ABSTRACT

Background To gain maximum insight from large administrative healthcare datasets it is important to understand their data quality. Although a gold standard against which to assess criterion validity rarely exists for such datasets, internal consistency can be evaluated. We aimed to identify inconsistencies in the recording of mandatory International Statistical Classification of Diseases and Related Health Problems, tenth revision (ICD-10) codes within the Hospital Episodes Statistics dataset in England.

Methods Three exemplar medical conditions where recording is mandatory once diagnosed were chosen: autism, type II diabetes mellitus and Parkinson’s disease dementia. We identified the first occurrence of the condition ICD-10 code for a patient during the period April 2013 to March 2021 and in subsequent hospital spells. We designed and trained random forest classifiers to identify variables strongly associated with recording inconsistencies.

Results For autism, diabetes and Parkinson’s disease dementia respectively, 43.7%, 8.6% and 31.2% of subsequent spells had inconsistencies. Coding inconsistencies were highly correlated with non-coding of an underlying condition, a change in hospital trust and greater time between the spell with the first coded diagnosis and the subsequent spell. For patients with diabetes or Parkinson’s disease dementia, the code recording for spells without an overnight stay were found to have a higher rate of inconsistencies.

Conclusions Data inconsistencies are relatively common for the conditions considered. Where these mandatory diagnoses are not recorded, there is potential for this to impact on the care provided. This study should motivate the improvement of clinical coding for all conditions with mandatory diagnosis recording.

INTRODUCTION

Decision-making by clinicians and healthcare service managers is increasingly being informed by large-scale administrative healthcare data. Although such data are observational and often lack clinical details, they can support decision-making, particularly in cases where other research methods (eg, randomised controlled trial) may be considered unethical or impractical. Where such data cover an entire population of interest, they can also help minimise the risk of bias due to unrepresentative patient selection criteria (collider bias). However, it is important to have a clear understanding of data quality and the strengths and limitations of any dataset prior to analysis.

In England, the Getting It Right First Time (GIRFT) programme is a National Health Service (NHS) England and NHS Improvement initiative with a remit to reduce unwarranted variation in clinical practice that negatively impacts on patient outcomes. The GIRFT programme is one of the largest users of administrative healthcare data for clinical outcome measurement in the UK and has a particular interest in data quality. A key data resource for the GIRFT programme is the Hospital Episodes Statistics (HES) dataset, which contains data for all hospital admissions of NHS patients in England.

The aim of this exploratory study was to identify the extent of, and data features associated with, data inconsistencies within the HES administrative dataset for England.
METHODS

Study design and data collection

This was a retrospective exploratory analysis of HES data. HES data are collected by NHS Digital for all NHS-funded patients admitted to hospitals in England. Hospital trusts run all NHS hospitals in England. A hospital trust is an administrative unit of, typically, one to four hospitals which provides secondary and/or tertiary care for all people living in a geographically defined catchment area. HES includes data for patients funded by the NHS but receiving treatment in a non-NHS hospital. Data collection and reporting is mandatory for NHS funded patients. Data are taken from clinical notes and discharge summaries and data are entered by trained clinical coders at each trust working to a national data standard.6 Extracts from HES data are audited against clinical audit in a small number of trust each year.

Data regarding pre-existing diagnoses would only be recorded by a coder if detailed in the medical notes or discharge summary, and all clinicians receive training in the importance of accurately recording data. Although autocoding of data is becoming more common in the NHS, its use in the period covered by our study was very limited.

HES data are primarily collected for the purposes of reimbursement. However, their value as a research resource and to inform policy decisions is being increasingly recognised.7

In HES, a hospital spell is defined as a continuous period in hospital from admission to discharge. A spell can include multiple smaller episodes of care in various hospital settings and under different consultants. As an example, following an emergency department attendance, a patient may initially be under the care of acute medicine (episode one), then transferred to a critical care setting (episode two) and then to a care of the elderly ward (episode three) prior to discharge. Spells involving transfers to other trusts were analysed as separate spells.

Timing, case ascertainment, inclusion and exclusion criteria

Data were taken from HES for all patient discharges during the period 1 April 2013 to 31 March 2021. Using International Statistical Classification of Diseases and Related Health Problems, tenth revision (ICD-10) codes three separate exemplar datasets were extracted for patients with a diagnosis of: childhood autism (F84.0), atypical autism (F84.1) and Asperger’s syndrome (F84.5); type II diabetes mellitus with peripheral circulatory complications (DMPC; E11.5) and Parkinson’s disease dementia (PDD; F02.3). ICD-10 codes allow data to be captured and defined consistently over time and across settings. There have been no major changes in ICD-10 coding guidance for these conditions over the study period.

