BMJ Health & Care Informatics

Communicating exploratory unsupervised machine learning analysis in age clustering for paediatric disease

Joshua William Spear ,[1,2] Eleni Pissaridou,[1,2] Stuart Bowyer,[1,2] William A Bryant,[1,2] Daniel Key ,[1,2] John Booth ,[1,2] Anastasia Spiridou,[1,2] Spiros Denaxas,[3,4] Rebecca Pope,[5] Andrew M Taylor,[1,6] Harry Hemingway,[1,3] Neil J Sebire [1,5]

To cite: Spear JW, Pissaridou E, Bowyer S, *et al*. Communicating exploratory unsupervised machine learning analysis in age clustering for paediatric disease. *BMJ Health Care Inform* 2024;**31**:e100963. doi:10.1136/bmjhci-2023-100963

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjhci-2023-100963).

Received 10 November 2023
Accepted 01 July 2024

Check for updates

© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

[1]DRIVE, Great Ormond Street Hospital for Children, London, UK
[2]NIHR GOSH BRC, London, UK
[3]Institute of Health Informatics, University College London, London, UK
[4]BHF Data Science Centre, London, UK
[5]Institute of Child Health, University College London, London, UK
[6]Institute of Cardiovascular Science, University College London, London, UK

Correspondence to
Professor Neil J Sebire; neil.sebire@gosh.nhs.uk

ABSTRACT

Background Despite the increasing availability of electronic healthcare record (EHR) data and wide availability of plug-and-play machine learning (ML) Application Programming Interfaces, the adoption of data-driven decision-making within routine hospital workflows thus far, has remained limited. Through the lens of deriving clusters of diagnoses by age, this study investigated the type of ML analysis that can be performed using EHR data and how results could be communicated to lay stakeholders.

Methods Observational EHR data from a tertiary paediatric hospital, containing 61 522 unique patients and 3315 unique ICD-10 diagnosis codes was used, after preprocessing. K-means clustering was applied to identify age distributions of patient diagnoses. The final model was selected using quantitative metrics and expert assessment of the clinical validity of the clusters. Additionally, uncertainty over preprocessing decisions was analysed.

Findings Four age clusters of diseases were identified, broadly aligning to ages between: 0 and 1; 1 and 5; 5 and 13; 13 and 18. Diagnoses, within the clusters, aligned to existing knowledge regarding the propensity of presentation at different ages, and sequential clusters presented known disease progressions. The results validated similar methodologies within the literature. The impact of uncertainty induced by preprocessing decisions was large at the individual diagnoses but not at a population level. Strategies for mitigating, or communicating, this uncertainty were successfully demonstrated.

Conclusion Unsupervised ML applied to EHR data identifies clinically relevant age distributions of diagnoses which can augment existing decision making. However, biases within healthcare datasets dramatically impact results if not appropriately mitigated or communicated.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Misuse of machine learning (ML) models, Bayesian analysis of uncertainty and the application of unsupervised ML models to healthcare domains are all widely studied areas of research. However, these had not been consolidated in a digestible format, aimed at data scientists not au fait with the details of each area. With the growing availability of "plug and play" computer software for performing ML analysis, the requirement for practical workflows providing a starting point for deeper technical analysis is growing.

WHAT THIS STUDY ADDS

⇒ This study has begun to address the aforementioned gap in existing literature by providing an exemplary piece of unsupervised learning analysis on healthcare data, demonstrating the potential benefit from such analysis as well as common pitfalls and how these should be mitigated. Unsupervised analysis was chosen since asking 'what can I learn from data' is a common starting place for practitioners interested in utilising ML methods. This study equips readers with an unsupervised ML workflow for performing analysis and considering uncertainties through a Bayesian viewpoint. Whilst simple, this workflow can be developed to incorporate more complex statistical analysis as the maturity of the practitioner grows.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Practitioners reading this study should be equipped and motivated to perform ML analysis which more robustly accounts for uncertainties and thus prevents decision making and policies from being miss-guided.

INTRODUCTION

The increasing availability of electronic healthcare data has presented an array of opportunities to improve healthcare services by enabling data-driven decision-making.[1–3] Despite this, and the discourse surrounding artificial intelligence (AI) and machine learning (ML), the adoption of data-driven tools within routine hospital workflows has remained limited. This study addressed two fundamental questions for using data-driven insights, derived from electronic healthcare record (EHR) data, using an example analysis. The first question concerns whether ML

BMJ Health & Care Informatics: first published as 10.1136/bmjhci-2023-100963 on 29 July 2024. Downloaded from https://informatics.bmj.com on 10 September 2024 by guest. Protected by copyright.

analysis is possible with EHR data and if so what kind. The second question concerns how the results of such ML analysis should be effectively communicated.

