
Proceedings Paper:
Ruddle, R.A., Hama, L., Wochner, P. et al. (Accepted: 2025) Using a Set-based Approach to Explore Patterns of Lost Autism Diagnoses in Hospital Data. In: Proceedings of the 19th International Joint Conference on Biomedical Engineering Systems and Technologies. 19th International Joint Conference on Biomedical Engineering Systems and Technologies, 02-04 Mar 2026, Marbella, Spain. (In Press)

Using a Set-based Approach to Explore Patterns of Lost Autism Diagnoses in Hospital Data

Roy A. Ruddle[1,2][a], Layik Hama[1][b] and Pamela Wochner[3][c] and Oliver T. Strickson[4][d]
[1]*School of Computer Science, University of Leeds, Leeds, UK*
[2]*Leeds Institute for Data Analytics, University of Leeds, Leeds, UK*
[3]*Department of Computing, Delft University of Technology, Delft, Netherlands*
[4]*Alan Turing Institute, London, UK*
r.a.ruddle@leeds.ac.uk, unlisted, p.wochner@tudelft.nl, ostrickson@turing.ac.uk

Keywords: Visualization, Data Quality, Electronic Health Records, Diagnostic Persistence.

Abstract: Health conditions such as autism do not remit (i.e., some symptoms always persist). If a patient with one of those conditions is treated as an in-patient then the hospital health record for that admission should contain a relevant diagnosis code (i.e., the diagnosis should 'persist'), irrespective of the reason for admission (e.g., cancer). However, for years the UK's National Health Service has been concerned that diagnostic persistence is poor for non-remitting conditions. This paper describes an investigation of diagnostic persistence for autism that used a dataset containing 25,152 hospital episodes from 6383 autism patients who were treated in 224 hospitals. Machine learning models were not accurate for predicting when autism diagnoses would be 'lost', but did indicate that the number of diagnoses, the primary diagnosis and hospital were the most important features. Those features provided a starting-point for a visual analytic investigation that uncovered seven patterns characterizing when autism diagnoses become lost. One pattern applied nationwide (99.6% of autism diagnoses were lost with an R69 primary diagnosis). Half of the lost diagnoses occurred when patients were admitted frequently (every 7 days or less, on average), and that included patients who had repeated treatment for a D61 primary diagnosis. A common theme in other primary diagnosis patterns was that autism diagnoses were always (or almost always) lost in some hospitals, but always or mostly persisted in other hospitals. That provides an opportunity to learn from pockets of good clinical coding practice and significantly improve persistence for autism diagnoses.

1 INTRODUCTION

Electronic health records (EHRs) are digital versions of patients' medical records. One of the main EHR datasets that is collected by National Health Service (NHS) hospitals in England is called Admitted Patient Care (APC) data. APC data is used for service planning, calculating payments to hospitals, providing general management information and secondary research (Herbert et al., 2017). APC data is stored centrally by NHS Digital, which uses a well-established lifecycle to perform hundreds of correction and validation rules on the data that hospitals collect and provide feedback about any data quality issues that occur.

Health conditions such as autism, dementia, diabetes, ischemic heart disease, Parkinson's disease and schizophrenia do not remit. Therefore, if a patient has been diagnosed with one of those diseases then a relevant International Classification of Diseases (ICD) code should always appear ("persist") in subsequent APC data records, even if the patient is being treated for something else (e.g., cancer). However, for years the NHS has been concerned that diagnostic persistence is poor, with an average of 40% of episodes missing those diagnoses for non-remitting conditions (NHS-Digital, 2018a). NHS Digital asked us to investigate this because the reasons are not known.

The volume of data that is involved in each of those conditions is modest by modern standards. Half a million different patients are admitted to English hospitals each year for the most common conditions (diabetes and heart disease) though fewer patients for

[a] https://orcid.org/0000-0001-8662-8103
[b] https://orcid.org/0000-0003-1912-4890
[c] https://orcid.org/0000-0003-4066-8614
[d] https://orcid.org/0000-0002-8177-5582

the other non-remitting conditions. Analysis of that data is challenging because of its complexity, which includes the large number of variables, inevitable differences between the 200+ hospitals that provide inpatient treatment in England and a host of other factors (e.g., see (Hardy et al., 2022). However, that is also the type of complexity for which visual analytics is well-suited.

In this paper we focus on one non-remitting condition (autism). The investigation was performed in Jupyter notebooks, because that matches the type of environment that is typically used by NHS analysts and so would simplify future adoption of our approach. APC data includes both numerical variables (e.g., a patient's age, the deprivation index for where they live and the number of medical conditions they have) and categorical variables (e.g., method of admission, hospital and primary diagnosis). Those categorical variables were hypothesized to be important in diagnostic persistence patterns, which suggested that a good approach could be to treat the data as set-based (e.g., hospitals and diagnoses would each be sets, and combinations of hospitals and diagnoses would be set intersections). Our main contributions are: (1) Identifying seven patterns that characterize the circumstances under which the autism diagnoses had been lost, and (2) Describing the method we used to perform the investigation. In future work we plan to refine the method by applying it to the other non-remitting conditions.

2 RELATED WORK

The present research investigated diagnostic persistence (a data quality issue) in an EHR dataset that records information about patients who were admitted to hospital. This section is divided into parts that briefly review previous research that used APC data, and then previous research into the visualization of EHRs and data quality visualization in general.

2.1 Research Using APC Data

As well as being used for planning, payments and general management information in hospitals, APC data is used as an important data source for research (Herbert et al., 2017). Recent examples include critical care use by dementia patients (Yorganci et al., 2023), acute heart failure outcomes (Cannata et al., 2025), admissions trends for learning disability and autism patients (Zylbersztejn et al., 2023).

There has also been a considerable body of research into aspects of data quality in the health domain (Lewis et al., 2023), including some specifically about APC data. A key finding of one study was the coding differences between hospitals for a particular type of unit (neurosurgery) (Wahba et al., 2023) and another investigated limitations of linking APC and educational datasets (Libuy et al., 2021). Two others have direct relevance to the present research. The first showed how heatmap visualizations of missing values revealed rare and widespread data quality issues that were previously unknown, and ranged from issues that could affect hospital payments to a data preparation error that meant that the researchers could not perform some of the survival analysis that they had planned (Ruddle et al., 2022). The second used random forest models to investigated diagnostic persistence between hospital spells for autism, diabetes and Parkinson's patients (Hardy et al., 2022). The model predicted that diagnoses were more likely to be present if a patient was treated in the same hospital as before and when less time had elapsed since the first spell with a diagnosis.