DMPC and PDD were selected as representative of patients within the broader disease categories of diabetes mellitus and dementia, respectively.

These conditions were chosen for several reasons:

1. Recording of these conditions is mandated by NHS Digital and NHS England for all subsequent hospital episodes once a diagnosis has been made.8
2. The conditions have typical onset in childhood (autism), midlife (DMPC) and late-life (PDD) and so cover a range of demographic groups.
3. All tend to be lifelong once present, accepting that DMPC and PDD are representatives of broader conditions and that the details of the diagnosis may change within these broad definitions.

The first use of the specified code in the diagnostic record for a hospital spell during the study period was identified (index spell) and data for all subsequent spells for the same person extracted.

Spells were removed from the datasets if:
1. The only ICD-10 code present in the record was R69 (unknown and unspecified causes of morbidity) or there was no valid entry in the diagnostic code field.
2. The spell was a regular attendance for renal or liver dialysis (Office of Population Censuses and Surveys Classification of Interventions and Procedures version 4 code X40 or X43; or other regular attendance with ICD-10 code N185 (chronic renal failure) present. Regular day-attendances are usually for a specific procedure, and in most cases only that procedure and the related diagnosis is coded. Inclusion of these spells would unduly bias the dataset.
3. Patients were in age bands where the initial coding diagnoses were most likely miscoded: we removed patients with PDD younger than 40 years, and patients with DMPC younger than 18 years. The data extraction and cleaning procedure for each dataset is summarised in online supplemental figures S1–S3.

Identification of data inconsistencies

All data inconsistencies are reported at the spell level. A subsequent spell was considered consistent with the first spell if at least one of its constituent episodes mentioned the ICD-10 codes listed below for that condition:

- Autism: F84.0, F84.1 or F84.5.
- DMPC (representing the broader disease category of diabetes mellitus): E10-, E11-, E14-.
- PDD (representing the broader disease category of dementia): F00-, F01-, F02-, F03-, F05.1, G30.1, G30.8, G309.

Further details on the definitions of these conditions are summarised in online supplemental table S1. In the case of DMPC and PDD, a broader definition of the condition was used for subsequent spells than for the first spell. This was in recognition of the fact that details may not be recorded regarding the diabetes subtype or its presentation or the exact role of Parkinson’s disease in the development of dementia.

Covariates and data features/characteristics

Patient characteristics: sex, age in years, ethnicity (white, black or black British, Asian or Asian British, mixed, other and not stated), comorbidities (Charlson Comorbidity Index or the exact role of Parkinson’s disease in the development of dementia.

Patient characteristics: sex, age in years, ethnicity (white, black or black British, Asian or Asian British, mixed, other and not stated), comorbidities (Charlson Comorbidity Index or the exact role of Parkinson’s disease in the development of dementia.

Patient characteristics: sex, age in years, ethnicity (white, black or black British, Asian or Asian British, mixed, other and not stated), comorbidities (Charlson Comorbidity Index or the exact role of Parkinson’s disease in the development of dementia.

Patient characteristics: sex, age in years, ethnicity (white, black or black British, Asian or Asian British, mixed, other and not stated), comorbidities (Charlson Comorbidity Index or the exact role of Parkinson’s disease in the development of dementia.

Patient characteristics: sex, age in years, ethnicity (white, black or black British, Asian or Asian British, mixed, other and not stated), comorbidities (Charlson Comorbidity Index or the exact role of Parkinson’s disease in the development of dementia.

Patient characteristics: sex, age in years, ethnicity (white, black or black British, Asian or Asian British, mixed, other and not stated), comorbidities (Charlson Comorbidity Index or the exact role of Parkinson’s disease in the development of dementia.

Patient characteristics: sex, age in years, ethnicity (white, black or black British, Asian or Asian British, mixed, other and not stated), comorbidities (Charlson Comorbidity Index or the exact role of Parkinson’s disease in the development of dementia.

Patient characteristics: sex, age in years, ethnicity (white, black or black British, Asian or Asian British, mixed, other and not stated), comorbidities (Charlson Comorbidity Index or the exact role of Parkinson’s disease in the development of dementia.

