Digital Analytics and Research Team - Innovation Branch
Annual Report | December 2022

Jonny Pearson
Lead Data Scientist
jonathanpearson@nhs.net

Dan Schofield
Senior Data Scientist
daniel.schofield1@nhs.net

Paul Carroll
Senior Data Scientist
paul.carroll9@nhs.net

Kevin Fasusi
Senior Developer
kevin.fasusi@nhs.net

Martina Fonseca
Senior Analyst
martina.fonseca@nhs.net

1 Summary

This report focuses on the learning for NHS England and the wider NHS from the research generated in the digital analysis and research team innovation branch.

The Digital Analytics and Research Team (DART) innovation branch aims to increase the value we can obtain from our current data through demonstration of data science techniques and approaches. During 2022, the DART innovation branch delivered and supported:

- Seven PhD Data Science Intern Projects covering single-table tabular synthetic data generation, creation of combined text and image embeddings to generate text from images, topic modelling of survey responses, application of hypergraphs to linked comorbidity data, demonstration of applying data science techniques to population health management data, and investigation of the representation space when applying large scale language models
- Two commissions around the "privacy of unstructured data" and a "synthetic adversarial suite" have paved the way to understanding the considerations around curating and sharing data such as synthetic data and unstructured text data for healthcare.
- Open code for travel times, geo-mapping, forecasting, an agent based simulation engine, and a Streamlit allocation tool, all shared in the open through GitHub, to highlight our work and start collaboration discussions
- UK-US Privacy Enhancing Technology Challenge and DHSC Ambulance Hackathon
- Initiated scoping into a whole integrated system dynamics model (WISDM), homomorphic encryption and XR visualisations as new frontiers of our research.

PhD Data Science Intern Scheme

Current PhD students undertaking quantitative disciplines join NHS England on an industrial placement for up to five months, working on a specific project aligned to business need and the students individual experience/knowledge. Currently six interns in two batches of three are recruited each year into permanent band 6 roles.

The Learning section of this report and the appendixes provide project specifics for the listed work above.

The innovation team's work lays the foundations for future augmented intelligence functions in health and care. We have identified key areas required for progress towards augmented intelligence and have a range of projects in support of bridging these gaps.

A: Representation of structured and unstructured data

Why - A key issue for the NHS to extract value from it's data is in how we reduce extremely complex data (e.g. millions of patient letters or multiple years of activity data relating to multiple morbidities) down to simpler representation which can then be used for our models. These representations can then be used for many data science tasks, but each comes with certain limitations which should be understood, as they will lead to different types of information loss, or if used incorrectly could magnify bias present in the data.

Specifics for healthcare - This is especially relevant for healthcare where rare and edge cases need to be dealt with and where even small drops in performance of models can result in indirect harm to patients.

What we are doing - We are investigating this area through three routes:

- Understanding the appropriate application of large scale language models (which attempt to discern how language is used in specific contexts).
- Exploring multi-modal representations of text and images together in order to unlock the potential of learning from linked data in different domains.
- graph representations of data (rather than row/column storage) to capture entity relationships rather than ordered correlations.

Where this is going - The vision for this work is to support data capture and curation so that the NHS has suitable data assets for reliable machine learning to be applied. These assets could feed multiple machine learning models including classification and retrieval models such as 111 online question and answer tooling.

B: Privacy of Unstructured data

Why - To use or share our large and unstructured data for innovation and development we need to consider the sensitive content within the data.

Specifics for healthcare - This often goes beyond private healthcare information (PHI) and includes domain knowledge and context (especially when dealing with sentences/paragraphs). To get the most value out of these data we need to understand the content of the sensitive healthcare components of these data in a more structured way (e.g. comparison against known ontologies) which can be evaluated against a standard.

What we are doing - Our work has focused so far on understanding the problem and currently available options. We are now looking to commission a structured calculation of the risk of disclosure for a free text data set.

Where this is going - Eventually, this work would support a standard way of assessing and curating unstructured text data. This would then be used to collect different data sources relating to how both patient and clinician talk about their healthcare into a text corpus which in turn can be used for training of NHS language and other data science models. The successful application of natural language processing (beyond text analysis) in the NHS depends on these NHS specific training libraries and validations principles. The specific use case for this work is to support secure data environments to be able to ingest, curate and share free text data with appropriate levels of access based on the sensitive content of these data.

