Development of a Semi-Automated Database for Adult Congenital Heart Disease Patients

Short Title: ACHD semi-automated Database

Authors: Shourya Verma BSc[*], Muhammet Alkan BSc[*], Dr Fani Deligianni PhD[+], Dr Christos Anagnostopoulos PhD, Gerhard Diller MD PhD, Lisa Walker, Fiona C Johnston, Mark Danton, Hamish Walker, Lorna Swan MD, Amanda Hunter PhD, Alex McGuire, Martin Dawes, Sharon Stott, Mitchell Lyndsey, Dr Niki Walker PhD, Dr Gruschen Veldtman MBChB FRCP[+]

--

From: Scottish Adult Congenital Heart Disease Service, Royal Brompton Hospital, King's College London, Golden Jubilee National Hospital, and the School of Computing Science at University of Glasgow

Keywords: Congenital Heart Disease, Database, Registry, Machine Learning, outcomes

Word Count: 4103

Correspondence:

Dr Gruschen R Veldtman, gruschen.veldtman@gjnh.scot.nhs.uk, SACCS, Golden Jubilee National Hospital, Glasgow, G814DY, Scotland, United Kingdom.

Shourya Verma BSc[*] and Muhammet Alkan BSc[*] have contributed equally.

Fani Deligianni PhD[+] and Dr Gruschen Veldtman MBChB FRCP[+] are equal senior authors.

Abstract:

Background. Databases for Congenital Heart Disease (CHD) are effective in delivering accessible datasets ready for statistical inference. Data collection hitherto has however been labour and time intensive and has required substantial financial support to ensure sustainability. We propose here creation and piloting of a semiautomated technique for data extraction from clinic letters to populate a clinical database.

Methods. PDF formatted clinic letters stored in a local folder, through a series of algorithms underwent data extraction, pre-processing and analysis. Specific patient information (*diagnoses, diagnostic complexity, interventions, arrhythmia, medications, and demographic data*) was processed into text files and structured data tables, used to populate a database. A specific data validation schema was pre-defined to verify and accommodate the information populating the database. Unsupervised learning in the form of a dimensionality reduction technique was used to project data into two dimensions and visualise their intrinsic structure in relation to the diagnosis, medication, intervention, and ESC classification lists of disease complexity. Nine-three randomly selected letters were manually reviewed for accuracy.

Results: 1409 consecutive outpatient clinic letters were used to populate the Scottish Adult Congenital Cardiac Database. Mean patient age was 35.4yrs, 47.6% female with 698, 49.5% having moderately complex, 369, 26.1% greatly complex, and 284, 20.1%, mildly complexity lesions. Individual diagnoses were successfully extracted in 96.95%, and demographic data was extracted in 100% of letters. Data extraction, database upload, data analysis and visualisation took 571 seconds (9.51 minutes). Manual data extraction in the categories of diagnoses, intervention and medications yielded accuracy of the computer algorithm in 94%, 93%, and 93% respectively.

Conclusions: Semi-automated data extraction from clinic letters into a database can be successfully achieved with a high degree of accuracy and efficiency.

Brief Abstract

Semi-automated databases integrating artificial intelligence may be more efficient, less prone to error, and enable processing of large and diverse origin data. In this analysis of 1409 outpatient clinic letters we demonstrate successful data extraction from clinic letters, and data upload into a commercially available database. Accuracy of 94%, 93% and 93% were demonstrated respectively in diagnosis intervention and medication identification. Semi automated clinical databases can be achieved with a high degree of accuracy.

Introduction:

Congenital heart disease (CHD) is the most common form of congenital malformation affecting 1 in every 180 live births in the United Kingdom. With [1]better surgery, early overall survival has improved to more than 94%, leading to a greater number of affected individuals reaching adult life. Other developed countries have reported similar improvements and prevalence of ACHD is now documented at approximately 4 per 1000 of adults[1]. The need for clinical databases which can identify and characterise local and regional ACHD patient populations has become a more urgent necessity. Some developed countries have already invested for several years in national ACHD databases that have served as an important resource for service delivery planning, research, and tracking trends in outcomes. Among the most notable databases are included the CONCOR registries[1], BELODAC (Belgian Congenital Heart Database), SWEDCON (Swedish Registry of Congenital Heart Disease)[2], and NICOR (the National Institute for Cardiovascular Outcomes Research in the United Kingdom)[3], and more recently the Congenital Heart Disease Initiative in the US. These databases, though highly effective, are labour intensive, require manpower, require extensive time, and may be prone to error.