The use of EHR data to derive insight into healthcare processes is not new. Recently, a study used outpatient data along with k-means clustering to identify the most frequent diagnoses in a children's hospital in China.[4] However, numerous factors, including hospital processes and external socioeconomic influences can result in significant diversity across healthcare populations. As such, the extent to which any part of an ML pipeline or results are transferable remains uncertain. Furthermore, ML-based analysis has become increasingly accessible, with the introduction of easy-to-use APIs[5] and a technology stack which has been abstracted away from companies and analysts.[6] Additionally, the use of robust 'simple' algorithms, for example, k-means clustering, has meant that producing reasonable findings based on ML analysis has been readily achievable. However, often the assumptions underpinning such analyses are not communicated and the results may therefore be over interpreted.

To address the first query, the problem of defining disease age clusters as a real-world use case, was considered. It is well-established that many diseases, while affecting a wide age range, have characteristic age distributions. Furthermore, several such distributions have already been described for specific conditions such as uveitis, asthma and thyroid disorders.[7–9] Providing quantitative insights into these distributions would directly enable operational and policy level, data-driven decisions to be made, such as through accurate resource forecasting and planning. Furthermore, such analyses would enable clinicians to better understand the likelihood of presenting with different conditions at various ages and, for example, enable clinicians to understand the age atypicality of a given patient. Despite the clear advantages of understanding the age distributions and clusters across different diseases, the focus of the present study is predominantly on the modelling methodology within the context of healthcare, rather than any novelty of clinical insights. To address the second question, explicit focus was placed on the effect of various preprocessing decisions and how these could be effectively understood and communicated, for example, by examining the preprocessing of ICD-10 diagnosis codes. ICD-10 diagnosis codes assigned as part of a hospital's billing process inform a large majority of statistical analysis and have been the subject of interest for recent research into applying ML to healthcare.[10] Despite this, the inherent bias within the code assignments has seldom been addressed.

METHODS

Data sources

Great Ormond Street Hospital for Children (GOSH) is a tertiary children's hospital in London, specialising in children with rare and/or complex conditions over a wide range of clinical specialities including oncology, paediatric transplantation, immunology and complex paediatric surgery but covering all paediatric specialities. The hospital has a dedicated digital research environment (DRE) which integrates harmonised routine data from legacy clinical systems as well as data derived from a comprehensive electronic health record system (Epic)[11] since 2019. Deidentified routinely collected data within the DRE were made available for this study through standardised processes within the Trust. Once appropriate research ethics committee and other approvals were in place, non-identifiable datasets were analysed within a secure audited workspace inside the GOSH-Aridhia DRE; all data remained in place, analysis was performed in situ and only aggregate results and charts were exported.[12]

Data processing

Available ICD-10 diagnoses were extracted from Epic for all patients (inpatients and outpatients) seen at GOSH between 1 October 2019 and 1 October 2021, using diagnoses provided through routine clinical coding. The full ICD code assigned to the patient was retained, resulting in the processed dataset containing codes with a mixture of lengths 3, 4 and 5 digits. ICD-10 codes were used due to the standardised coding process that is followed (assumption 1) and, since main diagnoses per admission were investigated, ICD-10 codes were considered when defined as the primary diagnosis by the clinical coding process (assumption 2).

Data exclusions

As part of standard disclosure control procedure, extremely rare diagnoses (those with fewer than five patients with the condition in the dataset) were excluded from further analysis as well as those with invalid ICD-10 codes (assumption 3). Prenatal diagnosis referrals were excluded (assumption 4). ICD-10 codes from chapters 'XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified', 'XIX Injury, poisoning and certain other consequences of external causes', 'XX External causes of morbidity and mortality', 'XXI Factors influencing health status and contact with health services' and 'XXII Codes for special purposes' were excluded from the analysis since the aim was to examine age distributions and associations of only primary medical conditions (assumption 5). For the purposes of clustering only, diagnoses with a prevalence less than or equal to 0.01% (by unique number of patients) were excluded to reduce noise (assumption 6). Refer to online supplemental appendix for a more detailed breakdown of the exclusion process. For the purposes of the study the term 'diagnoses' refers only to those selected through the cohort and data specification described earlier; no independent or objective phenotyping was performed.

Analysis methods

Using unnormalised age distributions for diagnoses (see online supplemental appendix for further discussion), clustering analysis was performed to determine which

BMJ Health & Care Informatics: first published as 10.1136/bmjhci-2023-100963 on 29 July 2024. Downloaded from https://informatics.bmj.com on 10 September 2024 by guest. Protected by copyright.

diagnoses had similar age distributions. The kml[13] clustering R package was used to provide an implementation of the k-means clustering algorithm with standard Euclidean distance metrics. Diagnoses specific first age distributions (assumption 7) defined the model input (clustering was performed over $\mathbb{R}^{n\times m}$, where n defined number of diagnoses and m defined number of age brackets). A single observation thus represented the marginal probability of a given diagnosis occurring for the first time at a given age.

