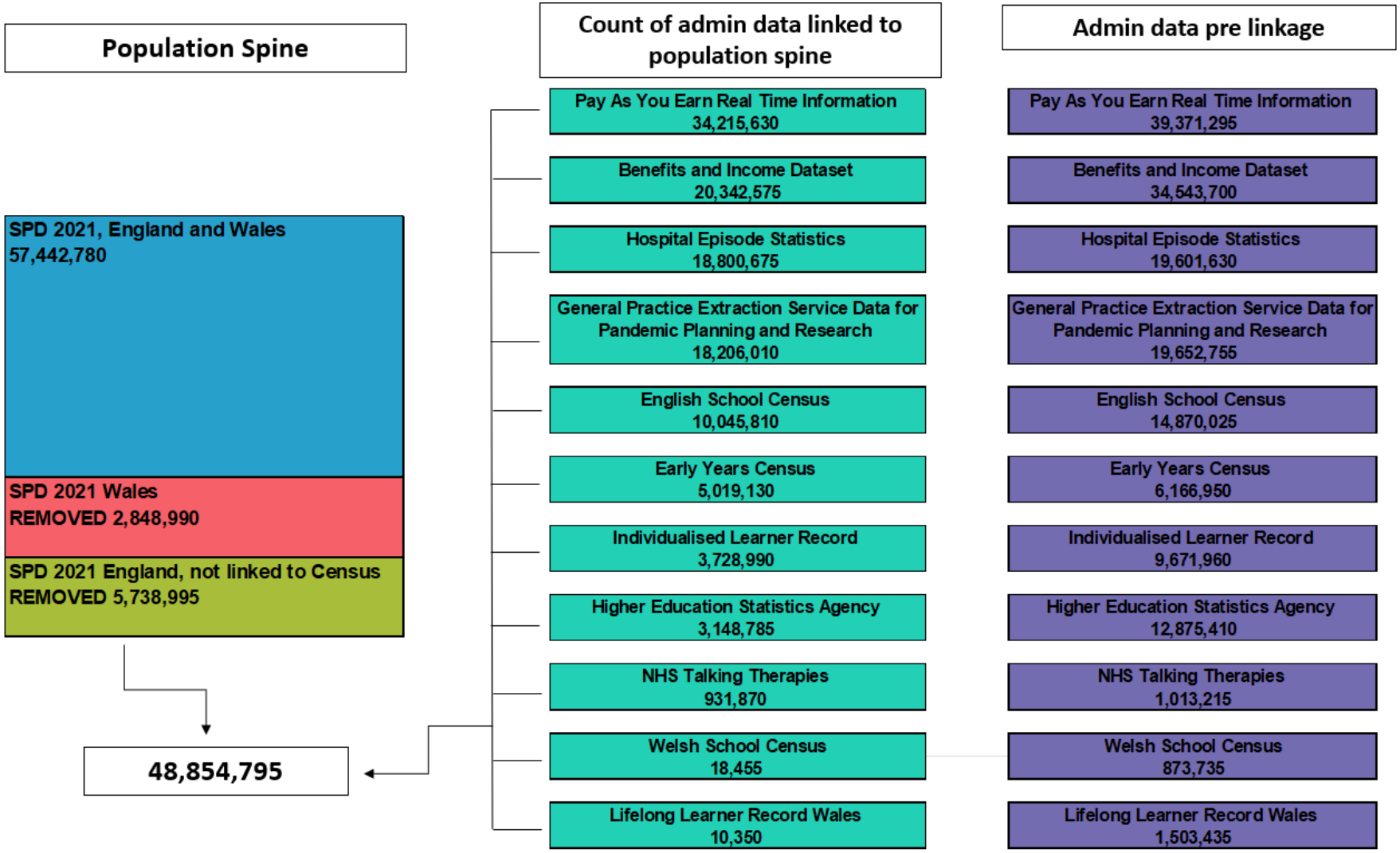
Data source	Time period used
Statistical Population Dataset v4.3	2021 (30th June 2021)
Census 2021	21st March 2021
Demographic Index v4.0.1 and v4.2	N/A
Demographic Index – Census 2021 lookup	N/A
Early Years Census	2001/02 to 2020/21 academic year
English School Census	2010/11 to 2020/21 academic year
Welsh School Census	2010/11 to 2020/21 academic year
Individualised Learner Record	2015/16 to 2020/21 academic year
Lifelong Learning Wales Record	January 2000 to August 2021
Higher Education Statistics Agency	2010/11 to 2020/21 academic year
Hospital Episode Statistics	March 2020 to March 2021
NHS Talking Therapies	2012/13 to 2020/21 tax year
General Practice Extraction Service Data for Pandemic Planning and Research	March 2020 to March 2021
Benefits and Income Dataset	2010/11 tax year to December 2021
Pay As You Earn Real Time Information	2014/15 to 2020/21 tax year

Population spine and linkage

The population used for the model development was England-based residents in the 2021 Statistical Population Dataset version 4.3 (ONS 2023b) who had a valid Census 2021 disability status. The population spine contained 48.9 million individuals after 5.7 million England-based residents were removed for either not linking to a Census 2021 record or having an invalid Census 2021 disability status recorded.

Administrative datasets were linked to this population spine using the Demographic Index (ONS 2023c), either version 4.0.1 or 4.2 depending on the data source. Figure 3.1 demonstrates the size of the population spine and administrative datasets as well as the count of people in the population spine that could be linked to each administrative dataset.

Figure 3.1: *Sample flow*



Outcome and predictor variables used for the predictive modelling

The outcome variable used was the binary disability status (disabled or non-disabled), aggregated from the four-category variable recorded on Census 2021 (ONS 2023a) (see also Appendix 7.1 for further information).

Age (left as a linear term in years), sex (female or male) and local authority district were the three predictor variables used from the Statistical Population Dataset 2021. All the other 337 predictor variables were sourced from the administrative data sources outlined in Table 3.1. Various versions of predictor variables were trialled, and the final selection was based on subject knowledge, distributional checks and checks for collinearity. All but one of the predictor variables from the administrative data sources were in binary format; body mass index being the one exception (see Appendix 7.2 for further information). For some of the administrative data sources listed in Table 3.1, only the latest record available for an individual was used to provide information for the predictive model, while for other administrative data sources, records for an individual were used from a specific restricted period. The use of a restricted time period was due to the date of Census 2021, time period limitations in the available administrative data, computational resource constraints and contrasts in the length of and reason for the duration that an individual is likely to be present in different data sources.