With the evolving sophistication, processing power and ease of computer language programming, and the rapid development of machine learning, the concept of greater automation of data extraction from electronic health records and specifically clinic letters is appealing. It has the potential to reduce error, obviate the requirement for multiple individuals to be involved, and may only take a fraction of the time to process data. Previous work has shown that it is possible to accurately extract clinical entities and relations from radiology reports[4]. We hypothesized that we would be able to process clinic letters, produced by a variety of physicians, with different writing styles, and extract clinical entities into a database along with demographic and other clinical characteristics. Secondly, we hypothesized that we would be able to apply unsupervised machine learning techniques to the datasets to test the ability to create an "intelligent" database. Our objectives were to test the performance of data extraction on various grammatically and syntactically diverse sets of clinical letters and store them on a local database.

In Scotland, care for adults with congenital heart disease is centralized, being commissioned through the National Service Division and is hosted by the Golden Jubilee National Hospital in Glasgow, where the present work is being undertaken. The Scottish Adult Congenital Cardiac Service (SACCS) currently provides care for an estimated 3000 moderate or complex patients, and a further 7-8000 patients with simpler forms of congenital heart disease who are seen less frequently, and generally are primarily cared for by cardiologists in local cardiac centers. The authors undertook to create a clinical database, with a

minimal dataset of around 1400 patients. This would allow for accurate identification and analysis of all SACCS ACHD patients' diagnoses, demographic data, and medication in this pilot project.

Methodology:

Data Extraction from clinic letters

An initial 60 randomly selected letters (approximately 12 per individual cardiology consultant) were used to develop algorithms in Python (a high-level computer programming language that is in common use) for data extraction that was sensitive to different writing styles. Consecutive outpatient clinic letters of patients seen in the Golden Jubilee National Hospital between 1 June 2016 to 1 June 2018 were selected for analysis, allowing for sufficient follow up time to discern outcomes. One thousand four hundred and twenty outpatient clinic letters in PDF format, were manually placed in a "screening" folder. The Python programming algorithm first converts all the PDF letters into text files and stores them into a secondary folder. Next, using keywords appearing in the text, and regular expression techniques, the algorithm extracts demographic and clinical information including hospital number, date of birth, name, gender, diagnosis, intervention, medication, clinic date, post code, health board, height, weight, and arrhythmia class. This is a common approach to natural language processing (NLP)[5] and uses pattern matching techniques to retrieve textual data. The extracted data is then pre-processed i.e., removal of unwanted characters, bullet points, numbers, double spaces etc. Incomplete letters, i.e., not containing the diagnosis, intervention and medication information are excluded through the algorithm. The European Society of Cardiology ACHD lesion complexity classification is then assigned by matching keywords in the diagnosis section with a matched permutation of different cardiac condition names that appear under each classification. The European Society of Cardiology (ESC)[6] classification of mild, moderate, and great complexity of CHD, was used in the present study. Aortopathy was incorporated as a separate category. An additional category was created for patients that did not have CHD. Medication names were obtained from the British National Formulary (BNF)[7] published list of medication and matched for all medications listed in the patient letter. Intervention names were taken from published CHD treatment lists[8]. Similarly, arrhythmias were identified in broad standardised categories including supraventricular and ventricular arrhythmias, devices, and ablation procedures.

Data Processing and Populating the Database

In total 17 clinical variables were extracted into columns of a structured table: hospital record number, name, date of birth, complete diagnosis, list of diagnosis, complete intervention, list of interventions, complete medication, list of medication, ESC classification, arrhythmia, clinic date, health board, post code, gender, height, weight. These 17 columns of data are stored as text file and uploaded to a commercially

available database (NoSQL/MongoDB)[9]. To ensure that the information entered in the database is not corrupted and does not already exist, certain information is required in order for a patient record to be inserted successfully in the database. This information includes the hospital number as an integer, the patient name and health board as text variables, and the clinic date as date (dd-mm-yyyy). Furthermore, the information provided to the database should be compatible with its predefined structure. The database upload algorithm reads the text file created by the data extraction algorithm and cross-checks it with the data already stored on the database. It identifies new patients if the hospital number is not already stored on the database, and updates information for patients already present in the database. This allows only the single latest version of the patient information to be present on the database at any one time. Once the data is uploaded to the database, the algorithm downloads the latest complete data from the newly updated database as a text file, ready to be inspected by the clinicians and to be used for various data analysis and data visualization tasks. Eighteen letters were rejected and removed by the algorithm, thus uploading 1409 letters to the database. This strategy allowed us to obtain a diverse list of letters dictated by successful different cardiologists over a span of time.

Unsupervised machine learning - Dimensionality Reduction for Visualisation

For unsupervised learning we chose a technique known as t-distributed Stochastic Neighbour Embedding (t-SNE)[10]. In essence, this technique simplifies more complex non-linear or multidimensional data into 2 or 3 dimensions that are easier to visualise and interpret. Such techniques can be considered as projections of the data, similar to how photos encode information about a three-dimensional world into 2 dimensional visual displays. T-SNE transforms the similarity, assigned as a distance, as a probability distribution over all the potential patient pairs[17]. Thus, if two patients' clinical characteristics are close in high dimensional space, i.e. in the initial complex dataset, then they are going to be assigned with a high probability of being neighbour points (i.e. located in close proximity). The aim of the low dimensional embedding/projections is to approximate the data distributions as closely as possible with the original, high-dimensional dataset.