Patient characteristics: sex, age in years, ethnicity (white, black or black British, Asian or Asian British, mixed, other and not stated), comorbidities (Charlson Comorbidity Index or the exact role of Parkinson’s disease in the development of dementia.

Patient characteristics: sex, age in years, ethnicity (white, black or black British, Asian or Asian British, mixed, other and not stated), comorbidities (Charlson Comorbidity Index or the exact role of Parkinson’s disease in the development of dementia.

Patient characteristics: sex, age in years, ethnicity (white, black or black British, Asian or Asian British, mixed, other and not stated), comorbidities (Charlson Comorbidity Index or the exact role of Parkinson’s disease in the development of dementia.

Patient characteristics: sex, age in years, ethnicity (white, black or black British, Asian or Asian British, mixed, other and not stated), comorbidities (Charlson Comorbidity Index or the exact role of Parkinson’s disease in the development of dementia.

Patient characteristics: sex, age in years, ethnicity (white, black or black British, Asian or Asian British, mixed, other and not stated), comorbidities (Charlson Comorbidity Index or the exact role of Parkinson’s disease in the development of dementia.

Patient characteristics: sex, age in years, ethnicity (white, black or black British, Asian or Asian British, mixed, other and not stated), comorbidities (Charlson Comorbidity Index or the exact role of Parkinson’s disease in the development of dementia.

Patient characteristics: sex, age in years, ethnicity (white, black or black British, Asian or Asian British, mixed, other and not stated), comorbidities (Charlson Comorbidity Index or the exact role of Parkinson’s disease in the development of dementia.

Patient characteristics: sex, age in years, ethnicity (white, black or black British, Asian or Asian British, mixed, other and not stated), comorbidities (Charlson Comorbidity Index or the exact role of Parkinson’s disease in the development of dementia.

Patient characteristics: sex, age in years, ethnicity (white, black or black British, Asian or Asian British, mixed, other and not stated), comorbidities (Charlson Comorbidity Index or the exact role of Parkinson’s disease in the development of dementia.
Features of hospital stay: Spell length of stay, admission method (emergency or elective), main specialty, number of days since the first spell with the diagnosis recorded (reported as the difference between the discharge date of the first spell and the admission date of the subsequent spell), change of trust between the first and subsequent spell, change of clinical specialty between the first and subsequent spell.

Coding of underlying conditions: We identified spells where a related condition would be expected to also be diagnosed. For PDD this was Parkinson’s disease (ICD-10 code G20), and for autism, whether learning disability (ICD-10 codes F70-, F71-, F72-, F73-, F78-, F79-, F80-, F81-, F82- or F83-) was also mentioned in the diagnostic record. The Parkinson’s disease code is not mandatory, although the learning disability codes are mandatory.

**Outcome (target) variable**

For each condition, the target was described by a binary flag indicating whether a code was recorded in the subsequent spell.

**Data analysis**

Data were extracted onto a secure encrypted server controlled by NHS England and NHS Improvement. Analysis within this secure environment took place using Alteryx 2019.3 (Alteryx, Irvine, California, USA), Python V.3.9.6 and the scikit-learn machine learning library V.1.0.1 (Python Software Foundation, Beaverton, Oregon, USA).

Important predictors associated with data inconsistencies were identified using a random forest classifier algorithm (briefly described in online supplemental figure S4). Missing data values were handled by imputation with the mean or mode in each class. The datasets were separated into a training, validation and test sets with 70%, 15% and 15% of data respectively. Machine learning algorithms require the data to be randomly split so that the algorithm can learn the relationships between the data points and then apply this learning to an unseen part of the data set. The algorithm parameters were determined using the validation set by performing a randomised search on a grid of values and choosing the ones that led to the highest value for the area under the precision recall curve. The classifiers were then trained on the training set and evaluated on the withheld test set. The final parameters of each classifier are summarised in online supplemental table S2.

The models’ most important predictors were identified using the SHapley Additive exPlanation (SHAP) feature importance to minimise bias towards high-cardinality variables. Positive or negative correlations of predictors with coding inconsistencies were estimated by calculating the Kendall Tau-b correlation coefficients between the values of the variables, and their estimated Shapley values. These were calculated using TreeSHAP, an efficient estimation approach for tree-based models. Model performance was evaluated using the area under the receiver operating characteristics (AUROC) curve, precision-recall curves and precision gain—recall gain curves.