General approach to the predictive modelling

The population dataset was randomly split 50:50 into training and test datasets, both containing approximately 24.4 million records. The training dataset was used to train a logistic regression model with functions from the Spark Machine Learning Library sparklyr package in Cloudera Data Platform. Random forest and gradient boosted tree models were run as part of sensitivity analysis.

As part of the modelling, the training dataset was run through a regularisation and stratified k-fold cross-validation function. The training dataset was split into four folds with each fold containing the same proportions of the outcome variable. For each iteration of modelling, one fold was used as the validation set and the other three were used for training. A range of values of the regularisation parameter (lambda) and the elastic net mixing parameter (alpha) were applied in each iteration of modelling. The final evaluation metric score for each set of alpha and lambda parameters used was the average of the performance scores from all iterations. The area under the precision-recall curve score (PR-AUC) was used as the evaluation metric to decide the optimum alpha and lambda values. The highest PR-AUC was achieved when the alpha and lambda parameters were left unchanged from their default values of 0. The final model, using the optimum alpha and lambda values, was fitted on the whole of the training dataset. The model was used to estimate predicted probabilities of being disabled on the test dataset. To view the predictive modelling code, please click on the following link: predictive modelling code.

Two approaches were used to convert predicted probabilities into population estimates. After the model was trained, it was used to generate a probability of being disabled for each record in the test dataset. Using increments of a single percentage point, the optimum probability threshold for maximising the harmonic mean of precision and recall (F1) score was determined. The optimum threshold was used to create a binary disability status on the test dataset. The use of an optimum threshold to measure disability is referred to as the 'threshold approach'. Additionally, the individual predicted probability scores for disability were summed across various socio-demographic domains (this method is referred to as the 'expected value approach'). Evaluation of the modelling and the approaches involved comparing the predicted disability prevalences with those observed in the test dataset and calculating the following evaluation metrics:

- Sensitivity (disabled as the positive class)
- Specificity (non-disabled as the negative class)
- Positive predictive value (PPV)
- Negative predictive value (NPV)
- phi coefficient
- Harmonic mean of precision and recall score (F1)
- Binary cross-entropy
- Area under the precision-recall curve score (PR-AUC)
- Area under the receiver-operating characteristic curve score (ROC-AUC)

The 'observed' disability status of each individual in the test dataset is the disability status that each individual self-reported on Census 2021, and so the 'observed' disability prevalence relates to Census 2021 disability in charts in the results section. It is important to note that Census 2021 disability in the results charts does not refer to disability prevalence of the whole population of England and Wales measured by Census 2021.

Question for MARP Panel:

- *Our investigations have suggested that there is minimal difference to the results when we apply simple scaling weights to account for the removal of 5.7 million records on the Statistical Population Dataset. Do you have any concerns over presenting the results of our model without applying weights to account for the records removed?*

4. Results

Model evaluation metrics

Table 4.1 shows the evaluation metrics; these metrics provide an indication of how well the model fitted to the training dataset predicts self-reported disability status (from the

census) at an individual level on the test dataset. Metrics which are more focussed on the positive class (e.g., sensitivity, PPV, area under PR curve) are lower than metrics which include a larger focus on the negative class (e.g., specificity, NPV). For this work the positive class refers to disabled and the negative class non-disabled. The sensitivity and specificity metrics in Table 4.1 show a higher specificity rate of 0.929 (correctly identified as non-disabled out of everyone in the test dataset who reported being non-disabled), the sensitivity rate being lower at 0.628 (correctly identified as disabled out of everyone in the test dataset who reported being disabled).

Table 4.1. *Performance evaluation metrics for logistic regression analysis*

Evaluation metric	Value
Sensitivity	0.628
Specificity	0.929
Positive predictive value (PPV)	0.659
Negative predictive value (NPV)	0.920
Harmonic mean of precision and recall (F1) score	0.643
phi coefficient	0.568
Area under precision-recall curve (PR-AUC) (baseline = 0.179)	0.698
Area under receiver-operating characteristic curve (ROC-AUC)	0.877
Binary cross-entropy (baseline = 0.470)	0.303
Optimum threshold	0.27

Note: The positive class and negative class were disabled and non-disabled respectively.

Feature importance of different administrative data variables within the model

Feature importance scores are not readily available for logistic regression models (unlike for tree-based models). All predictor variables used in the model, except age and local authority, are binary. As a proxy for feature importance, we have calculated the absolute t-value for each predictor variable from the different administrative datasets as the magnitude of the division of its coefficient by its standard error. The predictors with the highest absolute t-values are listed in Table 4.2. The p value for all variables listed in Table 4.2 was less than 0.05.

Table 4.2. *The top ten absolute t-values for predictor variables used in the model*

Predictor variable	t-value
Receipt of Personal Independence Payment benefit	722
Receipt of Disability Living Allowance benefit	581
Age (one-year increase)	576
Receipt of Attendance Allowance benefit	474
Receipt of Employment and Support Allowance benefit	424
Hospital diagnosis classified as "General symptoms and signs"	419
Interaction with general practitioner service classified as "Depression diagnosis codes"	349
Interaction with general practitioner service classified as "Asthma-related drug treatment codes"	272
Interaction with general practitioner service classified as "Drug treatment for epilepsy"	266
Mental health condition or difficulties reported in higher education	226

Table 4.2 suggests that currently the receipt of disability benefits and age are the most important features for predicting disability. All the administrative data sources used provided at least some variables with an absolute t-value of 50 or higher, except for Early Years Census. Over 200 variables had absolute t-values less than 20, with most coming from Hospital Episode Statistics and General Practice Extraction Service Data for Pandemic Planning and Research (i.e. primary care records). The predictor variables with absolute t-values closest to zero, and thus the least important predictors within the model, were those related to body mass index. The seemingly low importance scores of many health-related variables could reflect the data quality/missingness and/or coverage of those variables for the individuals within our dataset.