A hyperparameter of the k-means algorithm, defining the number of clusters, was required to be set before the algorithm was run. A value of k between 2 and 15 was proposed as an appropriate range. 15 was chosen as the maximum value based on previous data,[4] and the assumption that; as the number of clusters increased to 18 (the maximum age of patients), the model would overfit and clinical validity would decrease. This assumption was empirically observed in clusters below 15. Of the proposed values of k, it was non-trivial to select a final value for reporting. In line with previous work[4] the Calinski-Harabasz index (CH index) was used to assess cluster quality. Rather than maximising the CH index, the 'elbow' method was used to select an appropriate value of k which balanced cluster granularity against cluster validity. The silhouette score was also used to obtain an interpretable measure of performance of the chosen clusters.

Uncertainty quantification of preprocessing assumptions

As discussed, the reporting of ML analysis in healthcare would benefit from a greater emphasis on communicating the uncertainty of results. Of particular interest was communicating the uncertainty due to the various preprocessing decisions. As such, a Bayesian view of the analysis was assumed to emphasise the importance of preprocessing assumptions (priors).[14] While explicit Bayesian inferences were not made, the framework was used to guide the nature of the assumptions made and guide the kind of analysis that should be performed. The assumed Bayesian model is provided in online supplemental appendix for references and an interesting line of future work would be to perform such inferences using the model. However, the aim of this paper was to demonstrate the efficacy of just considering the analysis performed via a Bayesian view of probability and intended to motivate analysts to explore Bayesian methods further.

The preprocessing assumptions (described in the previous section) were grouped according to whether the uncertainty was introduced as a result of:

- The analysis hypothesis being too broad or;
- Was inherent to the data generating process.

Patient and public involvement

The public and patients were not involved in the definition or analysis performed in this study since the study predominantly focused on the representation of patient conditions via anonymous EHRs using an established secure data environment (SDE) and how these can be used for analysis with ML models. However, patients and public were involved throughout in the development of the SDE, including formal lay representation in the governance group.

RESULTS

Data overview

Data on 61 522 unique patients and 3315 unique (3, 4, 5 digits) ICD-10 diagnosis codes were examined. Combined, the diagnosis and age elements represented 618 124 data points over a 20-month period. The study population comprised 28 427 (46.2%) female, 32 843 (53.4%) male and 252 (0.4%) undocumented sex patients. With respect to ethnicity, the population consisted of 8247 Asian (13.4%), 4014 black (6.5%), 2582 (4.2%) mixed, 29 416 (47.8%) white, 4178 (6.8%) other and 13 085 (21.3%) undocumented.

Age distribution clustering

Figure 1 displays the mean CH score (across three random seeds) for the different numbers of clusters. There existed a level of ambiguity when assessing which cluster to select using the elbow method since the 'elbow' is ill-defined. In figure 1, clusters 4, 5 and potentially 6 would have been reasonable choices. In this instance, k=4 was chosen.

The resulting clusters of diagnoses are presented as the solid lines in figure 2. These clusters broadly aligned with empirical evidence and previous data[4] including clusters representing 'young infants', 'infants to young children', 'children' and 'teenagers/adolescents'. A further discussion regarding the efficacy of these clusters is presented in online supplemental appendix.

Similarly to the previous work,[15] the most frequent 50 primary diagnoses in each age cluster (online supplemental figures 5–8) demonstrated broad biological and clinical relevance, with congenital conditions such as cardiac defects, infantile haemangioma and other congenital anomalies dominating the young infant category, whereas for example, malignancies such as leukaemia clustered mainly within children, and mental health and musculoskeletal conditions mainly affected older children and adolescents. The clusters also aligned with known disease progression patterns. For example, dyskinetic cerebral palsy was followed by neuromuscular scoliosis occurring in older clusters. While these results are not 'new', they provide qualitative validity to the clusters identified and thus the inferences that may be drawn from subsequent analysis.

Uncertainty quantification of preprocessing assumptions

The outcome of the analysis on the effect of the preprocessing assumptions is presented in table 1.

The determination of the nature of uncertainty for each assumption was a result of assessing:

BMJ Health & Care Informatics: first published as 10.1136/bmjhci-2023-100963 on 29 July 2024. Downloaded from https://informatics.bmj.com on 10 September 2024 by guest. Protected by copyright.