As a pilot exploration of applying artificial intelligence to the database, we applied unsupervised learning to the following domains including diagnosis, intervention, medication, and ESC classification lists. In this way, we qualitatively assessed the data-driven low-dimensional embeddings that form ‘notional clusters’ of patients with relation to our knowledge about the data.

Clustering analyses on the t-SNE embeddings are summarized in the Appendix.

Data Validation:

A random sample of 5% of patient letters were manually checked to assess the accuracy of the computer algorithm in terms of diagnosis, intervention and medications. For this purpose, a minimum of 70 patient letters were required, assuming by the nature of this study that the accuracy would be approximately 90%. A sample size in excess of 70 would allow the accuracy to be estimated to within +/- 6% with 95% confidence. We somewhat arbitrarily chose 93, a number in excess of the required 70.

The individual patient processed datasets were manually reviewed by an independent skilled observer (FCJ) who was a final year medical student working in the ACHD department. The computing algorithm was scored for accuracy in extracting diagnostic, interventional and medication related information.

Results:

From a total of approximately 10-11 000 Scottish ACHD patients, and 3000 under regular follow-up, 1409 patient letters were analysed as described above. Mean age was 35.4yrs (SD 14.6 and 75-25% IQR [45,24]). Seven hundred and thirty-seven patients were male (52.3%) whereas 672 patients female (47.6%). Two hundred and eighty-four patients had mild (20.1%), 698 moderate (49.5%), and 369 great (26.1%) complexity CHD, ESC classification. Seven patients had a primary Aortopathy (0.5%) such as Marfan syndrome, 43 had no classifiable (3.1%) congenital heart disease such as anomalous right subclavian artery, and 8 did not have CHD, for example carcinoid disease (0.5%). Completeness of data extraction was tested in each of the major categories summarized in Table 1.

It was possible to categorize patients by individual health boards, gender, diagnosis, intervention, medication, and ESC classification among other patient characteristics. Diagnoses, and diagnostic groups are summarized in Figures 1(a,b) (diagnoses, and then ESC classification), and the medication groups are summarized in Figure 2. The most common condition was Tetralogy of Fallot, in 197 patients, followed by Pulmonary Atresia, in 96 patients.

The time taken for the data extraction, database population and analysis processes are summarized in Table 2. The machine was running on a 64-bit system with 6 core, 3.00 GHz processor, 8 GB RAM, and 4GB Intel UHD Graphics 630.

Figure 3 shows the results of the dimensionality reduction, within the domains of diagnosis, intervention, medication, and ESC classification lists. The selection of optimal t-SNE parameters and how they affect results is shown in supplementary appendix figure S1. Figure S2 in supplementary material shows how the low-dimensional embeddings coloured according to the disease complexity compare with data-driven (k-means) clustering. The diagnostic components of the low-dimensional embeddings were primed with apriori data labelling by human design into mild, moderate or great complexity disease as seen in panel 3. The corresponding ESC classification lists are visually depicted as blue, orange and green respectively. ESC-based embeddings are directly linked to disease complexity, and this is also reflected in the almost perfect agreement between the embeddings and the classification. On the other hand, the diagnostic embeddings do not necessarily overlap with the diagnostic classification. Nevertheless, diagnostic embeddings resemble better disease complexity compared to interventions and medication embeddings. This is also shown quantitatively in the appendix.

Validation:

The screening and processing algorithms were able to accurately identify diagnoses, intervention and medications in 94%, 93%, and 93% accurately. Inaccuracies were due to the following factors: A genetic diagnosis not listed (DiGeorge syndrome) in 1case; in another letter diagnoses were listed with social history in the original letter, and this was extracted into the diagnostic list; in another 3 letters, clinical diagnoses and interventions were listed in the same section under diagnosis, and this was inappropriately listed as diagnoses only; in another 2 letters, that did not list either diagnoses or interventions, the algorithm incorrectly retrieved diagnostic/intervention history.

Discussion:

Clinical databases are required to inform clinicians, researchers and policy-makers about health outcomes and the economic burden of strategic decisions. With the advent of advanced computing algorithms and machine learning, data that was previously manually extracted from clinical records can now potentially be automatically or semi automatically extracted. This ability provides enormous time efficiency, avoidance of human error inherent in manual data extraction, and facilitates more affordable clinical database management. Furthermore, it enables reproducibility in research studies and it allows efficient application of machine learning algorithms that cannot be used directly on unstructured datasets.