Disability prevalence estimates

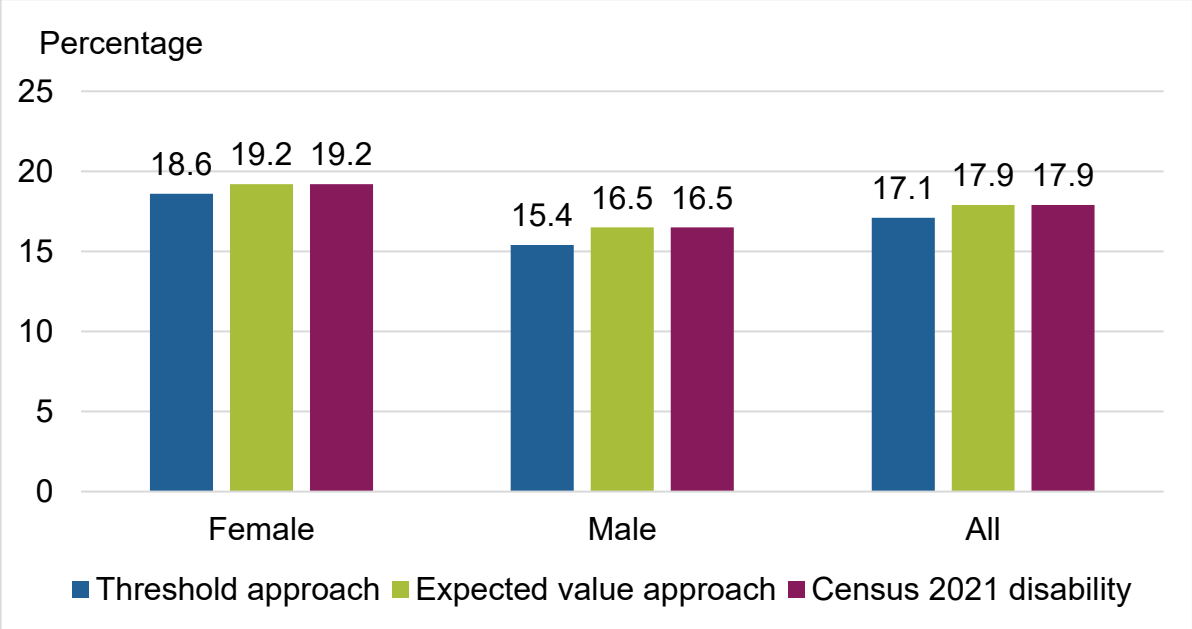
At population level in England our model estimates disability well, but there is variation between estimates produced from the model by the 'expected value' and 'threshold' approaches. The expected value approach matches the observed Census 2021 value in our study population, estimating disability prevalence to be 17.9%. The threshold approach estimates disability prevalence to be 17.1%.

Estimated prevalence of disability by sex

We would expect the predicted disability prevalence to be very close to the observed Census 2021 value when broken by socio-demographic variables included in the model (sex, age, local authority). For prevalence of disability by sex, this is true overall, though

the threshold approach underestimated disability prevalence for females, males and overall, by only 0.6, 1.1 and 0.8 percentage points respectively (Figure 4.1).

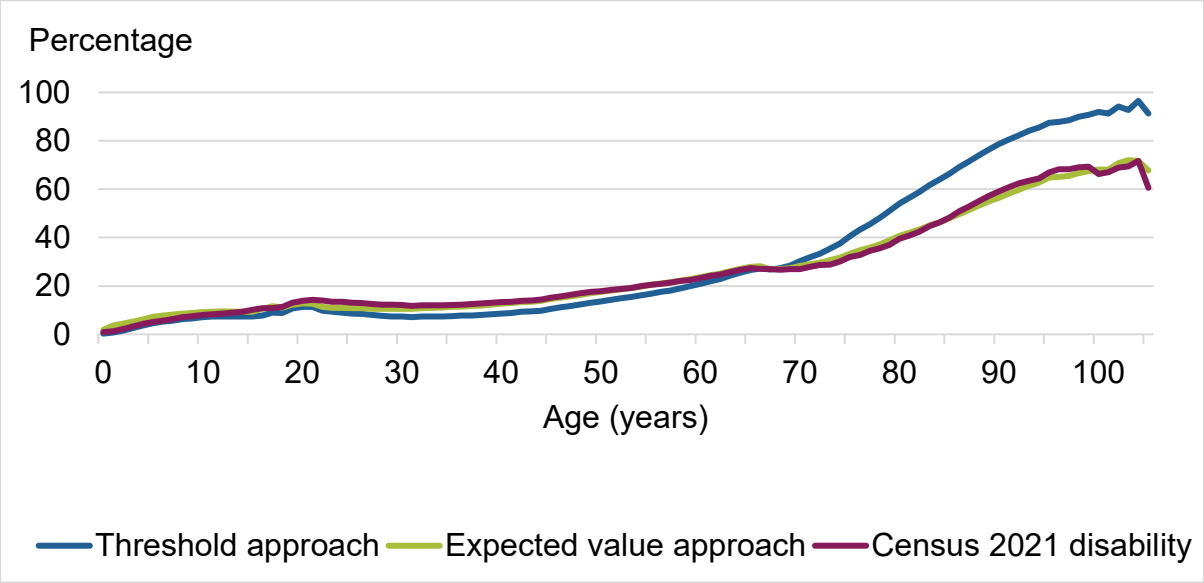
Figure 4.1. *Estimated prevalence of disability by sex*



Estimated prevalence of disability by age

For disability prevalence by single year of age, the biggest difference between the threshold approach and the Census 2021 disability values occurs from 74 years, with disability being overestimated in older ages by the threshold approach (Figure 4.2). The estimated disability prevalence for people aged 78 years and above was at least 10.9 percentage points higher than the observed value, with the biggest difference being for those aged 104 (24.8 percentage points). This difference possibly results from the type of health data sources used (i.e. Hospital Episode Statistics (HES)), as the majority of diagnosis codes originate from inpatient data and those above 75 are overrepresented in hospital admissions compared with the general population.