2.2 EHR Visualization

The 30 year history of research into visualization techniques for EHRs is superbly summarized in a state-of-the-art report (Wang and Laramee, 2022). The 51 main papers that were included in the report's review span applications in a very wide range of medical disciplines, from dementia, diabetes and heart disease (three of the conditions that do not remit), to others such as cancer, intensive care and public health. Also of note is the breadth of visualization techniques that the report identified as being used in EHR visualization. The most widely used techniques were bar, line and pie charts (grouped together under the term "standard 2D displays"). One or more of those and six other common visualization techniques (area chart, boxplot, heatmap, histogram, choropleth map and scatterplot) were used in 36 of the papers, demonstrating their utility.

While the report did not consider data quality as its own application area, data quality was highlighted as one of three major data challenges in healthcare. The report also noted the "significant amount of time and effort" that is spent verifying whether data is suitable for a research.

The life sciences is the most common application domain for set visualization techniques, particularly using implementations of UpSetPlot (Conway et al., 2017; Lex et al., 2014; Nothman, 2022). Examples include visualizing genomic variation (Lex et al., 2014; Conway et al., 2017), biomarkers and genes (Conway et al., 2017; Kalucka et al., 2020; Li et al., 2022), a

smell identification test for screening Parkinson's patients (Landolfi et al., 2022) and the staffing of pediatric palliative care (Rogers et al., 2021). However, the only prior application of set visualization to EHRs that we are aware of was to investigate missing values (Adnan et al., 2019; Ruddle et al., 2022).

2.3 Data Quality Visualization

Three of the main high-level types of data quality issue are accuracy, completeness and consistency (Institute, 2008). Lost diagnoses are a consistency issue, because they arise when one of a patient's APC episodes contains a non-remitting diagnosis that is missing in subsequent episodes. With some parallels to the state-of-the-art report (Wang and Laramee, 2022), data analysts working in 10 industry sectors reported in interviews that they used a variety of visualization techniques to investigate accuracy, completeness and consistency (Ruddle et al., 2023). For consistency, the analysts used bar, line and pie charts to check pairs of variables (e.g., the number of patients in different categories or over time). When 166 Masters students were given a practical assignment in which they had to identify and illustrate data quality issues, they also used a variety of visualization techniques (Ruddle, 2023). To visualize inconsistencies in the data, the students mainly used bar charts, line charts and scatter plots (e.g., to show inconsistencies between a car parking fine, the total paid and the balance remaining). However, twice as often they illustrated consistency issues by showing example data extracts rather than using visualizations.

Diagnostic persistence is a consistency issue, which occurs when some of a patient's records contain a diagnosis for a non-remitting condition (e.g., autism) but other subsequent records do not. In general, it is more difficult to visualize consistency than accuracy or completeness, because consistency usually involves more variables and/or more complex relationships. Our observation from those analyst and student assignment studies (Ruddle et al., 2023; Ruddle, 2023) is that visualizations of data inconsistency were most effective when a single simple visualization was used (e.g., showing new, derived variables) or a story was told with a few such visualizations.

Understanding when and how to simplify data, or being able to create a story (whether that is in an interactive tool or a notebook's workflow), is part of the skill set of good data analysts and visualization experts. Once the variables for each visualization are known then it is straightforward to select a suitable technique because the choice is constrained by well-known rules that map data types to chart types (Mackinley et al., 2007; Wongsuphasawat et al., 2015).

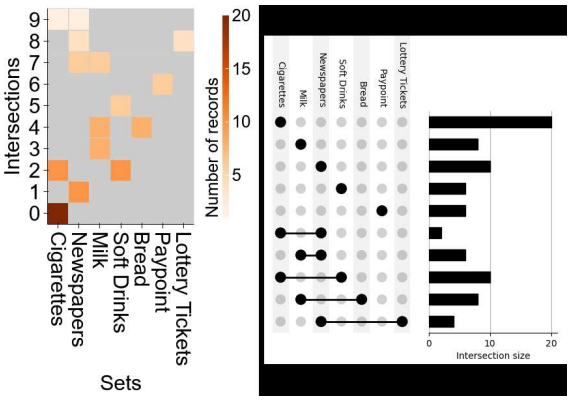

Figure 1: A heatmap (left) and UpSetPlot (right) showing the number of times people bought different combinations of product from a shop as set intersections.

Drawing on data quality exemplars, histograms, box plots and violin plots all have advantages for some types of numerical distribution, bar charts may be used to show any scalar against a discrete reference (e.g., the number of missing values for several variables or value counts for one variable), line and area charts are suitable for the same role if the reference is continuous, and heatmaps show a scalar against a pair of discrete references (e.g., value counts for values of pairs of variables) (Arbesser et al., 2016; Gschwandtner et al., 2014; Kandel et al., 2012; Ruddle and Hall, 2019).

Set-based data adheres to the same rules. The cardinality (i.e., count) of a set is a scalar, so it is typically visualized with a bar chart (Conway et al., 2017; Ruddle et al., 2024). Heatmaps (Ruddle et al., 2022) or UpSetPlots (Conway et al., 2017) are suitable for visualizing set intersections (see Figure 1). Both types of visualization have advantages, with the heatmap showing the make-up and counts of the intersections in a single chart, whereas an UpSetPlot uses two charts to show the same information (the matrix and the bar chart) but the bar lengths show the counts in a way that is perceptually more precise than a heatmap's use of color.

3 CASE STUDY

We had access to an anonymized APC dataset, which contained a whole year's episodes (20 million) from English hospitals. The following sections describe the autism data extract that the case study used, and then the analysis which is subdivided into sections to match the four rows of the flow chart that is shown in Figure 2.