In this paper, to the best of our knowledge, we introduced for the first time the development of a semi-automated database populated *by extracting* data from Adult Congenital Heart Disease patient clinic letters

into a working and live clinical database. This process was tested on 1409 clinic letters and achieved a high degree of efficiency and performance. Inaccuracy related mainly to how the initial letters were written, and we believe can be corrected with further modification of the natural language processing techniques currently employed. It requires minimal human resources to operate and performs the data extraction, pre-processing, and database population tasks within 40 seconds for 100 clinic letters. We were able to confirm that conventional, albeit basic, machine learning techniques can indeed be applied within the database and presented in a dynamic fashion as the data is updated. The database has the potential to facilitate the development of a rigorous framework to offer clinical decision support via descriptive, diagnostic, and subsequently predictive and prescriptive analytics.

Databases have been successfully used over the past 2-3 decades in ACHD. Some of the most successful and productive databases currently still in use include the Dutch CONCOR registry[12], the Belgian BELCODAC registry[2], SWEDCON (Sweden) and more recently the Adult Congenital Heart Disease Initiative in the United States[13]. The data collection and storage for CONCOR registry is performed by trained nurses who manually enter patient data into the database with unique patient identifying numbers. The German National Register[14] is one of the largest worldwide registry for CHD with approximately 60,000 patients. The data recorded in this registry is widely used for descriptive data analysis in research of CHD. Similarly, in the BELCODAC registry, data was obtained from electronic health records in 3 Belgian hospitals. Additionally, mortality and socio-economic data were derived from the Belgian Statistical Office. Clinical data is also manually obtained from the SWEDCON registry.

The strategy employed in the current database follows relatively basic strategies used in natural language processing to derive specific keywords which are tabulated in a structured fashion and then transferred to a versatile database manager. The technique is highly effective at extracting available information in letters that are “complete” and where information is organized into diagnoses, interventions, and medications as separate categories. In less well-organized letters, current performance is reduced. The investigators are currently working on natural language processing and annotation techniques to further improve performance of this strategy. Abbreviated informational letters are excluded by the algorithms as they do not contain the minimal data needed to populate and update the database. The success of such simple language processing is encouraging for further future work. This allows a high level of customizability in creating data extraction algorithms based on regular expressions. The data represented in this study was across a wide range of socio-economic status, reflected in the inclusion of all the different health boards within Scotland. We also were able to have a representation of varying disease complexity as indicated in the ESC complexity classification.

Broberg et al.[15] similarly investigated whether data can be extracted and consolidated from an EHR for ACHD patients. They collected multiple different data types from 9 variables which included demographics, medical history, medication, visits, nursing, imaging, obstetrics, interventions, and hospitalizations. Their algorithms successfully collected uniform existing data on CHD patients from different tertiary sectors but could not consistently extract the target variables since they were not universally available. The authors also highlighted the challenges of automating regular data updates from different institutions providing information on patients located in their regions. In the present dataset, continuity and sustainability of this database is likely to be more straightforward if referral centres provide letters with a quality requirement including diagnoses, intervention, and medications. Further evaluations are necessary to document sustainability of such semi-automated datasets. We anticipate maintenance of the database will require approximately 1-2 hours monthly, or shorter weekly updates, as new data becomes available. We also anticipate that letter deposition into the single common folder would become part of the workflow of the ACHD group practice. Currently, and for most practices, data is extracted separately to normal clinical workflow patterns to populate databases.

It was particularly gratifying to demonstrate the ability of unsupervised (dimensionality reduction) learning to provide intuitive interpretation of complex patients information related to the diagnostic, interventional and medication categories. The low-dimensional embeddings confirmed that diagnoses do not overlap necessarily with diagnostic classification in a specific manner, as patients carry multiple comorbidities, and also carry non-genital cardiac diagnoses. We were also able to demonstrate that medications do not follow a strict pattern of conformity to diagnostic classification, although it is evident that more moderate and complex patients are likely to receive medications. In this context we have used unsupervised machine learning as a visualisation and diagnostic tool demonstrating how powerful machine learning algorithms can draw useful insights from population-based data stored in a clinical database.

The ultimate intention of the database is to be extended with additional laboratory data, exercise test information as well as imaging data, which will be used to develop machine learning strategies to inform accurate risk prediction, like the work of Diller et al[16]. Artificial intelligence has already been successfully applied to several fields in medicine including functional MRI, ECG, and other time dependent and structured electronic health records. The next steps involve developing risk prediction models via a rigorous validation and monitoring framework. This will support clinical decision-making systems, which would help clinicians predict outcomes for ACHD patients and explain what the risk is based on.

Conclusions:

We were able to demonstrate successful semi-automation of data extraction for patient level information from clinic letters and importation into a working database with a high degree of accuracy, attractive time efficiency, minimal requirement for human resources and with the ability to apply widely used unsupervised machine learning techniques to inspect the current dataset and identifying meaningful clusters. This strategy holds promise for further enhancement of the database to include more sophisticated information and sets the groundwork for the development of "intelligent databases" that can promote high quality insight to clinicians for individualized care delivery in the clinical arena.