Figure 4.2. *Estimated prevalence of disability by single year of age*



Geographies

For predicted disability prevalence by local authority, the expected value approach aligns well with Census 2021, but the coherence of the threshold approach and the Census varies by local authority (Appendix 7.4). The threshold approach overestimated disability prevalence for those in East Lindsey and Blackpool (1.9 and 1.4 percentage points) and underestimated it for those in the Isles of Scilly, City of London and Tunbridge Wells (3.6, 2.9 and 2.4 percentage points).

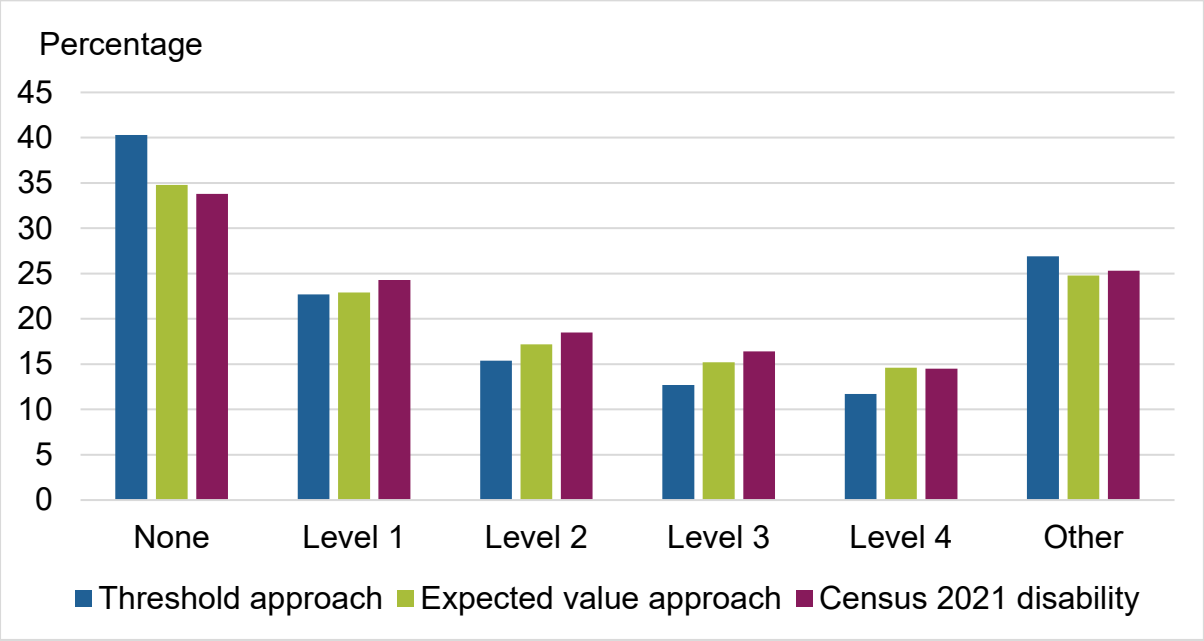
For region, the variation between the prevalence values from the threshold approach and the Census 2021 is smaller. The threshold approach underestimated disability by 1.8 for London and overestimated disability by 0.2 in the North East. Regional geographies were not included as a predictor variable in the model but regional variation will be captured by the inclusion of local authority (which can be aggregated to regions).

Ethnic group, general health and highest level of qualification

A number of additional sociodemographic variables from the 2021 Census were linked to the administrative data, but were not included in the model, so we could assess how the model performed for different sub-groups using the expected value and threshold approaches.

For disability by highest qualification, the expected value approach was closer to the Census 2021 disability values. Predicted prevalence for "Level 1" qualifications (Appendix 7.5) using the expected value approach had the biggest difference, underestimating disability by 1.4 percentage points. Whereas the threshold approach overestimated disability by 6.5 percentage points for those with no qualifications and underestimated disability for Level 1, 2, 3 and 4 (by 1.6, 3.1, 3.7 and 2.8 percentage points respectively).

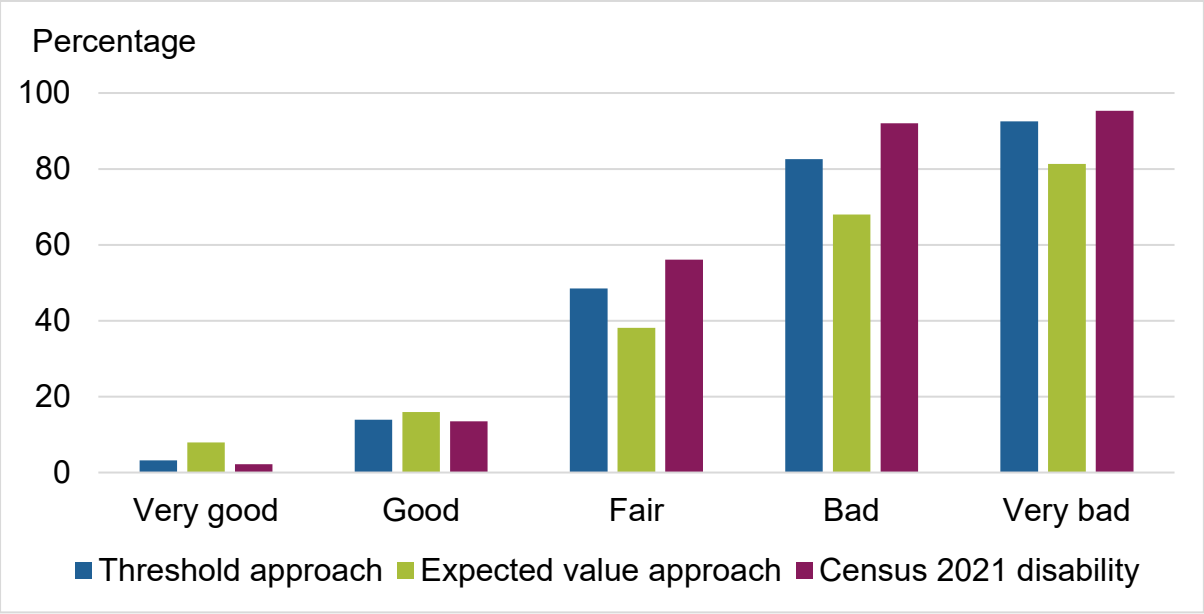
Figure 4.3. *Estimated prevalence of disability by highest level of qualification*



This overall pattern was true for all the additional variables assessed except self-reported health and ethnic group.