In subanalyses, we evaluated the impact of time from the first spell on the proportion of inconsistencies. Time from admission for the first spell where the diagnostic code was used to admission for a subsequent spell was calculated in days for the subset of patients where the first spell was prior to 1 April 2018. The follow-up period was set at 3 years for all patients. This was done to avoid a potential bias due to varying maximum follow-up periods for each patient.

**RESULTS**

Data were available for 172,324 unique patients with autism, 106,943 unique patients with DMPC and 27,794 unique patients with PDD. The characteristics of these patients on their first spell during the study period are summarised in table 1 together with the number of patients without data recorded for each feature. Autism patients had the youngest and patients with PDD the oldest age structure. The autism and DMPC dataset had a high proportion of patients from more deprived areas.

The number of subsequent spells for each patient within a 3-year follow-up period are shown in online supplemental figure S5 for each condition. High numbers of patients (more than 50% for patients with autism) had no subsequent spells within 3 years of their first spell. Patients with DMPC had the highest numbers of subsequent spells. Figure 1 summarises the number of data inconsistencies in these subsequent spells up to 3 years from the first spell where the diagnostic code was used. The number of data inconsistencies increased with time from the first spell, although the trend was less obvious after approximately 20 weeks. Figure 2 illustrates the percentage of subsequent spells with missing mandatory codes in the 3 years after the first spell. The consistency of the coding for PDD appeared to broadly improve over the study period, while for autism patients, consistency appears to have decreased slightly over time.

The number of subsequent spells with data inconsistencies were 170,447 (43.7%) for patients with autism, 46,679 (8.6%) for DMPC patients and 18,975 (31.2%) for patients with PDD. The number of subsequent spells with inconsistencies according to patient characteristics is summarised in table 2. For people with autism, data inconsistencies became more common with greater age. However, for PDD inconsistencies became less common with greater age. Females with autism and PDD had a noticeably higher proportion of inconsistencies than males. There was a modest trend towards a higher proportion of data inconsistencies in autism patients with...
increasing deprivation. White patients had the highest rate of inconsistencies for autism.

The variation in data inconsistencies across trusts in England is summarised in online supplemental figure S6.

There was substantial spread in terms of data inconsistencies across trusts.

Three random forest classifiers were optimised and trained to identify coding inconsistencies for each

---

**Table 1** Table of patient characteristics on first spell within the study period

|                          | Autism      | Diabetes mellitus with peripheral complications | Parkinson’s disease dementia |
|--------------------------|-------------|-----------------------------------------------|-----------------------------|
| No of patients           | 172324      | 106943                                        | 27794                       |
| Age band                 |             |                                               |                             |
| 0–17                     | 98591 (57.2 %) | 8 (0.01 %)                                   | 0 (0.0 %)                   |
| 18–39                    | 50682 (29.4 %) | 1085 (1.0 %)                                 | 12 (0.04 %)                 |
| 40–59                    | 16060 (9.3 %)  | 21745 (20.3 %)                               | 279 (1.0%)                  |
| 60–79                    | 6171 (3.6 %)  | 55050 (51.5 %)                               | 12375 (44.5 %)              |
| 80 years and over        | 820 (0.5 %)  | 28938 (27.1 %)                               | 15111 (54.4 %)              |
| Not recorded             | 0           | 117                                           | 17                          |
| Sex                      |             |                                               |                             |
| Female                   | 49414 (28.7 %) | 32854 (30.7 %)                               | 9828 (35.4 %)               |
| Male                     | 122616 (71.2 %) | 74089 (69.3 %)                               | 17961 (64.6 %)              |
| Not recorded             | 294         | 0                                             | 5                           |
| Deprivation quintile     |             |                                               |                             |
| 1 (most deprived)        | 48539 (29.1 %) | 27136 (25.4 %)                               | 4475 (16.1 %)               |
| 2                        | 38254 (22.9 %) | 23419 (21.9 %)                               | 5248 (18.9 %)               |
| 3                        | 31311 (18.28%) | 20714 (19.4 %)                               | 5815 (20.9 %)               |
| 4                        | 26332 (15.8 %) | 17008 (15.9 %)                               | 6084 (21.9 %)               |
| 5 (least deprived)       | 22275 (13.4 %) | 13757 (12.9 %)                               | 5790 (20.8 %)               |
| Not recorded             | 5613       | 4909                                         | 382                         |
| Ethnicity                |             |                                               |                             |
| White                    | 113146 (77.9 %) | 89084 (84.8 %)                               | 21402 (93.6 %)              |
| Asian                    | 6916 (4.8 %)  | 4778 (4.5 %)                                 | 730 (3.2 %)                 |
| Black                    | 4964 (3.4 %)  | 3371 (3.2 %)                                 | 426 (1.9 %)                 |
| Mixed                    | 3695 (2.5 %)  | 435 (0.4 %)                                  | 61 (0.3 %)                  |
| Other ethnic groups      | 16537 (11.4%) | 7325 (7.0 %)                                 | 240 (1.1 %)                 |
| Not recorded             | 27066   | 1950                                         | 4935                        |
| Most common specialties   |             |                                               |                             |
| Paediatrics (23.5 %)     | General medicine (33.4 %) | General medicine (33.1%) | Geriatrics medicine (20.7 %) |
| General surgery (7.3 %)  | General surgery (31.9 %) |                                        |                             |