C: Synthetic Data and Privacy Enhancing Technologies

Why - Accessing data is always an issue especially when it comes to some of the more sensitive data types e.g. patient pathways. Synthetic data is one solution (although not a silver bullet) which would allow access to a realistic form and content of the data. This could be used to support the development of solutions whilst awaiting data access or for the demonstration of that a proof of concept will work on an end-to-end example. Another solution is the application of privacy enhancing technologies which aim to access the data, without i.e. decrypting, or moving the raw data from source, and thus retaining privacy.

Specifics for healthcare - In healthcare this is even more emphasised through burdensome access processes as well as the data not being collected for secondary analysis and so the business case for sharing is often difficult.

What we are doing - We have been experimenting with how to create as high fidelity (measure of closeness to the ground truth data) data as possible whilst maintaining privacy. Our work has also investigated how to understand the fairness of any generated synthetic data and the explainability of the resultant dataset. This work has taken a few forms but in particular: - the generation of single table tabular data using a variational autoencoder - the generation of patient pathways through an agent based style simulation - explorations of homomorphic encryption for calculating aggregate statistics on two disparate secure data silos. - investigation into federated learning frameworks.

Where this is going - It is envisaged that these tool will be made into production ready code to enable NHS England and wider NHS organisations to create a wide variety of synthetic data and technology enabled options for accessing data privately.

D: Validation of model application

Why - Data science models are designed to find a minimum in a complex dataset to a query. Area A already picks up on the representation of the complex data but the application of the chosen algorithm needs to be investigated and validated for insights to be applied with any confidence. Mostly this comes down to asking, "does the technique applied perform well for all questions in the use-case and is there any induced bias being created"?

Specifics for healthcare - Complex cases (e.g. those interacting with multiple pathways and long stayers) are often the ones we are trying to understand using these models. Whilst the model may state that it's performing well for the majority of patients it can still miss the key dynamics which are behind serious concerns for a minority of patients. Thus, going beyond aggregate or global metrics is needed to properly validate and identify weaknesses in different approaches. Further, as discussed in Area A, there is a need for validation across multiple modalities which commonly occur together in healthcare.

What we are doing - Three projects over the last year highlighted the need for gold standard downstream tasks (and associated metrics and datasets) that would support the validation of the models across different healthcare contexts. We are now looking at a project investigating fairness in our models.

Where this is going - Standard tests and downstream tasks are required for a variety of healthcare specific use-cases. Isolated validation data is needed to run models against without the developer ever seeing this data. The combination of held-out data and suitable use-cases would allow for robust validation of models, and if further coupled with open task definitions and NHS specific datasets would lead to more transparent and useful benchmarks. This work needs to support the AI Lab and MHRA efforts in this area.

E: Reproducible analysis

Why - A significant amount of energy and time is lost to duplication of analysis across the NHS. Reproducible analytical pipelines (RAP) are a suggested standard that would allow advanced analytics to be reapplied in consistent ways across different localities and situations.

Specifics for healthcare - NHS Digital has defined levels of RAP which we encourage all analytical teams to be considering. Building reproducible components and pipelines which are easily shared would support (commonly repeated) small scale tasks such as mapping functions, text analysis, and links to clinical coding groupings, as well as large scale models for optimisation and forecasting.

What we are doing - All of our work is being pushed into the open and produced in ways adhering to RAP and good development practices. We have a particular focus on finding collaborations for our data science interns to work with applying open and shareable code in different datasets and infrastructures

Where this is going - A suite of example code which can be reused by out interns or other trust/ICS analysis. Further a culture shift to model open and RAP workflows would be supported increasing knowledge sharing and interoperable analysis.

2 Background

2.1 Team Setup

Alongside supporting the wider work of the Digital Analytics and Research Team, the innovation branch aims to work as a research and development team for secondary use of data in the NHS. This includes working in a high risk way on exploratory work but also needs a strong connection to business need and a defined research to operations/policy pipeline.

The team is made up of three lead/senior data scientists, a senior developer and a senior analyst alongside three data science intern posts.


(a) Jonny Pearson


(b) Dan Schofield


(c) Paul Carroll


(d) Kevin Fasusi


(e) Martina Fonseca

The key aim of the innovation branch are to increase the value that we can get from our current data through applying data science and advanced analytical techniques in an open and transparent way. We focus our outputs not only on NHS England but also to support the wider trust and ICS analysis especially in the area of population health management.