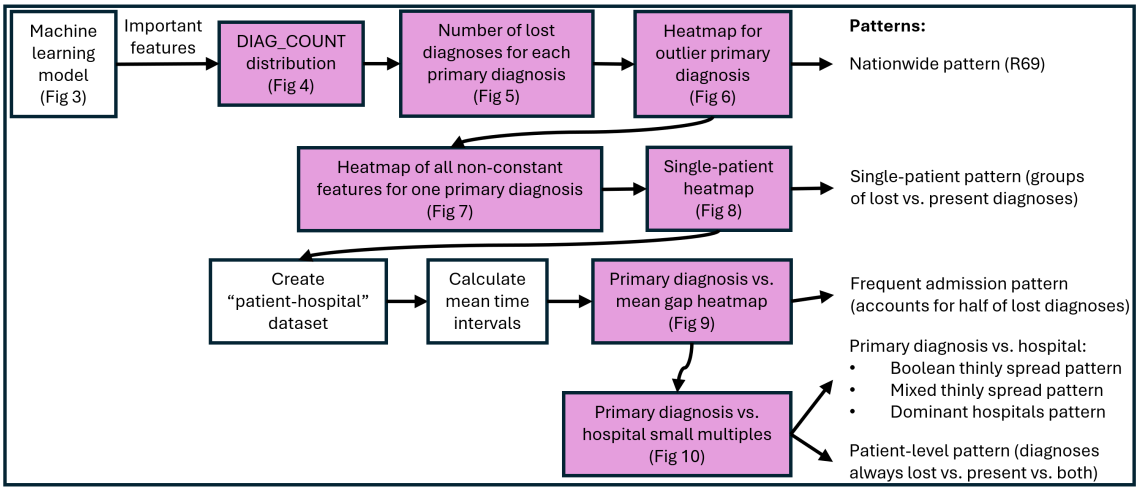

Figure 2: The approach that was used, showing the set-based visualizations (purple boxes) and the steps that led to the discovery of each of the seven lost autism diagnosis patterns.

3.1 Dataset

In APC data there is one record for each episode of patient care, and a new episode starts each time that a patient is admitted to a hospital or moves to the care of another consultant (NHS-Digital, 2018b). APC data is provided as a single data table, which contains hundreds of columns that cover patient demographics, diagnoses and operations, and the organizations that are involved.

APC data includes 20 columns for diagnoses, with each column either containing an ICD code or being empty. The first of those columns is the primary diagnosis, which should always be present, and the other 19 columns are secondary diagnoses (Herbert et al., 2017). The ICD codes that are recorded should reflect the information that a clinician has documented about a patient during a specific admission. Therefore, a key reason for autism diagnoses being missing may be the absence of information about autism in the clinical notes for a given episode.

Episodes for autism patients were extracted as follows. First, we identified the patients who had an autism diagnosis (F840 and F841 ICD codes) in any of their episodes. Then we extracted all of those patients' episodes, added a new variable (PERSIST, with values of "present" or "lost") to indicate whether each episode contained an autism diagnosis, sorted the patients' episodes into date order, and removed any episodes that occurred before the first autism diagnosis. If there was only one episode left for a patient then that was removed, so for all of the remaining patients there were at least two episodes, the first of which always contained an autism diagnosis.

The resulting dataset contained 59 columns (see Appendix). There were 25,152 episodes for 6383 patients, including 6654 "lost" episodes from 2524 patients. There were from 1–19 ICD diagnosis codes in each episode, 893 different primary diagnoses (DIAG_3_01), 218 different hospitals (PROCODE3) and 119 treatment specialties (TRETSPEF).

3.2 Analysis

Prior to the present research we used machine learning to try to investigate diagnostic persistence with the same dataset. Random forest and decision tree models were both very inaccurate. A LightGBM model (Ke et al., 2017) was somewhat better than the others but still missclassified 30% of the lost diagnosis episodes, perhaps due to having to use one-hot encoding to handle categorical variables such as DIAG_3_01 and PROCODE3 that had hundreds of levels. However, the model did indicate that the number of diagnoses, the primary diagnosis and hospital were the most important features (see Figure 3).

The rest of this section describes how we used visual analytics to explore the data and find patterns that characterize when diagnoses do not persist. The computational aspects of our approach involved on-the-fly wrangling of additional data frames, various types of descriptive statistics, slicing & dicing the data and performing set-based calculations (degree, exclusive set intersections, etc.) to provide input for different visualizations. Compared with ordinary data filtering, the advantage of our set-based approach was its capability to reduce dozens of variables into simple-looking heatmaps, and reveal patterns that would otherwise have been hidden in noise. Most of the visualizations were created using SetVis, which is much

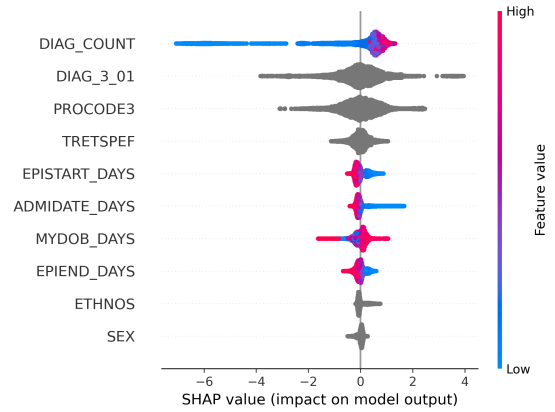

Figure 3: SHAP value (Lundberg et al., 2020) distributions for the 10 most important features in the LightGBM diagnostic persistence model.

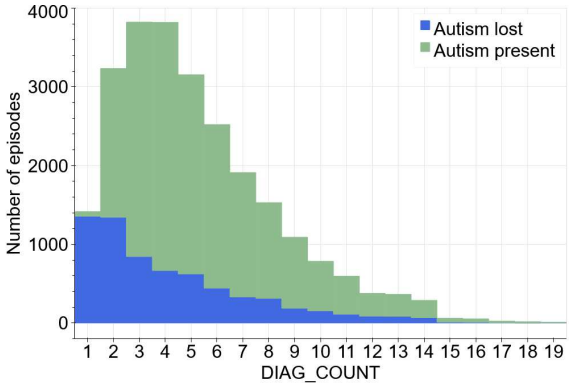

Figure 4: The number of lost and persisting diagnosis episodes for each value of DIAG_COUNT. DIAG_COUNT is the number of ICD diagnosis codes in each episode, where the first is the primary diagnosis (DIAG_3_01) and 2–19 are secondary diagnoses.

more memory-efficient that other set visualization software (Ruddle et al., 2024).