Where data are not recorded for deprivation, this is due to the lower super output area of residence not being recorded. In most cases this is due to the patient not having a permanent residence in England (typically they are residents of other parts of the UK). Percentages for each recorded category are calculated excluding any unrecorded data.

---

**Figure 1** Proportion of subsequent spells with inconsistencies over time up to three years after the index spell.
condition. The relative importance of each feature is shown in figure 3. Across all three conditions, features strongly associated with data inconsistencies included a change in specialty, a change in provider, shorter spell length of stay and female sex. Data inconsistencies were also associated with older patient age for autism and DMPC and younger patients age for PDD. Although deprivation score was an important predictor for all three conditions, the directionality of the relationship was unclear. For patients with PDD, emergency admissions and the absence of the diagnostic code for Parkinson’s disease were the most important features. AUROC curve values were 0.80 (95% CI 0.80 to 0.81) for autism, 0.76 (95% CI 0.76 to 0.77) for DMPC and 0.75 (95% CI 0.73 to 0.76) for PDD. online supplemental figure S7 reports the areas under the precision-recall curves and precision gain—recall gain curves, also suggesting the classifiers to have good performance. The performance of each model in Black, Asian, male and female patient subgroups is summarised in Online supplemental table S3o and indicates no significant drop in performance for these groups.

DISCUSSION

We used machine learning algorithms to analyse three large datasets to investigate the consistency of clinical coding of three mandatory health conditions within a large administrative healthcare dataset. Clinical coding of DMPC as a mandatory condition was relatively consistent. However, over two-fifths of subsequent spells for autism patients and almost a quarter of subsequent spells for patients with PDD had data inconsistencies. There was a high level of variation in the proportion of data inconsistencies between trusts, and there was no evidence that trusts are consistently poor at reporting mandatory codes across the three conditions studied.

In the HES dataset, inconsistencies related to mandatory clinical codes can arise from two main sources. A failure of the clinician to record the diagnosis in the medical notes or a failure of the clinical coder to code a diagnosis recorded in medical notes. In our analysis, data inconsistencies could also be due to misuse of the code of interest on the first spell (ie, a false positive in the index spell), although the numbers involved are likely to be small.

From the random forest classifier algorithms, age was strongly associated with data inconsistencies. A greater proportion of data inconsistencies were associated with increasing age for autism and DMPC, and with decreasing age for PDD. This confirms the pattern seen in the descriptive data and is likely to be due to expectations around the likelihood that a patient has the condition. This may also explain the relative importance of the association between female sex and more inconsistencies in the autism dataset. Although we identified a relationship between deprivation score and data inconsistencies in all three datasets, the nature of the relationship was unclear. This may suggest a bias towards continuous variables in the algorithms used.19 20

Change in provider, change in main specialty and time from first spell to the subsequent admission were also associated with a higher proportion of data inconsistencies across all datasets. Initiatives to allow easier cross-referencing of information across providers and settings and over an extended period of time should be encouraged.

For the PDD dataset, coding of Parkinson’s disease and emergency admission were associated with lower rates of inconsistencies. Elective admissions are generally of short duration and the case notes are likely to focus on the elective procedure being conducted, with limited coding depth.