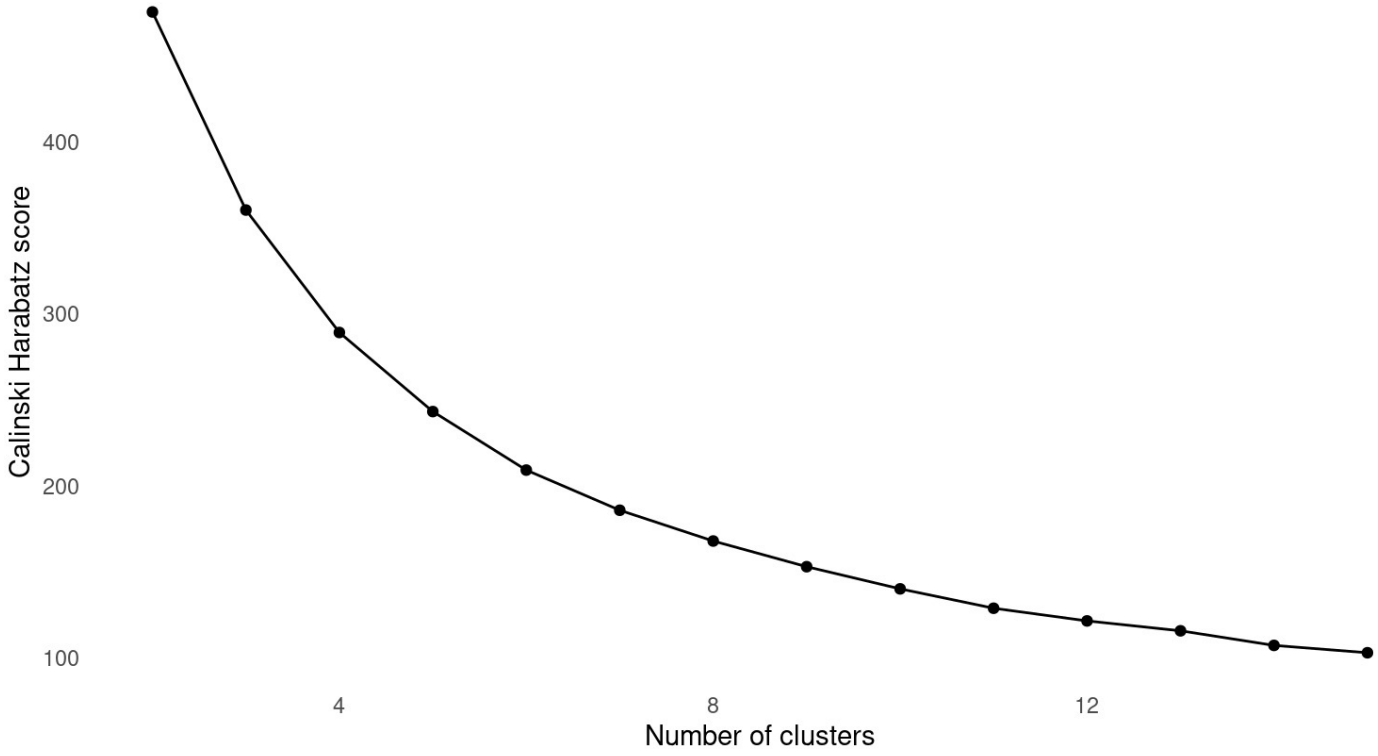

Figure 1 Mean Calinski-Harabasz score (over three random seeds) plotted against number of clusters that is, the value of k.