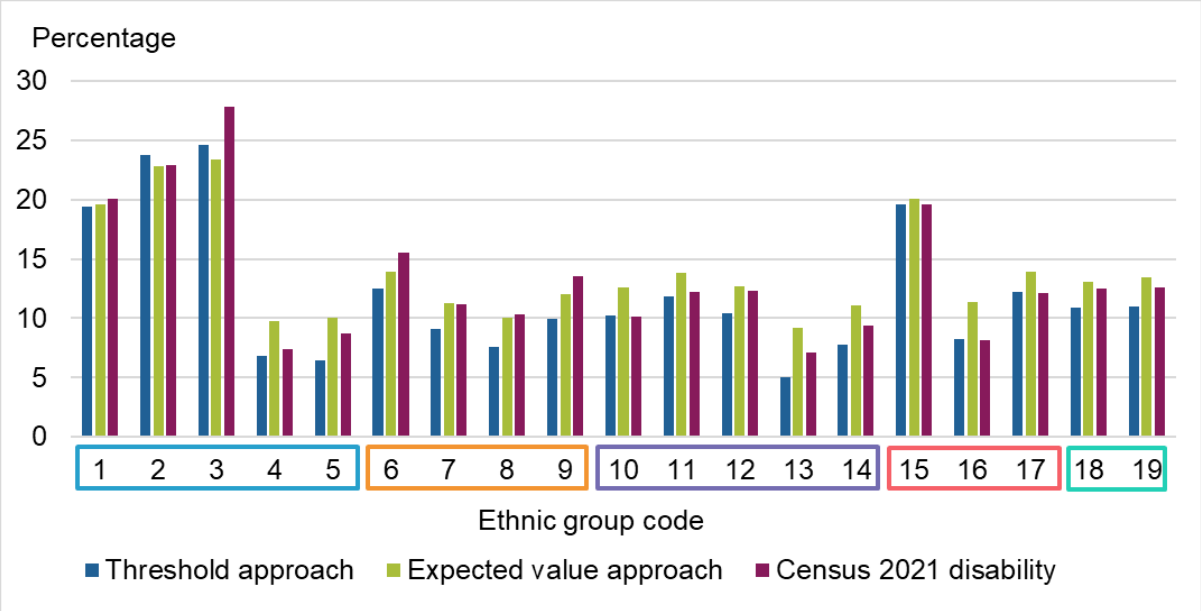
For self-reported health, the threshold approach is better at predicting disability prevalence (Figure 4.4). Across both approaches, the model is better at predicting disability prevalence for those who report "Very good" and "Good" health. The threshold approach is only 1 percentage point from the Census 2021 disability prevalence for "Very good" and 0.4 for "Good", but for "Bad" health the expected value approach has a 24 percentage points difference and the threshold approach a 9.4 percentage points difference.

Figure 4.4. *Estimated prevalence of disability by self-reported health status*



Overall, modelled estimates by ethnic group followed the pattern of distribution of disability prevalence between ethnic groups evident in Census 2021 for both approaches (Figure 4.5). However, with both approaches there were also differences between the predicted disability prevalence and Census 2021 disability for each ethnic group and the differences varied across the ethnic groups.

Figure 4.5. *Estimated prevalence of disability by ethnic group*



White	Mixed or Multiple ethnic groups	Asian, Asian British or Asian Welsh	Black, Black British, Black Welsh, Caribbean or African	Other ethnic group
1 - English, Welsh, Scottish, Northern Irish or British	6 - White and Black Caribbean	10 - Indian	15 - Caribbean	18 - Arab
2 - Irish	7 - White and Black African	11 - Pakistani	16 - African	19 - Any other ethnic group
3 - Gypsy or Irish Traveller	8 - White and Asian	12 - Bangladeshi	17 - Other	
4 - Roma	9 - Other	13 - Chinese		
5 - Other		14 - Other		

Note: Census ethnic group category was used: The ethnic group that the person completing the census feels they belong to. This could be based on their culture, family background, identity or physical appearance (ONS 2023d).

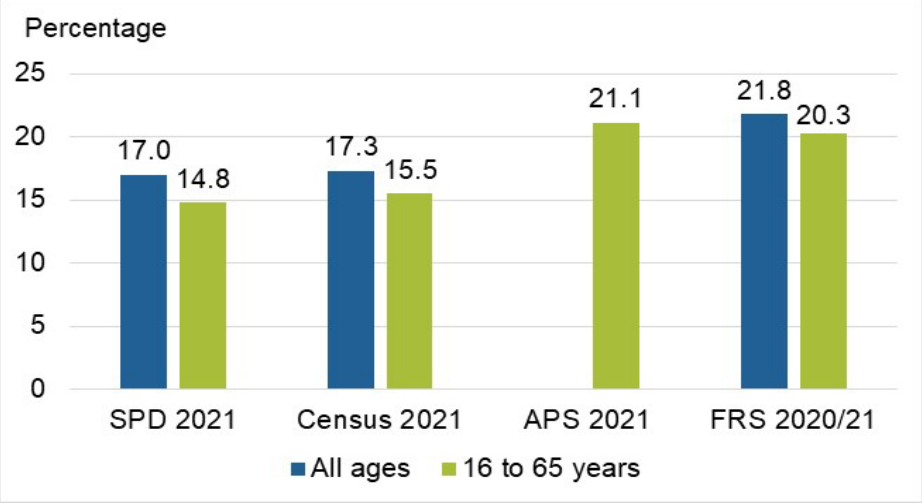
When looking across the predicted disability prevalence by the characteristics not included in the modelling, both the expected value and threshold approaches follow the overall trends of Census 2021. For example, both approaches find that those with no qualifications have the highest prevalence of disability.

However, overall the expected value approach generates disability prevalence estimates closer to the observed Census 2021 value for all levels of geography and for more socio-demographic characteristics than the threshold approach (with the notable exceptions set out above). Given that we are also primarily interested in generating population level estimates (aggregated counts/percentages rather than record-level classifications), this is our preferred approach to generating estimates from the model.