3.2.1 Nationwide Pattern

Consistent with the feature importance insight, SetVis showed that the number of diagnosis codes depended on the value of PERSIST. Autism episodes most often contained 3 or 4 ICD codes. However, episodes that had lost the autism diagnosis often only contained a primary diagnosis by itself or with one other non-autism code (see Figure 4).

That led us to filter the data to investigate PERSIST=lost episodes that only had a primary diagnosis. A SetVis value bar chart showed that R69 (the ICD code for "Unknown and unspecified causes of morbidity") occurred almost 10 times more often than any other primary diagnosis (see Figure 5). The other 246 primary diagnoses occurred in 1–50 episodes.

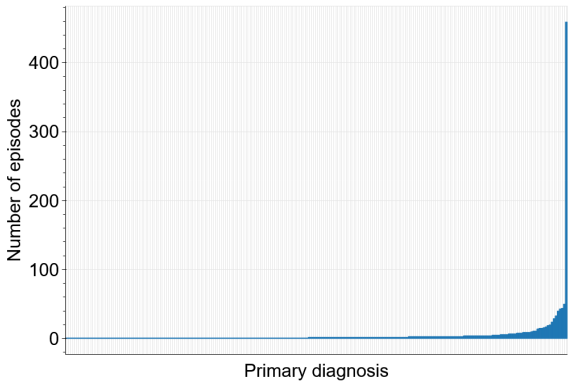

Figure 5: The number of episodes that occurred for each primary diagnosis (DIAG_3_01), in records that did not contain any secondary diagnoses. For confidentiality, the primary diagnoses are not labeled on the X axis. However, the spike at the right-hand side is for the R69 code.

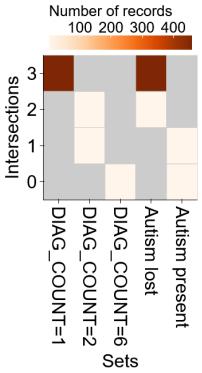

Figure 6: A heatmap for R69 primary diagnoses, showing the number of episodes that occurred for each combination of DIAG_COUNT and autism lost vs. present.

Clearly R69 was unusual. After selecting only the episodes for that primary diagnosis, a SetVis heatmap showed that all but two of the 462 episodes were PERSIST=lost and only three episodes contained secondary diagnoses (see Figure 6). Additional calculations showed that the R69 episodes took place in 68 different hospitals (the PROCODE3 variable). In other words, a nationwide pattern was that hospitals do not record autism for an R69 primary diagnosis. A domain expert with many years of NHS clinical coding experience was surprised that there were any R69 codes, particularly in the absence of any secondary diagnoses, because that indicated that an episode was effectively uncoded.

3.2.2 Groups of Lost and Present Diagnoses

Value count calculations showed that PERSIST=lost did not dominate for any other primary diagnoses. However, D61 ("Other aplastic anemias and other bone marrow failure syndromes") stood out as being

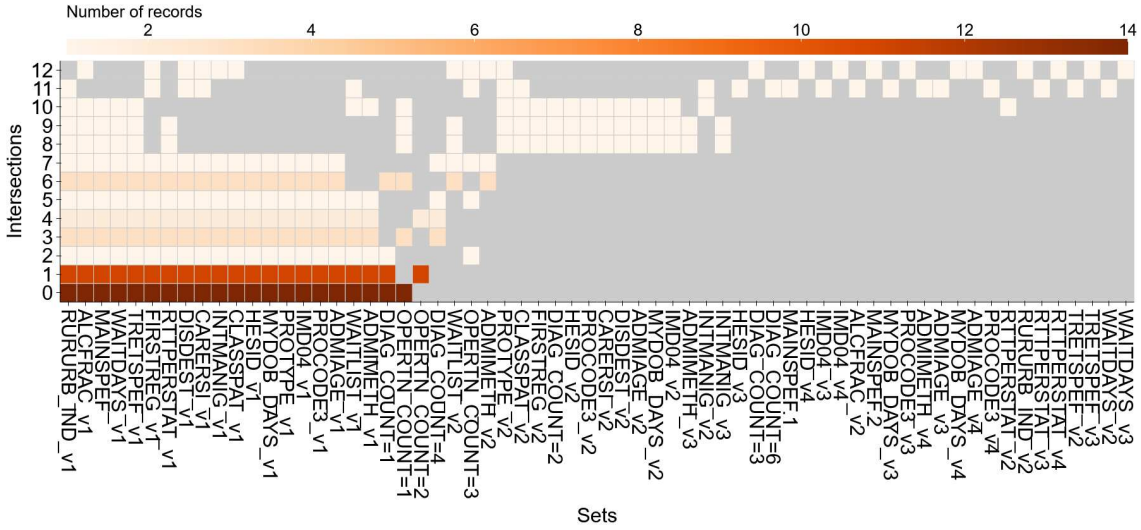

Figure 7: A heatmap for D61 primary diagnoses that had lost an autism diagnosis. The block of shaded cells (intersections 0–7) that are bottom left in the heatmap are a single patient.

balanced (41 out of 99 episodes had lost the autism diagnosis).

We selected all 41 of the lost episodes with a D61 primary diagnosis. Then, lacking other ideas, we cast the net wide and visualized intersections for all 21 of the variables that contained more than one unique value for those episodes (NB: admission and episode dates and duration were omitted because they were certain to change from between episodes). A sorted SetVis heat map revealed a group of intersections (see Figure 7) that always contained the patient with the most D61 episodes and 16 other sets (i.e., specific values of 16 variables, including the hospital). That homogeneity was unexpected, but was a breakthrough insight because it indicated that groups of PERSIST=lost episodes occur for specific patients who were admitted multiple times into a single hospital.

After filtering the data to only include that patient and the hospital where they were primarily treated, a heatmap for PERSIST and the episode start date showed that the 78 episodes occurred in 12 blocks that unexpectedly alternated between an autism code being present or lost (see Figure 8). Therefore, we hypothesized that autism diagnoses are often lost when a patient underwent repeated treatment in a specific hospital.