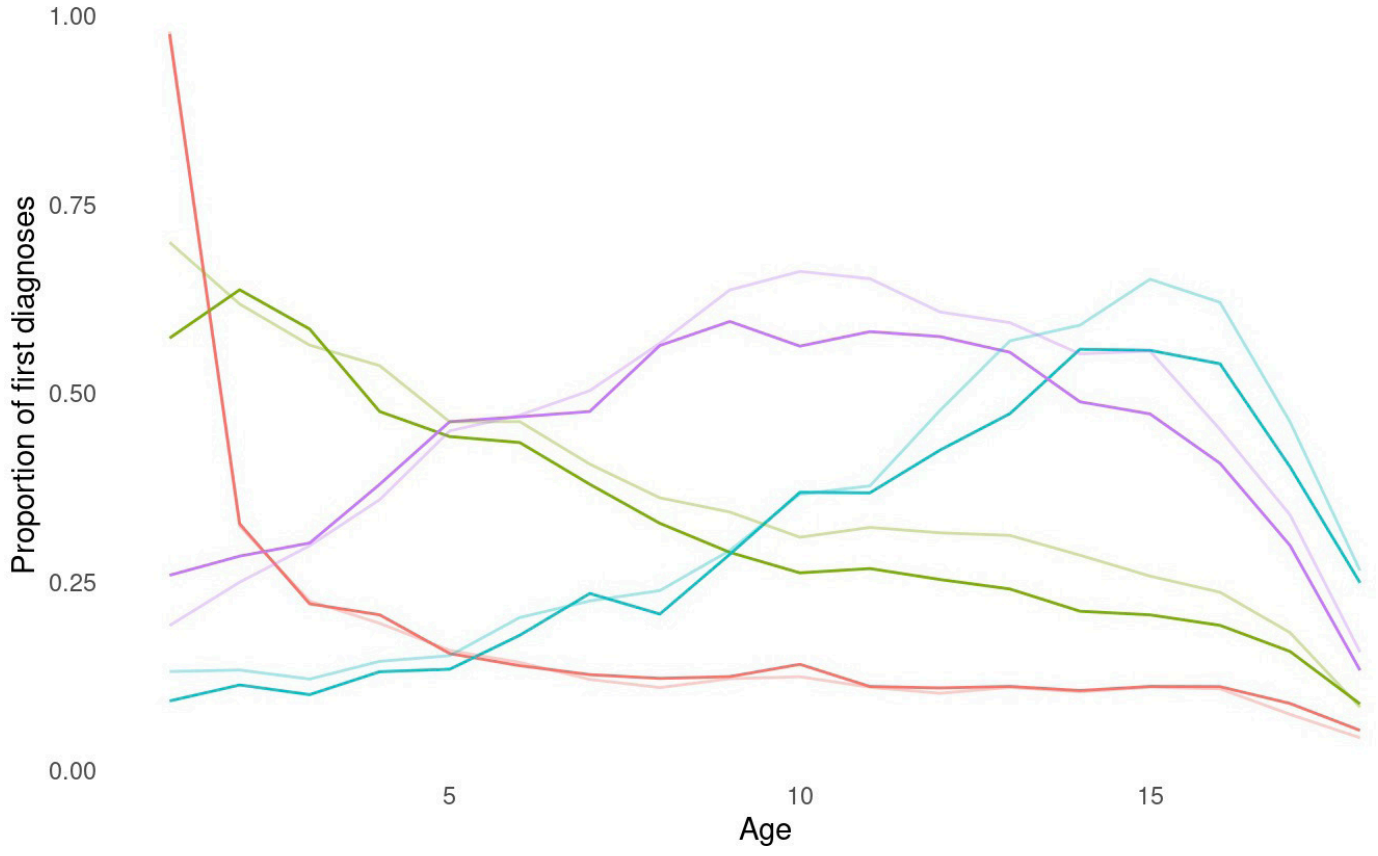

Figure 2 Clusters of first diagnosis age clusters. The solid lines define the clusters when applying assumption 2 as a preprocessing step, while the faint lines define the clusters when ignoring assumption 2. Ages 0–1 Ages 1–5 Ages 5–13 Ages 13–18.

BMJ Health & Care Informatics: first published as 10.1136/bmjhci-2023-100963 on 29 July 2024. Downloaded from https://informatics.bmj.com on 10 September 2024 by guest. Protected by copyright.

Table 1 Categorisation of preprocessing assumptions according to how each assumption contributes Bayesian uncertainty to the modelling process

Name	Description	Nature of uncertainty
Assumption 1	Use of ICD-10 codes assigned as part of the billing process	Inherent to data generating process
Assumption 2	Use of only codes defined as the 'primary diagnosis'	Inherent to data generating process
Assumption 3	Exclusion of rare diagnoses	Inherent to data generating process
Assumption 4	Exclusion of prenatal referrals	Too broad hypothesis
Assumption 5	Exclusion of diagnoses in chapters XVIII, XIX, XX, XXI, XXII	Too broad hypothesis
Assumption 6	Exclusion of diagnoses with a prevalence of less than 0.01%	Inherent to data generating process
Assumption 7	Use of 'first age distributions'	Too broad hypothesis

- The impact of altering the hypothesis of the analysis against;
- The complexity of accounting for, and communicating, the uncertainty of the assumption.

To address the uncertainty arising from assumptions with a 'too broad hypothesis' an example clinical query was required namely, 'Given a patient is aged X, what is the probability they will have condition Y?'. The query could thus be rephrased to address the uncertainty arising from each assumption, as follows:

- Assumption 4: 'Given a patient is aged X (*such that X>0*), what is the probability they will have condition Y?'.
- Assumption 5: 'Given a patient is aged X, what is the probability they will have condition Y (*such that Y is not in Z*)?'.
- Assumption 7: 'Given a patient is aged X, what is the probability they will *first present* with condition Y?'.