Limitations:

Though this work is prospective and was done in a proportionately large volume of patients, there were still some important limitations in this work. We used straightforward algorithms to parse data from clinic letters and they were confounded by letters that did not conform to standard formatting. Though we were able to program the algorithm to cope with a variety of different styles amongst our local physicians, it is not possible to account for all the different permutations that may arise including less well organized letters. Human error also may confound this process through mislabelled or missing information from the relevant sections in the letter. Therefore, in the next iteration of this work, we propose to use greater sophistication of natural language processing using annotation tools to overcome this limitation. Further, we also used basic unsupervised learning techniques. This was however purposeful to simply test the hypothesis that we could incorporate learning within the database.

Moreover, unsupervised learning is used as a visualisation technique to highlight the potential of incorporating advanced machine learning algorithms to extract novel information about data stored in the database. Further work is required to test the risk prediction models constructed based on the extracted information.

Disclosures and Funding

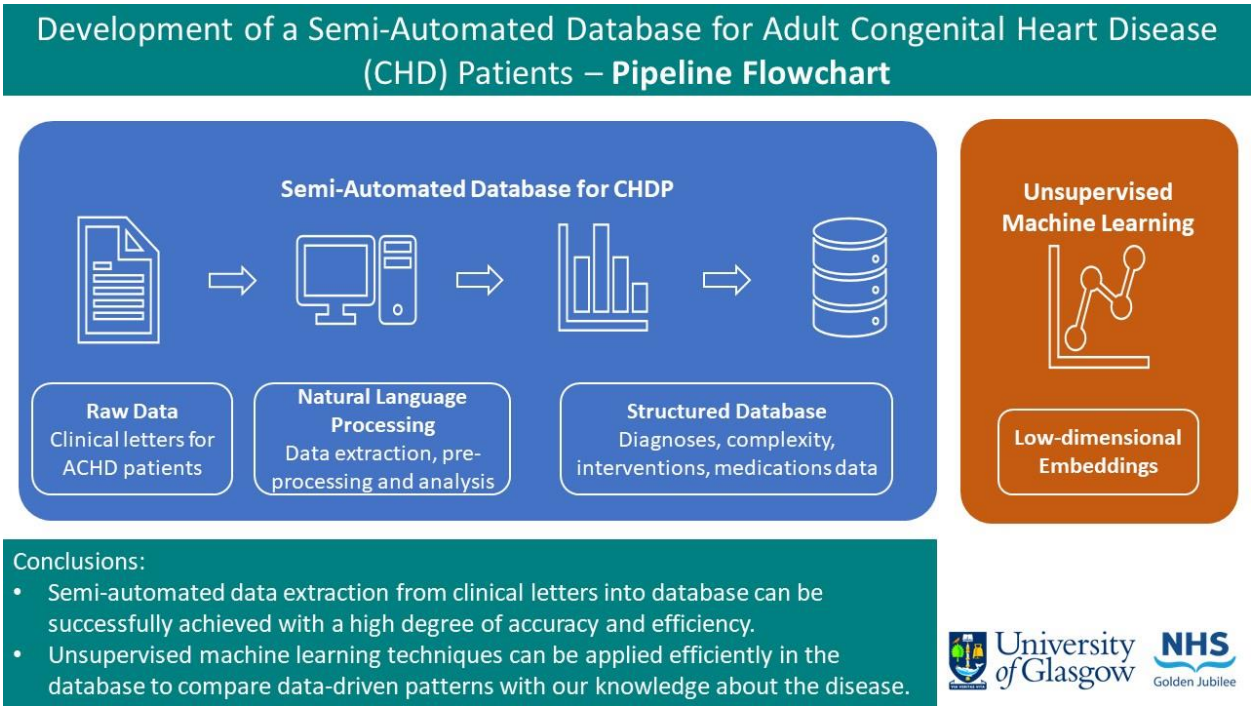
This work was self funded and supported by the School of Computing Science at University of Glasgow and Golden Jubilee National Hospital leadership and governance teams. None of the authors have anything to disclose relevant to this work.