3.2.3 Frequent Admissions

To investigate the repeated treatments, we transformed the autism extract into a new "patient-hospital" dataset. First, we filtered out episodes with a R69 primary diagnosis, and any patients that then only had one episode or did not have any PERSIST=lost episodes. Next, we grouped episodes by patient/hospital, sorted the episodes into order and calculated the mean interval between the episodes. A total of 2245 PERSIST=lost episodes occurred in groups with a mean gap of up to 1 day, 1222 occurred in groups with a 2–7 day mean gap, 1099 occurred in groups with a 8–30 day mean gap, and 934 occurred in groups with a mean gap that exceeded 30 days. In other words, the majority of lost autism diagnoses occurred when patients were admitted for multiple episodes to the same hospital at short time intervals (up to 7 days apart). That is the direct opposite of findings from researchers who used a random forest model to conclude that autism and other diagnoses were lost more often when patients were subsequently treated at a different hospital and significant time had elapsed between admissions (Hardy et al., 2022).

For each of the 617 unique primary diagnoses in the patient-hospital dataset, we calculated the number of PERSIST=lost episodes and the percentage that occurred within each mean gap band ($\leq$ 1 day, 2–7 days, 8–30 days and $>$ 30 days). There were no PERSIST=lost episodes for 52 of the primary diagnoses, and another 457 only had a few (1–9) lost episodes each. However, the other 108 primary diagnoses had 10 or more PERSIST=lost episodes (see Figure 9) and the patients with those primary diagnoses were treated in 207 hospitals. One pattern that immediately popped out was that 100% of the lost autism diagnoses occurred in a single band for some primary diagnoses. For the $\leq 1 day$ band, those were the ICD codes C64 (malignant neoplasm of kid-

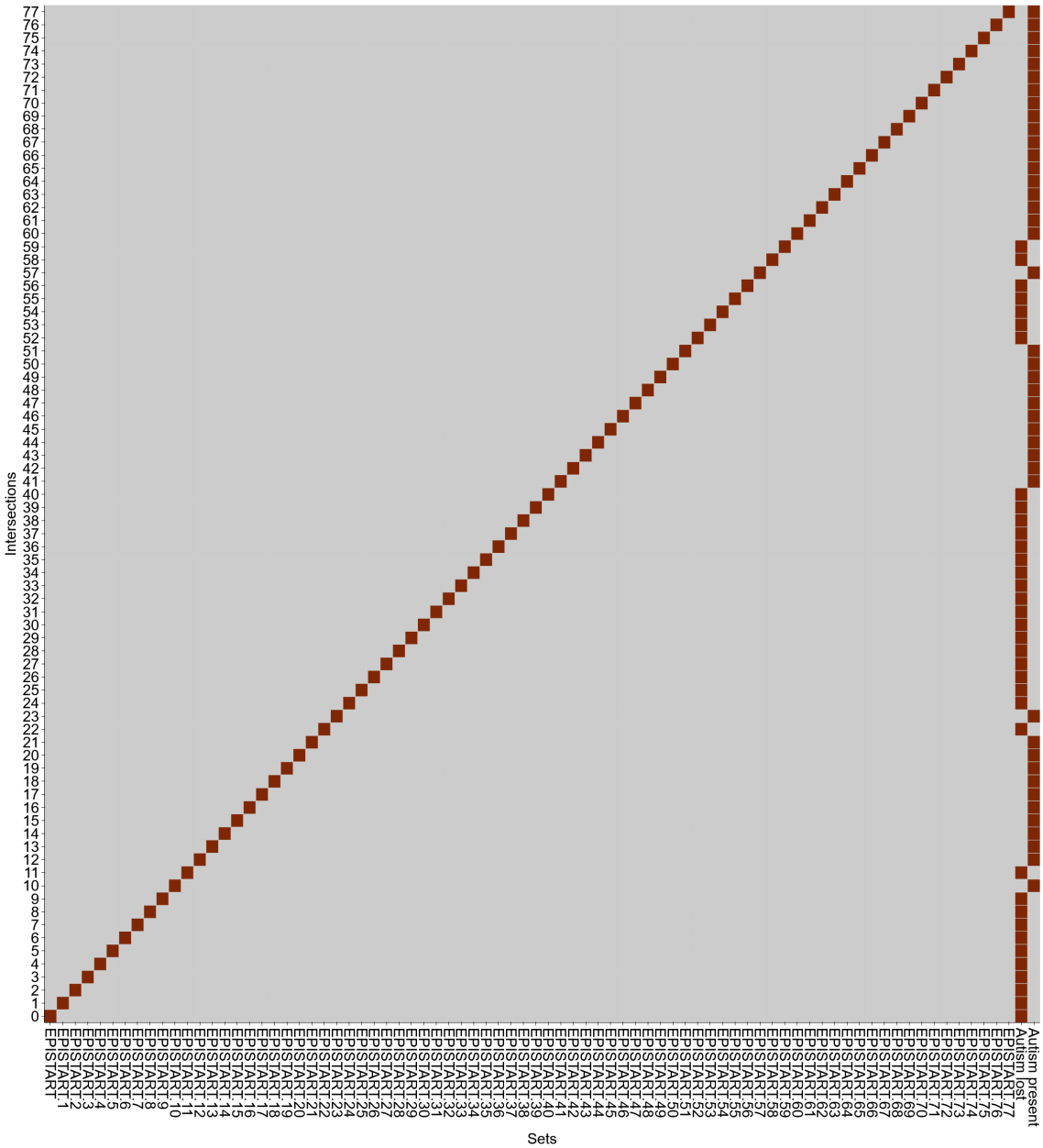

Figure 8: A heatmap showing 12 alternating blocks of autism lost vs. present episodes (the two right-hand columns) for one patient in one hospital and involving a single primary diagnosis (D61). The diagonal line of shaded cells are the patient's episodes in date order.

ney, except renal pelvis) and Q78 (other osteochondrodysplasias). For the the $\leq 7 days$ band it was the ICD codes C81 (Hodgkin lymphoma), D56 (thalassaemia), D83 (common variable immunodeficiency) and K58 (irritable bowel syndrome).