The uncertainty arising from the remaining assumptions was considered a result of either:

- Uncertainty regarding the data sampling process or;
- Uncertainty regarding the ICD 10 billing process.

In particular, the uncertainty arising from assumptions 3 and 6 was understood to be a result of the stochastic sampling process while assumptions 1 and 2 were related to the ICD billing process. Since the sampling uncertainty associated with assumptions 3 and 6 was not unique to healthcare, it was not investigated in this study.

Uncertainty arising from the ICD-10 billing process

As mentioned, the ICD-10 billing process is standardised and as such, ensures the assignment of a code is consistent (ignoring coder variance). However, the process also includes several nuisances. ICD-10 codes are defined as the 'primary diagnosis' if (generally speaking) they are the focus of treatment during episode of care. For example, assuming a patient was undergoing dialysis for chronic kidney disease then the dialysis would be the primary diagnosis. The uncertainty associated with 'primary diagnoses' could have been mitigated via a change in hypothesis however, the cognitive burden of requiring clinicians to consider the implications of only considering a diagnosis 'if it was of primary investigation' was deemed too high. The assumption 'ICD-10 codes' refers to codes never being assigned the primary diagnosis position because of additional coding rules.[16] Capturing the associated uncertainty of both assumptions was considered as defining *the probability that a diagnosis would never be assigned 'primary diagnosis' during a given patients interaction with the hospital, at a given age.*

Many codes that would have never been defined in the primary position, appear in chapter 'XVIII' and as such, had already been excluded. However, this did not account for all instances. Figure 3 defines a boxplot of the proportion of episodes where a diagnosis is ever defined in the primary position, conditional on age bucket. The median is around 0.5 suggesting that for 50% of diagnoses, for 50% of the time, the diagnosis is never 'primary' for a given episode of care and thus is not captured in the clustering analysis performed.

The faint lines in figure 2 define the clusters when ignoring assumption 2 and table 2 defines the transition probabilities between the two sets of cluster results. While the cluster centres themselves appeared relatively stable (based on figure 2), the diagnoses which defined the different clusters were less so (based on the transition probabilities in table 2). This analysis suggested that population level inferences regarding the derived clusters might have been robust to assumption 2, however, diagnosis level inferences were not. Furthermore, this analysis only accounted for the uncertainty arising from assumption 2 and not assumption 1. Additionally, using all diagnoses rather than just the primary diagnoses may have resulted in artificially inflated counts of diagnoses because of assumption 1. That is to say that accounting for the uncertainty in assumption 2 by removing the assumption, introduced a different source of uncertainty.

DISCUSSION

The results presented have demonstrated the efficacy of using EHR data for deriving quantitative insights for clinical practice. This was demonstrated quantitatively, since a silhouette score[17] of 0.15 was achieved and qualitatively, since the findings were broadly consistent with expected clinical and biological pathophysiologies. Given this, the result was considered as a validation of previous work

BMJ Health & Care Informatics: first published as 10.1136/bmjhci-2023-100963 on 29 July 2024. Downloaded from https://informatics.bmj.com on 10 September 2024 by guest. Protected by copyright.