References:

1. Marelli AJ, Ionescu-Ittu R, Mackie AS, Guo L, Dendukuri N, Kaouache M. Lifetime prevalence of congenital heart disease in the general population from 2000 to 2010. Circulation 2014;130:749-56.
2. Ombelet F, Goossens E, Willems R et al. Creating the BELgian COngenital heart disease database combining administrative and clinical data (BELCODAC): Rationale, design and methodology. Int J Cardiol 2020;316:72-78.
3. Rashid M, Ludman PF, Mamas MA. British Cardiovascular Intervention Society registry framework: a quality improvement initiative on behalf of the National Institute of Cardiovascular Outcomes Research (NICOR). Eur Heart J Qual Care Clin Outcomes 2019;5:292-297.
4. Jain S, Agrawal A, Saporta A et al. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. arXiv preprint arXiv:210614463 2021.
5. Pudasaini S, Shakya S, Lamichhane S, Adhikari S, Tamang A, Adhikari S. Application of NLP for Information Extraction from Unstructured Documents. Singapore: Springer Singapore, 2022:695-704.
6. Baumgartner H, De Backer J, Babu-Narayan SV et al. 2020 ESC Guidelines for the management of adult congenital heart disease: The Task Force for the management of adult congenital heart disease of the European Society of Cardiology (ESC). Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC), International Society for Adult Congenital Heart Disease (ISACHD). European heart journal 2021;42:563-645.
7. J. C. British national formulary. BMJ Publishing and the Royal Pharmaceutical Society, 2021.
8. Warnes CA. Congenital heart disease in adults. Mayo Clinic Proceedings: Elsevier, 1992:505.
9. Krishnan HE, M.Sudheep & Santhanakrishnan, T. MongoDB – a comparison with NoSQL databases. International Journal of Scientific and Engineering Research 2016;7:1035-1037.
10. Pedregosa F, Varoquaux G, Gramfort A et al. Scikit-learn: Machine learning in Python. The Journal of machine Learning research 2011;12:2825-2830.
11. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426 2018.
12. Vander Velde E, Vriend J, Mannens M, Uiterwaal C, Brand R, Mulder BJ. CONCOR, an initiative towards a national registry and DNA-bank of patients with congenital heart disease in the Netherlands: rationale, design, and first results. European journal of epidemiology 2005;20:549-557.
13. Bodell A, Björkhem G, Thilén U, Naumburg E. National quality register of congenital heart diseases–can we trust the data? Journal of Congenital Cardiology 2017;1:1-8.
14. Helm PC, Koerten M-A, Abdul-Khaliq H, Baumgartner H, Kececioglu D, Bauer UM. Representativeness of the German National Register for Congenital Heart Defects: a clinically oriented analysis. Cardiology in the Young 2016;26:921-926.
15. Broberg CS, Mitchell J, Rehel S et al. Electronic medical record integration with a database for adult congenital heart disease: early experience and progress in automating multicenter data collection. International journal of cardiology 2015;196:178-182.

16. Diller GP, Orwat S, Vahle J et al. Prediction of prognosis in patients with tetralogy of Fallot based on deep learning imaging analysis. Heart 2020;106:1007-1014.
17. Hinton GE, Roweis S. Stochastic neighbor embedding. Advances in neural information processing systems 2002;15.
18. Belkina AC, Ciccolella CO, Anno R, Halpert R, Spidlen J, Snyder-Cappione JE. Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. Nature communications 2019;10:1-12.
19. Van Der Maaten L, Weinberger K. Stochastic triplet embedding. 2012 IEEE International Workshop on Machine Learning for Signal Processing: IEEE, 2012:1-6.

Graphic Abstract



Tables and Figures

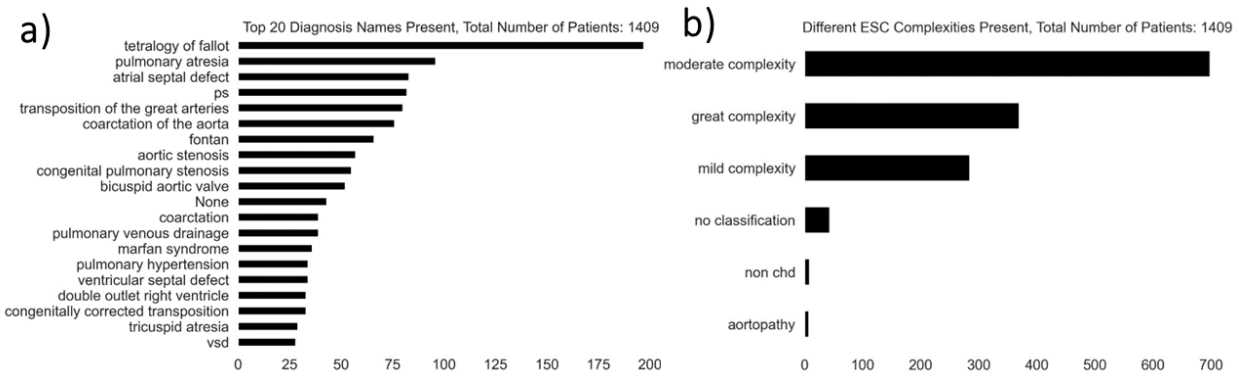
Table 1 : Completeness of data extraction for all attributes.