Comparing disability estimates from surveys

Figure 4.6 provides estimates of disability prevalence for England in 2021 from different sources. The results labelled as 'SPD 2021' represent the results from applying the predictive model, using the expected value approach, outlined in the methods section of this paper to the entire sample of the 2021 Statistical Population Dataset. These results are similar to the disability prevalence as measured by Census 2021 for all ages and ages 16 to 65 years. While the modelled and Census 2021 results are within a percentage point of each other, the Annual Population Survey and Family Resources Survey suggest disability prevalence in England in 2021 is higher. The stark difference in disability distribution as measured by Census 2021 and social surveys is a known observation which, at present, does not have an accepted explanation (ONS 2025b). The difference between estimates from census and survey data poses a potential challenge to monitor the accuracy of a predictive model for disability between census years.

Figure 4.6. *Disability prevalence as measured by various data sources, England, 2021*



Notes: SPD 2021 refers to the results of applying the predictive model outlined in this paper to the entire 2021 Statistical Population Dataset. Census 2021 refers to the measured disability status of usual residents as on 21st March 2021. APS 2021 refers to the weighted disability status of respondents to the Annual Population Survey (APS) for the calendar year 2021. FRS 2020/21 refers to the weighted disability status of respondents to the Family Resources Survey (FRS) for the financial year ending 2021. Data from APS and FRS were obtained from the UK Data Service. All data are non-age standardised.

Questions for MARP Panel:

- *We think this work represents a proof-of-concept for how new admin-based disability estimates may be obtained (with appropriate caveats about their limitations, such as they are not recommended for analysis by ethnic group). Does the Panel agree?*
- *Our preferred option for generating estimates from the model is the expected value approach. Does the Panel agree?*
- *Does the Panel think there is value in working to remove variables with lower 'feature importance' to simplify the model?*
- *There is a difference between Census estimates of disability prevalence and those from the APS and FRS, which suggests they are not suitable to monitor the model's performance in non-census years. Do you have any ideas on how we could monitor the performance of the model in the years between censuses?*

5. Next Steps

Comparison over time

As further evaluation of the model, work has begun to produce a time series from 2016 to 2021 of disability prevalence using the predictive model and the Statistical Population Datasets from the years of interest. One difficulty associated with producing such a time series is the health-related administrative data to which we currently have access. The health data are linkable via a Census 2021 identifier, which likely means the further away in time from 2021 that data cover, fewer people on the population spines will link to health data (e.g. due to migration). A possible solution to this is to filter the population spines to contain only people with a valid Census 2021 identifier. The downside of this solution is that it does not allow easy comparison with cross-sectional disability prevalence estimates from other data sources such as the Annual Population Survey or Family Resources Survey. A preferable solution would be to gain access to health-related administrative data which can be linked to the population spines by the Demographic Index, though this may not be a timely solution.

When relevant data becomes available for post-2021 years, we intend to apply the predictive model to post-2021 Statistical Population Datasets and administrative data. This work would be dependent on data availability and accounting for some changes to admin data sources planned by data producers.

Quantifying uncertainty

Quantifying uncertainty for the predicted estimates is challenging. An advantage of using functions within the sparklyr package of the Spark Machine Learning Library is the ability to use, and the efficient processing of, large volumes of data. Unfortunately, Spark Machine Learning Library functions focus more on modelling for the purpose of

predictive analytics than statistical inference, hence standard errors and confidence intervals are not routinely produced by the "off-the-shelf" applications.

Scoping work suggests it is theoretically possible to produce confidence intervals via a bootstrapping approach. However, given the size of the datasets involved, this would likely exhaust the computer memory available to us. An achievable solution to produce confidence intervals (using base R rather than Spark, facilitating the production of prediction intervals) would be to:

1. Take a random, computationally feasible sample of individuals from the spine
2. Fit the predictive model to the down-sampled dataset (with appropriate case weights applied) and obtain the predicted values (p_i) and their standard errors (s_i)
3. For each record on the down-sampled dataset, take a random draw from the normal distribution with mean p_i and standard deviation s_i; multiply by the case weight to convert the person-level outcome probability to corresponding number of people in the population who are estimated to have the outcome; sum these numbers across the dataset; and divide by population size (sum of the case weights) to obtain the estimated outcome prevalence in the population
4. Repeat step 3 many (e.g. 10,000) times to obtain the sampling distribution of the estimated population prevalence
5. Compute the lower and upper limits of a 95% confidence interval as the 2.5th and 97.5th percentiles, respectively, of sampling distribution

Whilst the calculated errors might not represent the full margin of error around the results, they should provide an indication of the level of uncertainty.

Inclusion of data from Wales

As it was not possible to link Welsh health data and the Statistical Population Dataset, Wales is not covered in the predictive model. If suitable linked data become available for use within the life of the project, the coverage could be expanded to include Wales. Consideration is being given to the alignment of equivalent Welsh administrative datasets to those used for England as the approach is developed, and the feasibility of extending the method to Wales will be included in reporting.

Disaggregating disability by degree of limitation of day-to-day activities

The current approach to the predictive model is based on a binary outcome for disability status (disabled or non-disabled). Scoping work will examine whether the predictive model can be expanded to a multinomial disability status outcome, perhaps with the disabled category split by the degree of limitation on day-to-day activities ('a lot' and 'a little') and the non-disabled category split by presence of health conditions ('no health condition' and 'health condition but no limitation on day-to-day activities').

Questions for MARP Panel:

- *Does the Panel think there is a need to produce a measure of uncertainty for the modelled estimates or can they be treated as 'predicted estimates' without such an uncertainty measure (i.e. point estimates only)?*
  - *And if they think a measure is needed, do they have suggestions for how this could be quantified?*
  - *What source(s) of uncertainty should we be attempting to measure? Values from the census are routinely treated as finite-population quantities; that is, they do not have a sampling error associated with them. If the same were to be applied to our predictive estimates, we would only need to capture the uncertainty inherent in using statistical models to estimate disability status, not the uncertainty associated with estimating unobservable "super-population" quantities from an observed sample (i.e. the finite population).*
- *Which 'next step' should we prioritise?*
- *Other than those listed above, are there any other refinements or research avenues that we should consider?*

6. References

Bosworth, M.L., Ayoubkhani, D., Nafilyan, V., Foubert, J., Glickman, M., & Davey, C. (2021). Deaths involving COVID-19 by self-reported disability status during the first two waves of the COVID-19 pandemic in England: a retrospective, population-based cohort study. *The Lancet Public Health, 6*(11), e817-e825. https://doi.org/10.1016/S2468-2667(21)00206-1

Government Statistical Service (GSS), released 25 June 2019 (last updated 26 April 2023), Government Analysis Function website, Guidance, Measuring disability for the Equality Act 2010 harmonisation guidance.