The clinical coding domain expert commented that it is often challenging to capture comorbidity information for patients who are regular attenders, and success would depend on the process that a provider has in place for such admissions. Sometimes a patient's codes might be copied from the previous admission, so lost diagnoses could be propagated into the block pattern that we found. Also, on some occasions a regular attendee may be treated by a nurse rather than a consultant, so the patient's full medical history may not be recorded.

Gap <= 1 day for at least 75% of lost diagnoses

DIAG_3_01	Number of lost diagnoses	% episodes with given mean gap			
		<= 1 day	<= 7 days	<= 30 days	> 30 days
C64		100			
Q78		100			
D61		97			2
D57		92	3	3	
N18		90	8	0	0
C71		88	8		2
C49		84		15	
D70		81	9	9	
D64		81	4	4	9
T86		80	8	2	8
K31		76			23
Z75		75	19	5	

Gap <= 7 days for at least 75% of lost diagnoses

DIAG_3_01	Number of lost diagnoses	% episodes with given mean gap			
		<= 1 day	<= 7 days	<= 30 days	> 30 days
C81			100		
D56			100		
D83			100		
K58			100		
T18		39	51	7	1
D59		63	27	9	
Z34		45	45	9	
I47			90		9
O36		50	40	10	
T83		26	63	5	5
C85		66	20	12	
T42		50	30	11	7
F60		63	15	12	9
C91		59	19	14	6
K50		26	50	21	1
I50		46	30	23	
L97		23	53	7	15
R07		52	23	16	7

Gap <= 30 days for at least 75% of lost diagnoses

DIAG_3_01	Number of lost diagnoses	% episodes with given mean gap			
		<= 1 day	<= 7 days	<= 30 days	> 30 days
C83		27	36	36	
J69		10	48	41	
M79		46	7	46	
O26		42	21	35	
S51		21	17	60	
A41		56	10	30	3
F10		36	32	28	4
M41		63		31	4
K91		28	33	33	4
Z46		46	13	33	6
T81		27	46	18	6
T82		35	21	35	7
S82			16	75	8
R50		17	56	17	8
M25		26	21	43	8
K22		56	12	21	9
F32		60		30	10
N20		30	10	50	10
R29		30	30	30	10
T50		30	30	30	10
G93		36	5	47	10
Z45		33	33	22	11
E10		30	33	23	12
R10		30	31	25	12
T39		42	17	27	13
T40		40	23	23	13
N17		47	23	14	14
T38		33	28	23	14
J45		31	39	15	15
M86		7	61	15	15
J18		43	23	16	15
M08			44	38	16
Z03		34	21	26	17
K85		11	41	29	17
I21		45	18	18	18
K62		9	27	45	18
K59		25	30	24	18
A09		31	34	15	18
F84		62	6	12	18
G40		28	21	30	19
K63		40	10	30	20
K92		14	8	57	20
R55		35	29	14	20
K51		28		50	21
N39		30	35	11	22
T43		38	12	25	22
R51		36	12	28	24

Other primary diagnoses

DIAG_3_01	Number of lost diagnoses	% episodes with given mean gap			
		<= 1 day	<= 7 days	<= 30 days	> 30 days
A08		5	55	11	27
B34		16	11	23	47
E16		70			30
E84		11	11	35	41
F20		18	27	18	36
G41		23	13	36	26
J03		24	10	29	35
J06		18	20	25	34
J22		32	23	13	30
J35		23	7	30	38
K02		8	6	26	58
K29		22	22	27	27
K56		6	42	24	27
K80		13	27	27	31
L02		14	14	28	42
L03		36	18	9	36
M54		20	33	20	26
R06			16	50	33
R11		28	21	12	37
R31		16	25	8	50
R33		28	23	4	42
R44		6	13	53	26
R45		25	18	30	25
R56		36	12	25	25
R63		11	23	34	30
S09		30		38	30
S61		27	36		36
T85		44	20	6	27
Z09		18	43	6	31
Z43		46	7	15	30
Z47		3	6	20	70

Figure 9: The number of PERSIST=lost episodes for each primary diagnosis (DIAG_3_01), and a heatmap showing the percentage for four bands of mean gap between all episodes for each patient/hospital combination. Note: The heatmap only shows the 108 primary diagnoses that had 10 or more lost autism diagnoses.

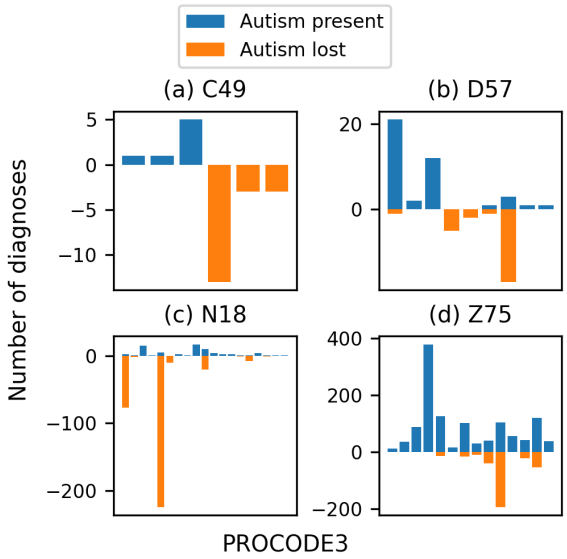

Figure 10: Each plot shows one primary diagnosis, with the bars showing the number of diagnoses that were lost and/or persist for each hospital (PROCODE3). The plots illustrate: (a) A boolean thinly spread pattern for the C49 (malignant neoplasm of other connective and soft tissue) primary diagnosis, (b) A mixed thinly spread pattern for D57 (sickle-cell disorders), (c) A dominant hospitals pattern for N18 (chronic kidney disease), and (d) A patient-level pattern for Z75 (problems related to medical facilities and other health care, e.g., holiday relief care).

3.2.4 Patterns at a Diagnosis/Hospital and Patient Level

Finding patterns in data that contain hundreds of primary diagnoses and hospitals is a daunting task. However, we found small multiple visualizations (see Figure 10) useful to identify where to look for the proverbial "needle in a haystack" insights that revealed four types of pattern in which diagnoses are lost.