2.2 Research Narrative

There is data in the NHS which is being underused due to sensitive access restrictions and complex storage structures resulting in lower value and evidence for patient care. These data include both structured and unstructured source such as population health management linked datasets, free text data (e.g. patient feedback, clinical letters, clinical papers), patient pathways and siloed electronic health records.

To increase the value we can obtain from these data we have developed our research programme to cover four areas:

- Increasing access to sensitive data – through either the development and evaluation of synthetic data or by applying Privacy Enhancing Technologies (PETs);
- Creation and use of validation and training libraries – through curating data sources together to making training corpora and by exploring options for identifying fairness and explainability of our models;
- Representation of complex data – through graph structures and understanding lower dimensional embeddings of complex data (including combined embeddings for multi-modal data);
- Insight Extraction – through text analysis, forecasting techniques as well as investigating system modelling and simulation methods.

Many of these topics need considerable investigation and understanding to adequately apply the techniques in a standardised way for the NHS and thus we are looking at three to five year timelines. To keep momentum and show progress during this time we aim to create frequent shareable outputs from individual short-term projects.

The majority of our research work is driven through the data science internship scheme aimed at current PhD students undertaking quantitative disciplines and looking for an industrial placement. This scheme then allows us to connect to academia and learn from some of the cutting-edge techniques whilst having focused projects demonstrating these techniques on our data in short-time scales. See the internship page on the transformation directorate site for more details.

2.3 Roadmap

Waves four and five of the data science internship scheme are currently being recruited to which will lead to an expected additional six projects to the programme. These projects include:

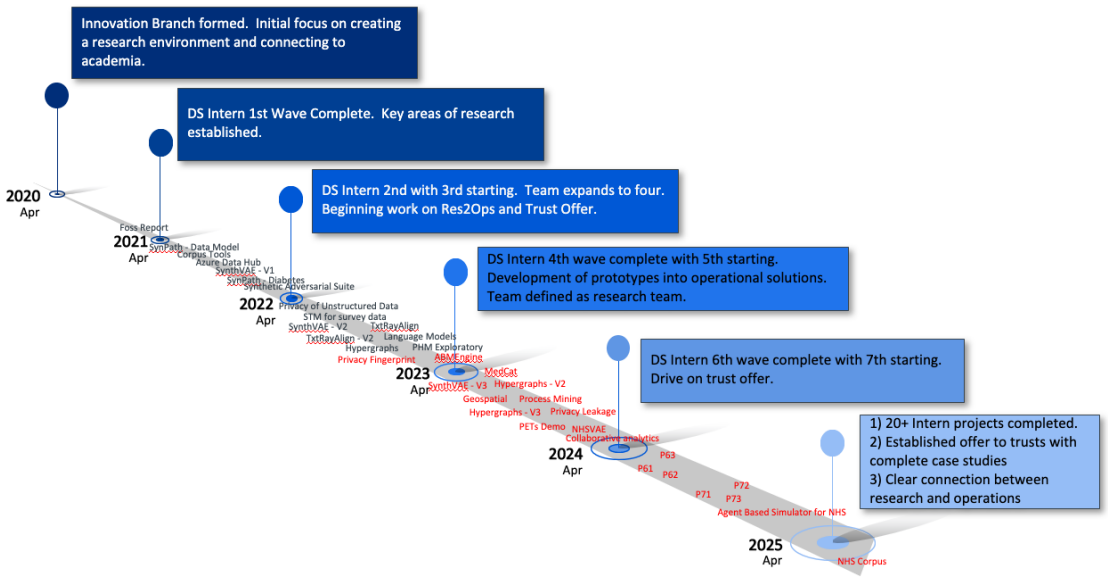

Figure 2: two years into five year road map

- SynthVAE - Production, causal model inclusion and multi-table
- Hypergraphs - Inclusion of demographics to comorbidity studies
- NLP applied to neurology letters - enriching clinical coding using MedCAT
- Graph Neural Networks - extension to our hypergraphs work
- Process mining - generating healthcare pathways from event logs
- Understanding privacy leakage from NLP models

The plan is then to continue an iterations of this scheme through 2024/25 resulting in 20+ intern projects completed and connected to operations. During this time we will need to consider options for expanding the scheme if desired.