Attributes	CHI[1]	Name	DOB[2]	Clinic Date	Post Code	Health-Board	Gender	Height/Weight
Completeness (%)	100	100	100	100	100	100	100	53.16/47.13
Attributes	Diagnosis	Diagnosis-List	Intervention	Intervention-List	Medication	Medication-List	ESC[3]-Classification	Arrhythmia
Completeness (%)	100	96.95	90.84	81.62	95.39	91.55	96.95	13.77

Table 2: Time related characteristics of the composite algorithms and memory storage space used for each Python notebook

Algorithm	**Data Extraction**	**Database Upload**	**Analyse Data**	**Visualize Data**
Time Taken (s): 100 letters	37.92	0.35	NA	
Time Taken (s): 1409 letters	570	0.55	1.92	
Space Taken (KB)	106	6.91	7.5	294

Figures 1: Depiction of (a) Diagnosis and (b) European Society of Cardiology Classification for each patient.



Figures 2: Summary of all medications used across clinic letters

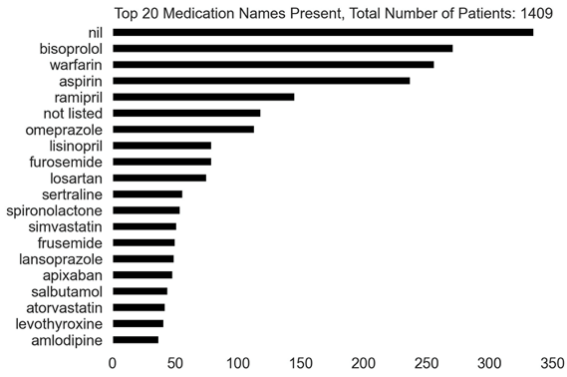
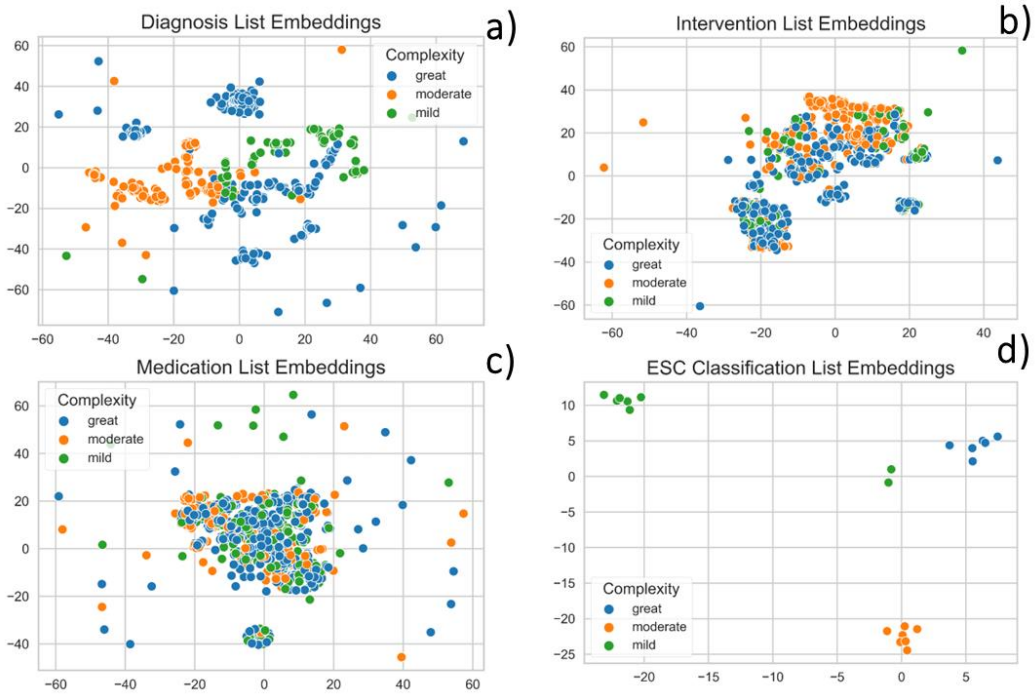


Figure 3: t-SNE dimensionality reduction of diagnosis, intervention, medication, and ESC Classification according to their respective ESC Classification of mild, moderate, and great complexity.