Large scale, administrative datasets, such as HES, are being increasingly used to inform decision-making in healthcare.21 22 Such data have helped inform the response to the COVID-19 pandemic23 24 and are being used to inform service structure postpandemic.25–27 Having data which is as reliable as possible will be invaluable. Understanding the source and structure of coding inconsistencies may also help the development of new quality improvement programmes, as well as inform the work of researchers, clinical coders and policy analysts.22 28 The
impact of the data inconsistencies identified in this paper will vary in importance depending on the nature and aims of the data analysis being undertaken. However, we recommend that researchers using HES and interested in long-term comorbidities should not rely on the coding of the index spell alone, but should look at prior spells for the same patient. Frailty/comorbidity indices, such as the Charlson Comorbidity Index and HFRS, if constructed from HES data, perform this function (to an extent) by looking back over 1 and 2 years of prior hospital spells, respectively.

The performance of the algorithms used to identify key features of data inconsistencies was similar in smaller subgroups of ethnicity and sex. There are concerns that artificial intelligence (AI) techniques can accentuate known biases against representation of smaller subpopulations of a dataset. Although the problem of fair data analysis is not unique to AI techniques, and can occur with more traditional forms of data processing and analysis, the ‘black-box’ element of AI methodology leads naturally to concerns over ‘fair AI’ and data equity. We used random forest classifiers in our analysis, allowing us

---

**Table 2** Characteristics of subsequent spells with data inconsistencies

|                        | Autism          | Diabetes mellitus with peripheral complications | Parkinson’s disease dementia |
|------------------------|-----------------|-------------------------------------------------|------------------------------|
| Total no of spells     | 583373          | 651458                                          | 91328                       |
| No of subsequent spells| 390220          | 544341                                          | 60822                       |
| No of subsequent spells with missing mandatory codes | 170447 (43.7 %) | 46679 (8.6 %)                                   | 18975 (31.2 %)              |

**Data inconsistencies by overnight stays**

- Overnight stay: 66251 (38.9 %) 16792 (5.0 %) 11134 (25.2 %)
- Day case: 104196 (47.3 %) 29887 (14.2 %) 7841 (46.8 %)

**Data inconsistencies by method of admission**

- Elective: 84845 (44.4 %) 19393 (5.6 %) 6214 (57.8 %)
- Emergency: 85380 (43.0 %) 27247 (13.9 %) 12752 (25.5 %)
- Not recorded: 222 (42.4 %) 39 (24.7 %) 9 (32.1 %)

**Data inconsistencies by age band**

- 0–17: 57317 (35.2 %) 0 0
- 18–39: 71936 (47.3 %) 594 (9.9 %) 0
- 40–59: 27441 (54.4 %) 9065 (7.3 %) 502 (60.5 %)
- 60–79: 12025 (53.7 %) 25409 (8.6 %) 9514 (34.0 %)
- 80 years and over: 1728 (65.2 %) 11597 (9.7 %) 8955 (28.0 %)
- Not recorded: 0 14 (6.6 %) 7 (23.5 %)

**Data inconsistencies by sex**

- Female: 64650 (46.7 %) 14150 (8.8 %) 6554 (32.9 %)
- Male: 105797 (42.0 %) 32529 (8.5 %) 12421 (30.4 %)
- Not recorded/other: 0 0 0

**Data inconsistencies by deprivation quintile**

- 1 (most deprived): 50739 (45.1 %) 11298 (7.3 %) 3466 (29.2 %)
- 2: 40305 (44.8 %) 10383 (8.1 %) 3622 (30.0 %)
- 3: 30392 (41.5 %) 9493 (8.8 %) 5000 (36.1 %)
- 4: 26390 (43.0 %) 8428 (9.9 %) 3587 (30.0 %)
- 5 (least deprived): 20478 (41.8 %) 6811 (10.1 %) 3213 (29.9 %)
- Not recorded: 2143 (50.1 %) 266 (9.7 %) 87 (30.5 %)

**Data inconsistencies by ethnicity**

- White: 114938 (42.9 %) 39129 (8.5 %) 14301 (31.4 %)
- Asian: 4733 (36.3 %) 1811 (7.0 %) 720 (35.1 %)
- Black: 4195 (41.3 %) 1301 (7.5 %) 295 (25.5 %)
- Mixed: 2362 (36.3 %) 385 (14.8 %) 48 (34.0 %)
- Other ethnic groups: 10155 (41.8 %) 2968 (9.8 %) 183 (32.6 %)
- Not recorded/stated: 34064 (42.9 %) 1086 (12.8 %) 3428 (30.0 %)

Where data are not recorded for deprivation, this is due to the lower super output area of residence not being recorded. In most cases this is due to the patient not having a permanent residence in England (typically they would be residents of other parts of the UK).
to understand the key features represented in our algorithm and allowing a degree of transparency.