There will be a push towards supporting and prioritising trust and local data collaborations through embedding data science colleagues in honorary contracts to apply reproducible analytical pipelines to their data and build knowledge of the application and value of this approach.

We are working on our research to operations route in order to establish appropriate development of these projects into tangible tools integrated into national platforms.

In particular we aim to produce some key outputs, these include:

- **Privacy Fingerprint** - a quantitative calculation of the private and sensitive content of a database of text data beyond personal healthcare information
- **NHS VAE** - production level code able to create high fidelity synthetic data for internal use with clear understanding of the fidelity, privacy and fairness of the generated asset.
- **Agent Based Simulations for NHS** - efficient reusable foundations with range of example implementations of individual level simulations.
- **NHS Corpus** - Framework and foundations for a text data store for the purpose of training and validating text and multi-modal models in the NHS.
- A suite of standalone code for data science in ICS/trust contexts to support population health management.

3 Learning

This section runs through our various projects developed over the course of the year and pulls out detailed learning.

Six points of highlighted learning from this year include:

Variational Autoencoders are suitable architectures for generating tabular single-table high fidelity data. Further work is ongoing to finalise our approach to addressing privacy, fidelity, fairness, and multi-table outputs. We are actively working to turn this into a integrated tool for NHS England. Note that this work would sit alongside the NHS Digital Artificial data solution [1] and the CRPD Bayesian approach [2] to generating publicly available synthetic data.

The NHS needs a clear understanding of how to connect the sensitive/private content in a dataset of unstructured text (beyond protected health information and into considerations around domain and context) to a risk assessment and mitigation actions. This work aims to continue to build a quantitative calculation of a privacy fingerprint of any dataset of unstructured text.

Applications of (un-)directed hypergraph structures to population health data has allowed complex directional relationships to be demonstrated for different groups of multi-morbidities, going beyond pairwise measures commonly studied in applications of simple graphs to comorbidities. This work sets the foundations for further exploration of alternative representations of healthcare data, which can model more of the inherent structures, and the application of (hyper)graph neural networks as an efficient approach to insight extraction.

Embedding spaces created by language models applied to healthcare data are very dependent on the specific model used and how they are trained. If the creation of the embedding space is not well understood, then there is significant risk of undue harm caused by these models. We have shown the performance increase versus computational cost of using fine-tuned models rather than frozen off-the-shelf ones. Overall it needs to be well understood that to use these models in the NHS appropriately, requires both suitable input data and computational resources to train them to a properly representative level.

Whilst using contrastive learning approaches to build shared representations of Chest X-ray images and their associated text notes can be exploited to produce reasonably meaningful sentences for unseen images, the performance falls short of acceptable levels for smaller model architectures. Similar to the language models work above this demonstrates the need for significant computation expense to create models with high enough performances for wide spread use, as well as the need for input data that captures the local variance seen within the NHS.

Our work with Trust data through an honorary contract has shown the value of data science methods to extract additional insight from their data. The developed code bases are being designed with reuse in mind and the aim is to reapply them to new data and datasets at the same and different trusts/ICSs.

More specific learning from each of our projects is set out below in the context of our research narrative described in 2.2. For further details about the project specifics and outputs then see 4.8 and 4.8.

3.1 Increasing access to data

Whilst supporting the development, testing and use of secure data environments in NHS England we focus our own access research on the use of privacy enhancing technologies (PETs) to increase the options for accessing data. The term PETs covers a large range of different technologies and we recommend reading the PETs adoption guide [3] from the CDEI and the recently published chapter 5 of the ICO consultation [4] for background material in this area.

This year we have continued our focus on synthetic data generation and evaluation but also extended our interests to include Homomorphic Encryption as well as monitoring a US-UK PETs challenge around Federated Learning.

3.1.1 SynthVAE

We have shown that Variational Autoencoders are an effective architecture for generating synthetic single-table tabular data. Our first project in this area demonstrated how to accommodate differential privacy (DP) [5] into the model. We observed that restrictive privacy budgets led to reduced synthetic dataset quality as expected however, the measure of privacy provided by DP did not always correlate directly with the observed level of privacy according to other privacy metrics. This led to a question around if DP is necessary at all or whether the noise inherent to a variational encoder is enough to disguise the real data.