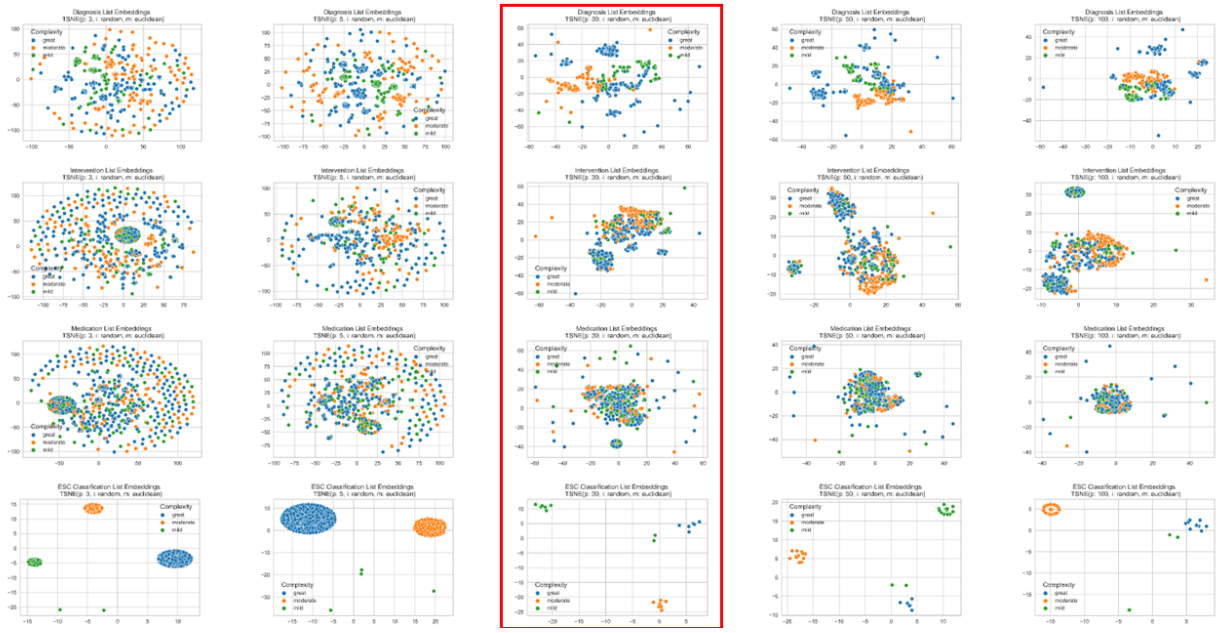

a)
Diagnosis List Embeddings
Complexity
great
moderate
mild
b)
Intervention List Embeddings
Complexity
great
moderate
mild
c)
Medication List Embeddings
Complexity
great
moderate
mild
d)
ESC Classification List Embeddings
Complexity
great
moderate
mild

Supplemental Appendix S1

We used t-SNE in its context of being an iterative non-linear dimensionality reduction technique. For each patient's characteristics a Gaussian distribution describes the similarity with all the other patients. In other words, t-SNE transforms similarity distances between 'patients' into a probability for each pair of patients features (Hinton et al. 2002). If two patients' clinical characteristics are close in high dimensional space then they are going to be assigned with a high probability. The aim of the low dimensional embedding is to approximate these distributions as close as possible with the original, high-dimensional. Mathematically, the similarity between patients x_i and x_j is the conditional probability $p_{i/j}$ with $p_{i/j} = \frac{exp(-|x_i - x_j|^2/2\sigma_i^2)}{\Sigma_{k \neq i} \; exp(-|x_i - x_k|^2/2\sigma_i^2)}$. An important parameter in t-SNE is the perplexity, that relates to the variance of the Gaussian distribution, and it can be interpreted as a smoothness factor of the number of neighbor points. Evidently, these probabilities are generated by quantifying and transforming norms (in this case we adopt the euclidean norm/distance) between data points reflecting the probabilistic similarity between them. It is evident that the changes in perplexity will result in changes in the t-SNE visualisation. Since these visualisations are interactive the user will adjust them manually. In Figure S1, we show how perplexity affects the t-SNE results. We also highlight with a red rectangle the results we choose to demonstrate in Figure 3.

Furthermore, we performed k-means clustering to compare the data-driven clusters with the default grouping based on disease complexity, Figure S2. We selected the optimum number of clusters based on 100 bootstrap repetitions of t-SNE followed by k-means. The agreement between the k-means clusters and the disease complexity grouping was estimated based on the Jaccard similarity index (JSI), which is the ratio of the intersection of two clusters, divided by their union. JSI gets values from 0 to 1 with higher values reflecting higher similarity between two clusters. The average JSI across 100 bootstrap repetitions was 0.28, 0.2, 0.14 and 0.97 for diagnosis, intervention, medication and ESC classification, respectively. This is in agreement with the qualitative results shown in Figure S2.

Figures S1: Selection of t-SNE initialisation parameters. Each column corresponds to a perplexity parameter of 3, 5, 30, 50 and 100, respectively. Each row corresponds to diagnosis, intervention, medication and ESC classification embeddings, respectively. We selected perplexity equal to 30 because this resulted in the most 'balanced' embeddings with respect to the disease complexity index.



Figures S2: Clustering Results. First row shows the stability score of k-means clustering based on 100 bootstrap repetitions. The middle row shows the low-dimensional embeddings colored based on the data-driven (k-means) clustering. The bottom row shows the same low-dimensional embeddings coloured according to the disease complexity.