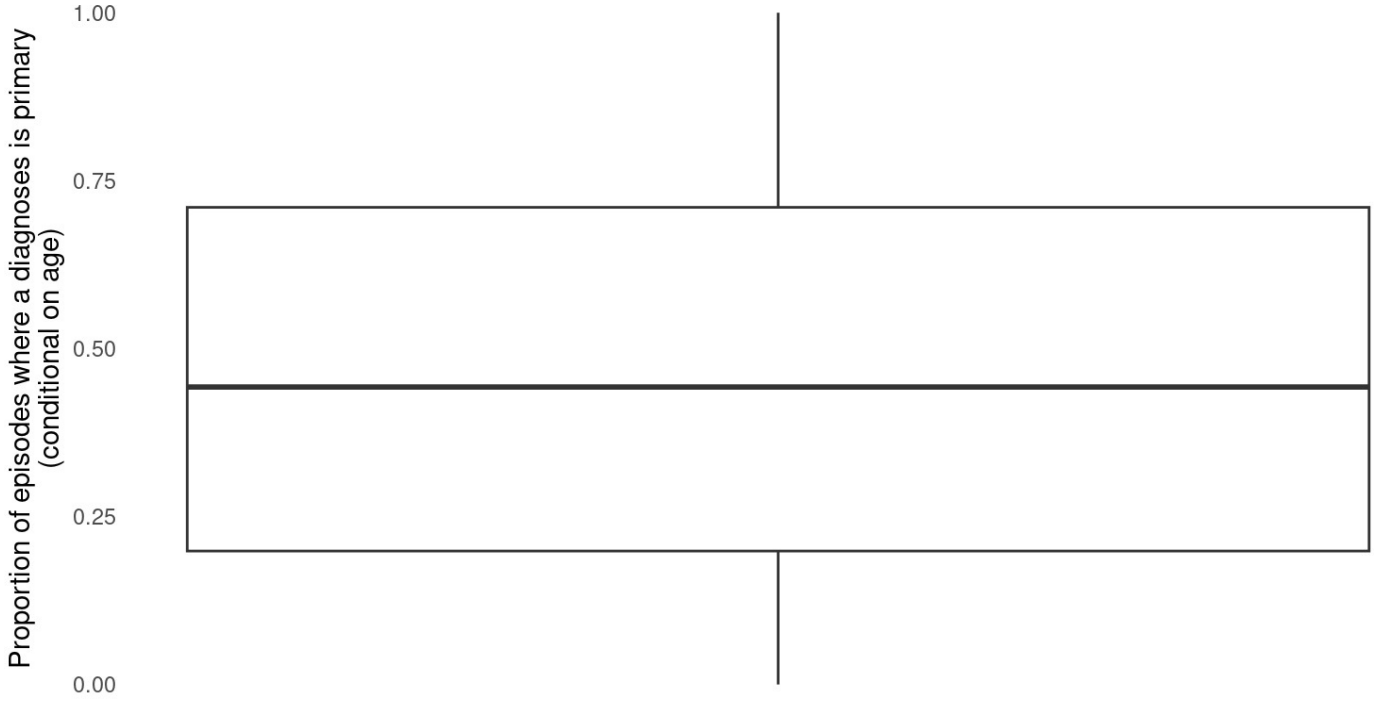

Figure 3 Boxplot of the proportion of patient episodes where a given diagnosis appears as the primary diagnosis at least once within age bucket where the episode started (or older).

using a completely independent dataset from a different population.[4] Even accounting (partially) for the bias induced by the ICD-10 coding process, the learnt clusters were identified as grouping conditions predominantly affecting: young infants (ie, broadly between the ages 0 and 1); infants to young children (ie, broadly between the ages 1 and 5) with a peak in very early years; children (ie, broadly between the ages 5 and 13) and; teenagers and adolescents (ie, broadly between the ages 13 and 18).

However, care should be taken before generalising the findings of this study beyond the assumptions made in the analysis. ML methods generally assume the data on which the model was trained follows the same distribution as the data on which the model/insights from the model are being applied. While the general methodology of clustering diagnoses by defining age distributions may be applied elsewhere, applying the findings of this analysis to populations from different hospitals or populations, may result in invalid conclusions. Furthermore, the relationships presented here are associations and no suggestion is made regarding causality.

Future work could consider additional patient features as well as the use of other clustering methods. By their construction, ML methods assign different assumptions on the data which are required to hold for the clustering to be optimal and valid. For example, k-means clustering assumes the data forms independent spherical groups with constant radius.[18] While the purpose of the current study was to demonstrate a proof of concept of the approach using routine EHR data and standard methods, given that the final clustering approach achieved a silhouette score of 0.15 (ie, less than 1), it may be possible to find superior separation of clusters with additional methods.

The preprocessing assumptions made when performing the modelling have been explicitly demonstrated to have a large effect on the results, an observation which has not generally been presented within the majority of applied ML for healthcare literature. Addressing uncertainty via altering the analysis hypothesis, while possible, required being extremely precise. Notably, the example of 'Given a patient is aged X, what is the probability they will have condition Y?' is not the only query a lay practitioner may have when presented with analysis defining patient age clusters. This suggested the need for further research into how ML tools and results can be better disseminated and reliably scaled and applied across healthcare organisations. Considering PedMap[4] as an example, uncertainty resulting from a broad hypothesis could be reduced by tailoring the application views even further, to be focused on 'clinical questions' such as the one considered in this paper. It should be noted that the challenges associated

Table 2 Transition probabilities of a diagnosis moving from a given cluster when using assumption 2 as a preprocessing step (rows) to a given cluster when not using assumption 2 as a preprocessing step (columns)

	Clusters without assumption 2			
Clusters using assumption 2	**A**	**B**	**C**	**D**
A	0.89	0.1	0.01	0.01
B	0.23	0.74	0.01	0.02
C	0.05	0.05	0.74	0.16
D	0.04	0.23	0.05	0.68

BMJ Health & Care Informatics: first published as 10.1136/bmjhci-2023-100963 on 29 July 2024. Downloaded from https://informatics.bmj.com on 10 September 2024 by guest. Protected by copyright.

with preprocessing ICD diagnosis codes are not unique to the analysis presented or dataset used and are prevalent throughout clinical informatic applications. Considering how to remove such biases and derive independent data representations that 'truly' reflect the holistic health status of patients (including diagnosis) is an interesting avenue of future work which may improve the results of the analysis presented in this paper.