Variational AutoEncoders (VAEs)

To reduce high-dimensional complex data down to a manageable representation we use an encoder/decoder pair trained so that the reconstruction loss (the error in recreating the original data from the simple representation) is minimised. This is called an autoencoder and has the aim of producing a good representation of the data but in a reduced number of dimensions.
The variational autoencoder is an autoencoder that has been regularised to avoid overfitting and ensure that the latent space has good properties that enable a generative process. This regularisation is enabled by encoding the raw data as a distribution rather than a set of single points. This means that when new data is generated is has been trained to have a higher probability of recreating likley events.[6]

3.1.2 Next Steps

Regarding the synthVAE work we plan to run one more piece of development work and then split the project into two; one to continue transparent exploration of methods; and one to develop the tool into a closed usable product for NHS England platforms. More work is needed to understand fairness in downstream task performance and a wider discussion around privacy and fairness is needed to ensure utility of the model in all potential NHS use cases. A comparison of differential privacy in alternative generator methods (e.g. GAN) to benchmark our VAE would be useful but so would a comparison of Private Aggregation of Teacher Ensembles (PATE) [7] vs DP to explore if there is a better approach to privacy that current. The Interpretability of VAEs could also be explored (as this is a key reason for choosing this architecture).

Homomorphic Encryption has potential to enable analysis and possibly linkage of siloed data without the requester ever seeing the raw data. Whilst the impact of this could be significant for remote analysis the implementation needs to be proven in a step wise fashion. We are aiming to work with the Office for National Statistics on a series of small incremental prototypes to demonstrate the feasibility and potential of this technique.

Homomorphic Encryption

Homomorphic encryption enables computation directly on encrypted data. Whereas traditional encryption schemes facilitate the encryption of data in transit and at rest, homomorphic encryption schemes additionally facilitate encryption in process. This means that data can be sent securely, acted upon and the results of that action sent back to the data owner to be then decrypted, without anyone but the data owner ever seeing the data or the results of the computation. [8]

The bilateral privacy enhancing technology challenge has announced the short list of chosen proposals for the challenge. These include multiple groups proposing to setup up federated learning with a secure multi-party computation alongside differential privacy. We are both assessors and observers for this work which will continue through to sprint 2023[9]. We will continue to monitor the UK-US PETs challenge and aim to address setting out a road map of how to turn the resulting learning and code into a useable tool for the NHS.

Federated Learning

The execution of training a model on distributed datasets without the data ever leaving it's home location. This area includes a variety of approaches to sharing the learning (e.g. sharing model weights) between the distributed datasets. [8]

Secure multi-party computation

When data is fragmented over multiple silos so that no single silo has knowledge of the full record. When a function is applied to the data then each silo computes the function on their share and then aggregates the answers to give the solution without any of the individual data being seen
e.g. the value 15 could be split into two values 17, -2 and then distributed across two nodes. To Double the value each node doubles their share into 34 and -4, and then sends the answers to a separate silo to aggregate together to give the answer of 30 [8]

Differential Privacy

A formal definition of privacy requiring that the output of any statistical analysis reveals no information specific to an individual in the dataset. This means that an individual asking for an output from the data would not be able to see a statistically significant difference if they included/excluded an individual from the data request and therefore cannot infer information about this individual, protecting their privacy. [8]

3.2 Validation and Training Datasets/Libraries

For the NHS to be confident in it's application of AI we need detailed understanding of the limitations and performance of our models. For data science models, this requires high quality training data suitable to different NHS use-cases including cases which require broad training data (e.g. from patient feedback, clinical notes and academic papers to understand how different users of NHS talk about health) and focused training data (e.g. focused data on a clinical condition with coverage for edge cases allowing high accuracy in tagging or classification). Accurate and understandable model validation needs to be available alongside clinical verification. There needs to be a balance of the technique applied and the explaianbility of the model and outputs. Further, there is a need to be able to independently compare models through more widely shareable benchmarking tasks to give extra confidence to perceived performance on local data.

In light of this we have started to investigate the curation of unstructured text data for the purpose of building an NHS specific corpus of training text data. The main hurdle for bringing this data together and allowing models to train on it is to understand the privacy content of the data and if any models trained on the data may leak privacy of the training data if shared openly.

Additionally, we are looking at what explanability techniques are most useful for NHS predictions and classification tasks.