A *boolean thinly spread pattern* (see Figure 10a) occurred where: (i) there were only a few episodes with a given primary diagnosis at each hospital, and (ii) in those episodes the autism diagnosis was either always present or always lost (never a mixture of the two). We hypothesize that there are different procedures for recording patient data in hospitals where the autism diagnosis was present vs. lost.

A *mixed thinly spread pattern* occurred where there were only a few episodes at each hospital (as above), but the hospitals with the most episodes had a mixture of present and lost autism diagnoses (see Figure 10b). We hypothesize that the hospitals with the greatest number of lost diagnoses could learn internally from pockets of good coding practice that already exist.

A *dominant hospitals pattern* occurred where one or two hospitals had lost diagnoses for a very large number of episodes (see Figure 10c). Those are the hospitals where the greatest improvements in diagnostic persistence data quality could be made, using NHS England's existing data quality lifecycle (Team, 2016).

At first glance, the final pattern (see Figure 10d) looks similar to the dominant hospitals pattern, with a few hospitals having large numbers of lost and present autism diagnoses. However, further investigation using SetVis heatmaps that were similar to those used for the D61 primary diagnosis (see Figure 7) provided a different explanation, because they revealed that the hospitals could be divided into groups where: (i) each patient's autism diagnoses were either always present or all lost, (ii) every patient had PERSIST=present and PERSIST=lost episodes, or (iii) a mixture of (i) and (ii). We term this the *patient-level pattern*.

4 CONCLUSIONS

In this visual analytic case study, we investigated the circumstances under which diagnoses were lost from autistic patients (a condition that does not remit) in a 25,152 record EHR dataset for 6383 autism patients. The dataset contained 59 variables, some of which were categorical with a large number of levels. Those included two of the three most important features that a LightGBM model (Ke et al., 2017) identified – the primary diagnosis (893 levels) and the hospital for each treatment episode (218 levels). One reason for the model's limited overall accuracy was probably the sparse distribution of the lost autism diagnoses across those features. However, for such data, a visual analytic approach did prove effective to explore the patterns of lost autism diagnoses.

The computational aspects of the analysis involved deriving additional data frames, value counts and other descriptive statistics, various subsets of the data and performing set-based calculations to provide input for different visualizations. Through the computations and visualizations we identified seven patterns that characterize the circumstances under which the autism diagnoses had been lost. One pattern applied to every hospital in the country that had admitted autism patients with a particular primary diagnosis. Other patterns were based on what appeared to be different clinical coding practices within and between hospitals, and the remaining patterns were based on differences at the patient level.

Finally, future work is planned to apply our method to five other non-remitting conditions: de-

mentia, diabetes, ischemic heart disease, Parkinson's disease and schizophrenia.

ACKNOWLEDGEMENTS

This research was supported by the Engineering and Physical Sciences Research Council (EP/N013980/1; EP/K503836/1; EP/X029689/1) and the Alan Turing Institute. The authors thank M. Harwood-Jones for discussions about clinical coding. The authors acknowledge the use of the LASER platform and assistance provided by the Data Analytics Team in Leeds Institute for Data Analytics (University of Leeds)

Data availability statement: The case study used a pseudonymized dataset from NHS Digital (NIC-49164-R3G5K) that, due to data governance restrictions, cannot be made openly available.

REFERENCES

Adnan, M., Nguyen, P. H., Ruddle, R. A., and Turkay, C. (2019). Visual analytics of event data using multiple mining methods. In *EuroVis Workshop on Visual Analytics (EuroVA) 2019*, pages 61–65. The Eurographics Association.

Arbesser, C., Spechtenhauser, F., Mühlbacher, T., and Piringer, H. (2016). Visplause: Visual data quality assessment of many time series using plausibility checks. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):641–650.

Cannata, A., Mizani, M. A., Bromage, D. I., Piper, S. E., Hardman, S. M., Sudlow, C., de Belder, M., Scott, P. A., Deanfield, J., Gardner, R. S., et al. (2025). Heart failure specialist care and long-term outcomes for patients admitted with acute heart failure. *Heart Failure*, 13(3):402–413.

Conway, J. R., Lex, A., and Gehlenborg, N. (2017). UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18):2938–2940.

Gschwandtner, T., Aigner, W., Miksch, S., Gärtner, J., Kriglstein, S., Pohl, M., and Suchy, N. (2014). Timecleanser: A visual analytics approach for data cleansing of time-oriented data. In *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business*, pages 1–8.

Hardy, F., Heyl, J., Tucker, K., Hopper, A., Marchã, M. J., Briggs, T. W., Yates, J., Day, J., Wheeler, A., Eve-Jones, S., et al. (2022). Data consistency in the English hospital episodes statistics database. *BMJ Health & Care Informatics*, 29(1):e100633.

Herbert, A., Wijlaars, L., Zylbersztejn, A., Cromwell, D., and Hardelid, P. (2017). Data resource profile: Hospital episode statistics admitted patient care (HES APC). *International Journal of Epidemiology*, 46(4):1093–1093i.

Institute, B. S. (2008). Software engineering - software product quality requirements and evaluation (SQuaRE) - data quality model.

Kalucka, J., de Rooij, L. P., Goveia, J., Rohlenova, K., Dumas, S. J., Meta, E., Conchinha, N. V., Taverna, F., Teuwen, L.-A., Veys, K., et al. (2020). Single-cell transcriptome atlas of murine endothelial cells. *Cell*, 180(4):764–779.

Kandel, S., Parikh, R., Paepcke, A., Hellerstein, J. M., and Heer, J. (2012). Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 547–554.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.

Landolfi, A., Picillo, M., Pellecchia, M. T., Troisi, J., Amboni, M., Barone, P., and Erro, R. (2022). Screening performances of an 8-item UPSIT Italian version in the diagnosis of Parkinson's disease. *Neurological Sciences*, pages 1–7.

Lewis, A. E., Weiskopf, N., Abrams, Z. B., Foraker, R., Lai, A. M., Payne, P. R., and Gupta, A. (2023). Electronic health record data quality assessment and tools: A systematic review. *Journal of the American Medical Informatics Association*, 30(10):1730–1740.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992.

Li, Z., Feng, J., Zhong, J., Lu, M., Gao, X., and Zhang, Y. (2022). Identification of FN1 as a biological marker for diabetic nephropathy by bioinformatics analysis. *Frontiers in Endocrinology*, page 1450.