Bayesian analysis has been generally conceived as difficult and subjective however, the results of the current study have demonstrated that even engaging with the Bayesian framework (without performing any Bayesian inferences) was intuitive and could prevent over interpretation of results. Rather than point estimates of clusters (such as those provided in PedMap), a distribution of clusters could be provided as output, such as that provided in figure 2. In the absence of techniques to directly remove the bias arising from preprocessing assumptions, a Bayesian inference approach to defining this uncertainty would be an interesting next step. Additional interesting avenues of future work might be to focus on expanding the 'plug and play' ML APIs to allow practitioners to incorporate Bayesian uncertainties over preprocessing steps, with only an introductory understanding of Bayesian inference.

Contributors JWS performed the core analysis with equal supervision by SB, WAB and NJS. DK and JB performed the data extraction and engineering for the project. EP, AS, RP, AMT and HH and all previously named authors contributed via discussions and in reviewing the manuscript.

Funding JWS is funded by CIRP via GOSHCC. The study and researchers were part funded through the NIHR GOSH Biomedical Research Centre and the NIHR UCLH Biomedical Research Centre.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Use of the routine deidentified data was approved by the appropriate research ethics committee and HRA approval (REC approval number 17/LO/0008) and analysis was carried out within the GOSH Digital Research Environment according to standard procedures. JWS is the is the project guarantor.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. Deidentified patient data not available but code available on request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, 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.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs
Joshua William Spear http://orcid.org/0009-0005-9366-8185
Daniel Key http://orcid.org/0000-0002-9559-1784
John Booth http://orcid.org/0000-0003-4357-1324
Neil J Sebire http://orcid.org/0000-0001-5348-9063

REFERENCES

1 Zhao J, Papapetrou P, Asker L, *et al*. Learning from heterogeneous temporal data in electronic health records. *J Biomed Inform* 2017;65:105–19.
2 Violán C, Foguet-Boreu Q, Fernández-Bertolín S, *et al*. Soft clustering using real-world data for the identification of multimorbidity patterns in an elderly population: cross-sectional study in a mediterranean population. *BMJ Open* 2019;9.
3 Zhang JHaipingZAnalysis of clustering algorithms in machine learning for healthcare data. *J Commer Biotechnol* 2022;27.
4 Li H, Yu G, Dong C, *et al*. Pedmap: a pediatric diseases map generated from clinical big data from Hangzhou, China. *Sci Rep* 2019;9.
5 Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn mach learn python. *J Mach Learn Res* 2011;12.
6 Amazon Web Services. AWS cloud products. Available: https://aws.amazon.com/ [Accessed 12 Feb 2024].
7 Al-Haddad C, BouGhannam A, Abdul Fattah M, *et al*. Patterns of uveitis in children according to age: comparison of visual outcomes and complications in a tertiary center. *BMC Ophthalmol* 2019;19.
8 Kaplan A, Hardjojo A, Yu S, *et al*. Asthma across age: insights from primary care. *Front Pediatr* 2019;7:162.
9 Simon M, Rigou A, Le Moal J, *et al*. Epidemiology of childhood hyperthyroidism in France: a nationwide population-based study. *J Clin Endocrinol Metab* 2018;103:2980–7.
10 Choi E, Bahadori MT, Searles E, *et al*. Multi-layer representation learning for medical concepts. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016
11 Epic|…with the patient at the heart. Available: https://www.epic.com/ [Accessed 13 Aug 2020].
12 Trusted research environments - HDR UK. Available: https://www.hdruk.ac.uk/access-to-health-data/trusted-research-environments/ [Accessed 4 Jan 2022].
13 Genolini C, Alacoque X, Sentenac M, *et al*. Kml and Kml3D: R packages to cluster longitudinal data. *J Stat Softw* 2015;65.
14 Winkler RL. Why Bayesian analysis hasn't caught on in healthcare decision making. *Int J Technol Assess Health Care* 2001;17:56–66.
15 Kuan V, Denaxas S, Gonzalez-Izquierdo A, *et al*. A chronological map of 308 physical and mental health conditions from 4 million individuals in the english national health service. *Lancet Digit Health* 2019;1.
16 Terminology and classifications delivery service, national clinical coding standards ICD-10. Leeds, 2021.
17 Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987;20:53–65.
18 Bishop CM. Pattern recogniton and machine learning, information science and statistics. 2006;738.

BMJ Health & Care Informatics: first published as 10.1136/bmjhci-2023-100963 on 29 July 2024. Downloaded from https://informatics.bmj.com on 10 September 2024 by guest. Protected by copyright.