3.2.1 Privacy of Unstructured Data

A report on the privacy of unstructured data [10] was generated off the back of a expert workshop in February 2022. This report highlighted the considerations around:

- Sharing unstructured text data
- The high level anonymisation process
- Considerations when assessing the re-identification risk of unstructured text
- Different anonymisation levels
- Available anonymisation tools
- Challenges to anonymisation
- Case studies of anonymisation of unstructured text data

The report summarised with a list of key qualities for a tool aiming to identify and deal with privacy of unstructured text data. These qualities included:

- structuring and data handling;
- tool use and validation;
- context;
- flexibility.

The report didn't deal sufficiently with aspects of how to structure the data and how to connect the privacy considerations with a risk assessment and mitigation actions. These aspects are now being considered for our current work in this area.

3.2.2 Model Class Reliance

Model class reliance is a novel technique to understand variable importance through perturbations of a machine learning model (regression or random forest). This technique was applied successfully and demonstrated the predictive power of including commercial data into a respiratory deaths models.

Model Class Reliance

A generalised calculation of the permutation importance from a predictive model (regression or random forest) which covers a class of models. [11]

3.2.3 LIME

Local Interpretable Model-Agnostic Explanations (LIME) were applied to a facial image prediction model for identifying Rosacea. The key limitation is that the superpixels constructed as an input for LIME to be applied are too granular for facial images.

LIME

A technique that aims to explain the predictions of any classifier, by learning an interpretable model locally around the prediction. In practice the technique splits the input into smaller chunks and then permutates masking these chunks in the input data and recording the impact on the output prediction accuracy. This then gives a broad and understandable indication of which parts of the input are most important for the predicitons.

3.2.4 NHS Corpus tooling

Our work on investigating the toolset required to ingest and curate healthcare data into an NHS Corpus has suffered from a lack of knowledge of the considerations that this tooling would need to address (see 3.2.1). However, whilst the toolset itself is not in a usable state, the learning from the project is of interest.

3.2.5 Next Steps

For the privacy work there are two future steps:

- Round table discussion to continue the conversation and bring a collective expert voice to state the need for this work in healthcare
- An additional commission called "Privacy Fingerprint" which would focus on the structuring and qualitative calculation of a privacy score for a block or data base of text.

Model class reliance should be considered in other settings when applying a regression or RF prediction model. there is also a possibility to extend our collaboration with the CIVIC II project depending on funding of round two.

The NHS Corpus tooling needs to be made into a static learnable state to inform the privacy fingerprint outcomes.

More generally we are considering further work on explainability, AI safety and fairness through exploratory projects and collaborations with the AI Lab.

3.3 Representation of complex data

When dealing with complex data we often need to translate a high number of dimensions to a much lower number which inevitably loses some information. We need to ensure we understand what has been lost and what is retained in order to use our data effectively. This section of research focuses on the embeddings of complex unstructured data and the opportunities that graph based representations have for applying efficient and flexible methods or representing these complex data.

Embedding

Embeddings are learned, low-dimension, continuous representations of high-dimensional complex categorical variables such as words in a document. Embedding are created in order for machine learning models to create meaningful clusters, inputs for supervised tasks and visualisations. There are a significant number of techniques to create embeddings as the the formation of these structures is often more important than the model applied.

3.3.1 Txt-Ray Align

Successfully created exploratory work for generating radiology reports from chest x-rays. TxtRayAlign exploits contrastive training to learn similarities between text and images, allowing a retrieval-based mechanism to find reports that are "similar" to an image. The results of our investigation indicate that smaller scale pre-trained models can be tuned to help retrieve clinically meaningful sentences, combining them into larger generated content, yet falls short in achieving this with sufficient accuracy. While improvements to the models could be made, our findings are corroborated by others in literature, where improvement in performance is often achieved by increasing the scale of model (orders of magnitudes larger in terms of number of trainable parameters).

Contrastive Learning

Data points are contrasted against each other to teach a model which points are similar and which are different. Those belonging to the same distribution are pushed towards each other in the embedding space whilst those belonging to different distributions are pulled away from each other.

Three potential uses for natural language generation models in the clinical radiological workflow have been identified;

1. use as a safety-net for radiologists to auto-fill positive findings if not included in the report by the radiologist;
2. provide preliminary reports for acute CXRs to support junior doctors interpreting scans on the wards in the first instance whilst awaiting the radiologists report communicating critical findings;
3. automate follow up oncology scans, e.g. CT, reporting to provide a faster indication if a malignancy has progressed / quantifying response to therapy.