Libuy, N., Harron, K., Gilbert, R., Caulton, R., Cameron, E., and Blackburn, R. (2021). Linking education and hospital data in England: Linkage process and quality. *International Journal of Population Data Science*, 6(1):1671.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):56–67.

Mackinley, J., Hanrahan, P., and Stolte, C. (2007). Show me: Automatic presentation for visual analysis. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1137–1144.

NHS-Digital (2018a). Data quality report on comorbidity diagnostic persistence. [Online]. Available: https://nhs-prod.global.ssl.fastly.net/binaries/content/assets/website-assets/data-and-information/data-quality/february_diagnostic_persistence_infographic_final_v0.1.pdf.

NHS-Digital (2018b). HES data dictionary - admitted patient care.

Nothman, J. (2022). Upsetplot.

Rogers, M. M., Friebert, S., Williams, C. S., Humphrey, L., Thienprayoon, R., and Klick, J. C. (2021). Pediatric

palliative care programs in US hospitals. *Pediatrics*, 148(1).

Ruddle, R. and Hall, M. (2019). Using miniature visualizations of descriptive statistics to investigate the quality of electronic health records. In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies-Volume 5: HEALTHINF*, pages 230–238. SciTePress.

Ruddle, R., Hama, L., Wochner, P., and Strickson, O. (2024). Setvis: Visualizing large numbers of sets and intersections. *Journal of Open Source Software*, 9(103):6925.

Ruddle, R. A. (2023). Using well-known techniques to visualize characteristics of data quality. In *VISIGRAPP (3: IVAPP)*, pages 89–100.

Ruddle, R. A., Adnan, M., and Hall, M. (2022). Using set visualisation to find and explain patterns of missing values: a case study with NHS hospital episode statistics data. *BMJ Open*, 12(11):e064887.

Ruddle, R. A., Cheshire, J., and Fernstad, S. J. F. (2023). Tasks and visualizations used for data profiling: A survey and interview study. *IEEE Transactions on Visualization and Computer Graphics*, 30(7):3400–3412.

Team, H. D. Q. (2016). The HES processing cycle and data quality, version 4. [Online]. Available: https://bit.ly/2HWgMeR.

Wahba, A. J., Cromwell, D. A., Hutchinson, P. J., Mathew, R. K., and Phillips, N. (2023). Assessing national patterns and outcomes of pituitary surgery: Is hospital administrative data good enough? *British Journal of Neurosurgery*, 37(5):1135–1142.

Wang, Q. and Laramee, R. S. (2022). Ehr star: The state-of-the-art in interactive ehr visualization. In *Computer Graphics Forum*, volume 41, pages 69–105. Wiley Online Library.

Wongsuphasawat, K., Moritz, D., Anand, A., Mackinlay, J., Howe, B., and Heer, J. (2015). Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):649–658.

Yorganci, E., Sleeman, K. E., Sampson, E. L., and Stewart, R. (2023). Survival and critical care use among people with dementia in a large English cohort. *Age and Ageing*, 52(9):afad157.

Zylbersztejn, A., Stilwell, P. A., Zhu, H., Ainsworth, V., Allister, J., Horridge, K., Stephenson, T., Wijlaars, L., Gilbert, R., Heys, M., et al. (2023). Trends in hospital admissions during transition from paediatric to adult services for young people with learning disabilities or autism: population-based cohort study. *The Lancet Regional Health–Europe*, 24.

APPENDIX

This appendix lists the 59 columns that were in the data extract. The columns names are in CAPTIALS. The descriptions follow and are from the HES Data Dictionary – Admitted Patient Care (2 August 2017), which is no longer provided on the NHS Digital website (NHS Digital merged with NHS England on 1 February 2023).

ACSCFLAG: Ambulatory Care Sensitive Condition flag
ADMIAGE: Age on admission
ADMIDATE: Date of admission
ADMIMETH: Method of admission
ADMINCAT: Administrative category
ADMISORC: Source of admission
ADMISTAT: Psychiatric history on admission
AEKEY: Whether there is a corresponding A&E attendance
ALCFRAC: Principal alcohol related fraction
BEDYEAR: Bed days within the year
CARERSI: Carer support indicator
CLASSPAT: Patient classification
DIAG_3_01: The first three characters of the primary diagnosis code
DIAG_COUNT: Count of diagnoses
DISDEST: Destination on discharge
DISMETH: Method of discharge
ELECDUR: Duration of elective wait
EPIDUR: Episode duration
EPIEND: Date episode ended
EPIKEY: Record identifier
EPIORDER: Episode order (within an admission spell)
EPISODE_SEQUENCE: Overall order of episode for patient (derived from ADMIDATE and EPIORDER)
EPISTART: Date episode started
EPISTAT: Episode status
EPITYPE: Episode type
ETHNOS: Ethnic category
ETHRAW: Ethnic character
FAE: Finished Admission Episode
FAE_EMERGENCY: Finished Admission Episode, emergency classification
FCE: Finished Consultant Episode
FDE: Finished In-Year Discharge Episode
FIRSTREG: First regular day or night admission
HESID: Encrypted patient identifier
IMD04: Index of Multiple Deprivation
INTMANIG: Intended management
MAINSPEF: Main specialty
MARSTAT: Marital status
MYDOB: Date of Birth - month and year
NEOCARE: Neonatal level of care
NEWNHSNO_CHECK: NHS Number valid flag
NHSNOIND: NHS number status indicator
OPERSTAT: Operation status code
OPERTN_COUNT: Count of procedures
PCFOUND: Postcode Found

PERSIST: Diagnosis of non-remitting condition (present or lost)
POSOPDUR: Post-operative duration
PREOPDUR: Pre-operative duration
PROCODE3: Provider Code (hospital)
PROTYPE: Provider type
RTTPERSTAT: Referral to Treatment period status
RURURB_IND: Rural/Urban Indicator
SEX: Sex of patient
SPELBGIN: Beginning of spell indicator
SPELDUR: Duration of spell
SPELEND: End of spell indicator
TRETSPEF: Treatment specialty
WAITDAYS: Duration of wait (referral to treatment period)
WAITLIST: Method of Admission - Waiting List
WELL_BABY_IND: Well baby flag