Haegele, J. A., & Hodge, S. (2016). Disability Discourse: Overview and Critiques of the Medical and Social Models. *Quest*, *68*(2), 193–206. https://doi.org/10.1080/00336297.2016.1143849

Inclusive Data Taskforce (IDTF), released 21 September 2021 (last updated 29 December 2022), UKSA website, publications, Inclusive Data Taskforce recommendations report: Leaving no one behind – How can we be more inclusive in our data?

Kuper, H., Shakespeare, T., & Mpanju-Shumbusho, W. (2025). Announcing The *Lancet* Commission on Disability and Health: Creating disability-inclusive health systems that leave no one behind. *The Lancet 406*(10500), 215-216. https://doi.org/10.1016/S0140-6736(25)01041-4

National Statistician's Inclusive Data Advisory Committee (NSIDAC), released 3 October 2024, UKSA website, minutes, National Statistician's Inclusive Data Advisory Committee minutes: 27 March 2024

ONS 2023a: Office for National Statistics (ONS), released 19 January 2023, ONS website, statistical bulletin, Disability, England and Wales: Census 2021

ONS 2023b: Office for National Statistics (ONS), released 28 February 2023, ONS website, article, Developing Statistical Population Datasets, England and Wales - Office for National Statistics

ONS 2023c: Office for National Statistics (ONS), released 28 February 2023, ONS website, article, Understanding quality of the Statistical Population Dataset in England and Wales using the 2021 Census - Demographic Index linkage - Office for National Statistics

ONS 2023d: Office for National Statistics (ONS), Last Updated 28 November 2023, ONS website, article, Ethnic group variable: Census 2021 - Office for National Statistics

ONS 2025a: Office for National Statistics (ONS), released June 2025, ONS website, article, ONS Survey Improvement and Enhancement Plan for Economic Statistics - Office for National Statistics

ONS 2025b: Office for National Statistics (ONS), released 16 September 2025, ONS website, article, Labour Force Survey quality update: September 2025

Zaks, Z. (2024). Changing the medical model of disability to the normalization model of disability: clarifying the past to create a new future direction. *Disability & Society*, *39*(12), 3233–3260. https://doi.org/10.1080/09687599.2023.2255926

7. Appendices

Appendix 7.1. Census self-reported disability categories according to whether flagged as disabled for our binary classification for predictive modelling

Flagged as disabled	Disabled under the Equality Act: Day-to-day activities limited a lot
	Disabled under the Equality Act: Day-to-day activities limited a little
Flagged as not disabled	Not disabled under the Equality Act: Has long-term physical or mental health condition but day-to-day activities are not limited
	Not disabled under the Equality Act: No long-term physical or mental health conditions

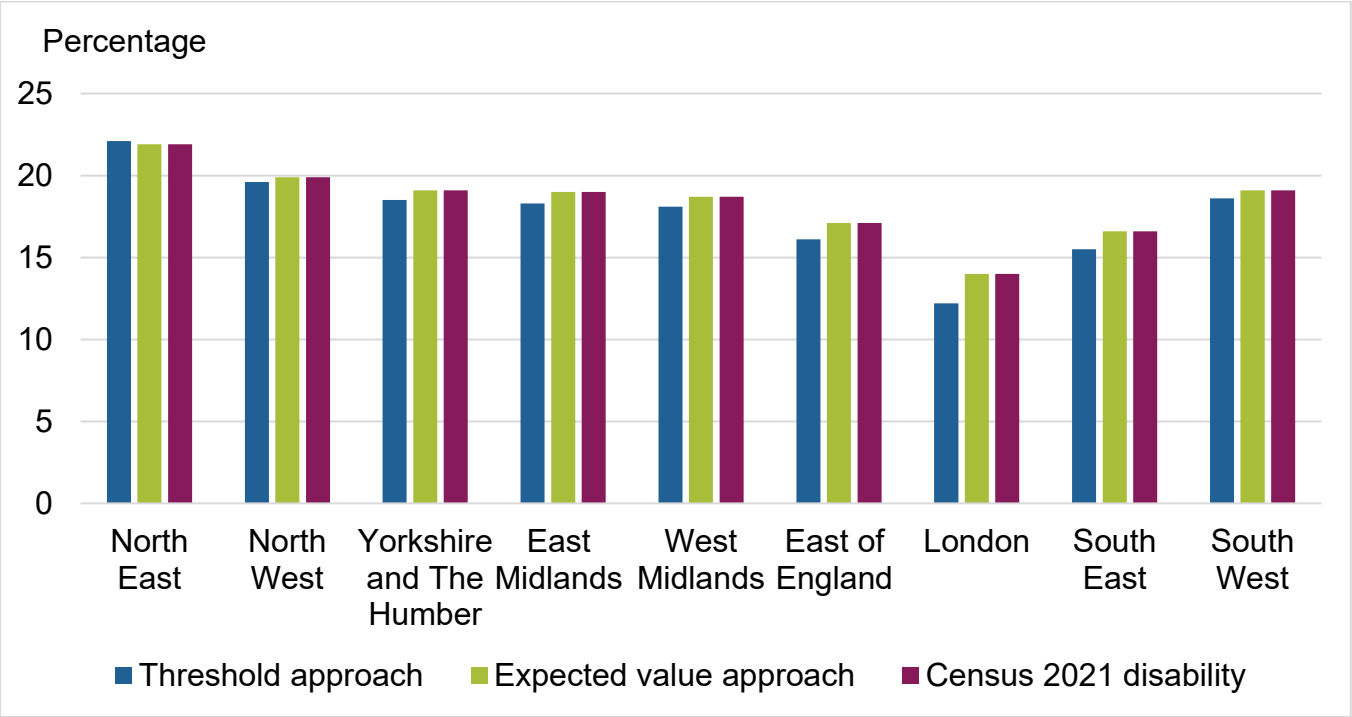
Appendix 7.2. Predictor variables from administrative data sources used in the predictive modelling