Our study has a number of strengths and limitations. We had access to one of the most extensive and complete healthcare datasets anywhere in the world. However, this meant that there was no ‘gold standard’ against which to externally validate the dataset. Difference in coding practice across trusts will have affected our assessment of data quality on the national scale, and we highlight the variability across trusts. We were not able to identify whether an inconsistency was related to a mandatory code being missed in a first spell or being missing in all subsequent spells. We recognise that patients with diabetes mellitus can go into remission, but the number involved across the time period investigated are likely to be very small indeed. We also acknowledge that some forms of dementia and autism may be mild and not impact on the clinical care. Nevertheless, all the conditions studied are mandatory and should still be recorded once diagnosed. The potential variation in the source and proportions of coding inconsistencies across all three conditions, the performance of the three classifiers should not be assessed by one single metric alone. For that reason, we opted to also use the precision-recall curves and the recall-aware precision gain—recall gain curves, particularly relevant for the coding of diabetes where the number of inconsistencies is much lower (ie, higher class imbalance). Our analysis highlights that the characteristics of coding inconsistencies can be particular to the condition under investigation. Although we selected conditions that tend to be present across the lifetime, extrapolation to other disease groups should be done with caution. More broadly, although we investigated inconsistent use of mandatory diagnostic codes in this study, it would be possible to investigate other types of inconsistencies using similar methods.

CONCLUSIONS
We have identified the extent of, and features associated with, data inconsistencies in the HES database for the three conditions studied, with autism having the highest rate of data inconsistencies. With the likely increased use of administrative data to inform healthcare decision-making, data quality will be of central importance if outcomes for patients are to be optimised. As such, improving data quality should be a priority.

Machine learning techniques, as well as providing insight into the characteristics associated with data inconsistencies, may also be of value in identifying potential data inconsistencies during data input, allowing inconsistencies to be corrected prior to finalisation of the data submission.

Acknowledgements We acknowledge NHS Digital for permission to use their data in this report. The GIRFT programme is providing a framework for examining contemporary clinical practice in unprecedented detail and breadth. We also thank all staff within individual NHS trusts who collected and entered the data used in this study.

Contributors This study was designed and organised by FH, JH, WKG, KT, MM, AH and JY. Data cleaning, and analysis was by FH, supported by JH, MM, KT, AH, AW, SE-J and WKG. Writing of the first draft was by FH and WKG. WKG acts as guarantor.

Funding This study was funded by Science and Technology Facilities Council (Grant numbers: Fellowship from DIRAC, ST/S003916/1, ST/W002760/1).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Consent from individuals involved in this study was not required for this analysis of the HES dataset. Data were pseudonymised for analysis. The analysis and presentation of data follows current NHS Digital guidance for the use of HES data for research purposes. Reported data are anonymised to the level required by ISB1523 Anonymisation Standard for Publishing Health and Social Care Data.

Provenance and peer review Not commissioned; externally peer reviewed.
REFERENCES

1. Agrawal R, Prabakaran S. Big data in digital healthcare: lessons learnt and recommendations for general practice. *Heredity* 2020;124:525–34.

2. Griffith GJ, Morris TT, Tuuball MJ, et al. Collier bias undermines our understanding of COVID-19 disease risk and severity. *Nat Commun* 2020;11:1–12.

3. Stulberg JJ, Haut ER. Practical guide to surgical data sets: healthcare cost and utilization project national inpatient sample (NIS). *JAMA Surg* 2018;153:586–7.

4. Benchimol EI, Smeeth L, Guttmann A, et al. Open access and/or omissions arising from translation and adaptation or otherwise. *BMJ* does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

5. Oswald M.

6. Platt J, Grossman D, Goldberg J, et al. Artificial intelligence: the next frontier in health care. *JAMA* 2018;319:240–1.

7. Reddy S, Fox J, Purohit MP. Artificial intelligence-enabled healthcare delivery. *J R Soc Med* 2019;112:22–8.

8. Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. *J Glob Health* 2019;9:e010318.