Statistical Population Dataset 2021	
	Age (single year)
	Sex (female or male)
	Local authority district
Early Years Census	
	Two binary variables indicating whether the individual received different types of special educational needs provision or not in their latest academic year on record
	Two binary variables indicating whether the education provider received disability-related funding for the individual or not (combined with the equivalent variables from the English School Census) in their latest academic year on record
English School Census and Welsh School Census	

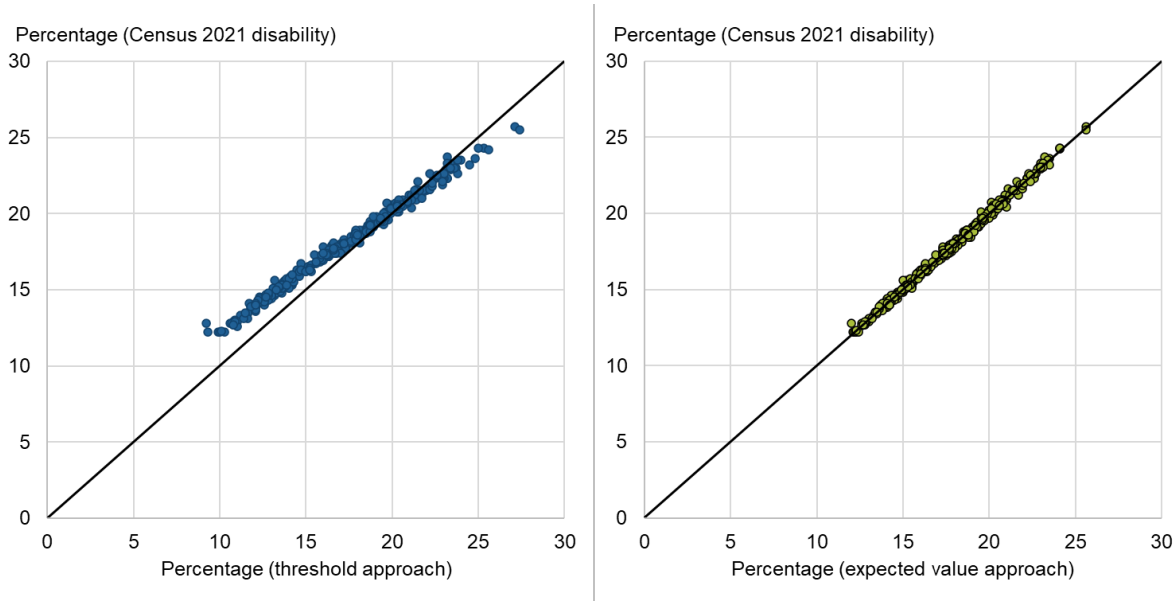
	13 binary variables indicating whether the individual was reported to have different types of special education needs or not in their latest academic year on record
	Two binary variables indicating whether the education provider received disability-related funding for the individual or not (combined with the equivalent variables from the Early Years Census) in their latest academic year on record
Individualised Learner Record and Lifelong Learner Record Wales	
	18 binary variables indicating whether the individual was reported to have or not have different types of disability, learning difficulty or health condition in their latest academic year on record
Higher Education Statistics Agency	
	Nine binary variables indicating whether the individual was reported to have different types of disability or not in their latest academic year on record
NHS Talking Therapies	
	11 binary variables indicating whether the individual was reported to have different types of disability or not in their latest tax year on record
Hospital Episode Statistics	
	195 binary variables using combined inpatient and outpatient data indicating whether the individual had been diagnosed or not with any health-related problems defined by blocks of the International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10) between March 2020 and March 2021
General Practice Extraction Service Data for Pandemic Planning and Research	
	78 binary variables indicating whether the individual has interacted or not with their general practitioner service for different reasons classified by Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) clusters between March 2020 and March 2021
	Body mass index category (underweight, healthy weight, overweight, obese, unknown) as measured between March 2020 and March 2021
Benefits and Income Dataset	
	Six binary variables indicating whether the individual has received a payment of any of six disability-related benefits or not (Personal Independence Payment, Disability Living Allowance, Attendance Allowance, Severe Disablement Allowance, Incapacity Benefit, Employment and Support Allowance) between March 2020 and December 2021
Pay As You Earn Real Time Information	

	A binary variable indicating whether the individual worked as an employee with the employment starting before April 2021 and being active at some point from March 2020
	A binary variable indicating whether the individual received a serious ill health lump sum payment from an employer at any point before April 2021 or not

Appendix 7.3. *Estimated prevalence of disability by r*egion



Appendix 7.4 *Estimated prevalence of disability by local authority*



Appendix 7.5 Highest level of qualification variable descriptions

None	No qualifications
Level 1	Level 1 and entry level qualifications: 1 to 4 GCSEs grade A* to C, Any GCSEs at other grades, O levels or CSEs (any grades), 1 AS level, NVQ level 1, Foundation GNVQ, Basic or Essential Skills
Level 2	Level 2 qualifications: 5 or more GCSEs (A* to C or 9 to 4), O levels (passes), CSEs (grade 1), School Certification, 1 A level, 2 to 3 AS levels, VCEs, Intermediate or Higher Diploma, Welsh Baccalaureate Intermediate Diploma, NVQ level 2, Intermediate GNVQ, City and Guilds Craft, BTEC First or General Diploma, RSA Diploma
Level 3	Level 3 qualifications: 2 or more A levels or VCEs, 4 or more AS levels, Higher School Certificate, Progression or Advanced Diploma, Welsh Baccalaureate Advance Diploma, NVQ level 3, Advanced GNVQ, City and Guilds Advanced Craft, ONC, OND, BTEC National, RSA Advanced Diploma
Level 4	Level 4 qualifications or above: degree (BA, BSc), higher degree (MA, PhD, PGCE), NVQ level 4 to 5, HNC, HND, RSA Higher Diploma, BTEC Higher level, professional qualifications (for example, teaching, nursing, accountancy)
Other	Other: vocational or work-related qualifications, other qualifications achieved in England or Wales, qualifications achieved outside England or Wales (equivalent not stated or unknown) and Apprenticeship