A range of common machine translation metrics (BLEU, ROUGE, Meteor), more focused clinical metrics (cheXpert, Mirqi), and the combinations of these alongside a clinical scoring system have been investigated for the evaluation of the generated language. The exact combination to produce a final output metric is discussed in relation to the downstream task.

3.3.2 Large Language model applied to Patient Safety Data

We trained and applied a RoBERTa based model [16] on a large subset of the incident reports data. Subsequently we introduced a contrastive loss objective to the pre-training stage using the DeCLUTR approach [12]. In order to investigate the changes in the embedding space produced by these different models we derived pseudo classification tasks, CheckList style test suites, and embedding similarity comparisons presenting aspects of the results in each case.

It is clear, though not surprising, that differently trained language or NLP models produce distinct embedding space structures. These embedding spaces are the crucial component to any downstream task and this is highlighted in our work using a number of probing methods and pseudo classification tasks. Another finding of interest is the usefulness and re-usability of language model embeddings without the need to fine-tune entire models for different downstream tasks. Whilst the performance in a frozen model did not match that of the full fine-tuned setting there are promising approaches which can better utilise frozen language models, such as prompt learning and parameter efficient fine-tuning.

3.3.3 Hypergraphs for Comorbidities

We have successfully shown that we can understand the observed progression of multi-morbidity in a population through the lens of morbidity prevalence. The morbidity progression and ordering can be included using centrality-based algorithms such as PageRank. The importance of the disease progression can then be highlighted using Eigenvalue centrality. We found that finality (mortality) is required when considering disease progression. We finally demonstrated that demographics such as age, sex and deprivation can be included when considering the impact of multi-morbidites.

Overall this demonstrates a powerful and flexible model for complex data sources to help inform policy and strategy decisions.

Hypergraphs

Graphs are a representation of data entities (e.g. patients, postcodes, disease codes) connected by edges representing their relationships (an edge between an individual patient and a certain postcode shows there is a relationship between these data elements.
Hypergraphs allow higher order relationships than pairwise edges. This means a entity can be connected to another entity whilst maintaining knowledge of all other entities they are connected to. This is helpful when trying to understand the combined and individual impact of confounding variables such as comorbidities.)

PageRank

PageRank was designed to rank webpages. Essentially it works by counting the number and quality of links to an entity in a graph to determine a rough estimate of how important the entity is.

Eigenvalue centrality

Eigenvector Centrality is an algorithm that measures the transitive influence of nodes. Relationships originating from high-scoring nodes contribute more to the score of a node than connections from low-scoring nodes. A high eigenvector score means that a node is connected to many nodes who themselves have high scores. This allows the impact of any entity in our graph to be assessed.

3.3.4 Next Steps

For our work into text and image dual embeddings we would like to consider generating the impression section of the report only, use both directional views of the chest x-ray if available together to generate the report and provide a variable length report to avoid repetition and the inclusion of erroneous information if a normal exam, as well as only include the key clinical diagnosis if a positive exam. In addition as this project focused on further investigating the retrieval based approach, future studies using an encoder-decoder model for a generation based approach to allow for the comparison of performance would be of interest. Further looking into how this sort of multi-modal approach can be extended to include further modalities in a shared space would be a direction of significant interest.

There are a wealth of alternative language models in the open which could be also explored, some of which are designed for biomedical texts such as BioBERT or clinicalBERT, knowledge enhanced transformers, and many more variations. The CheckList framework provided a valuable tool for proving and testing different aspects of a NLP model, and only a limited number were explored in this project. Future work could develop a much broader range of tests at a greater scale. Moreover, the integration of embedding cosine similarity expectation functions into the CheckList protocol would be desirable. With the help of easy to use frameworks such as Bulk one could attempt to label a sample of reports to provide a gold standard set of topic labels to allow more sophisticated approaches. We aim to continue our work on language models to investigate these alternatives and how to efficient use and evaluate pre-trained language models in the NHS.

We also aim to continue our work with SAIL for another one or two years access. Our main focus will continue to be the utilisation of hypergraphs to model healthcare data, and build off the first project. In particular, we would like to increase the investigation into adding demographic factors directly into the graph, and to investigate how to include episodic conditions into the graph. There is also a possibility to consider applying graph-based neural network approaches to the generated hypergraphs.

3.4 Insight Extraction

Once the data is accessible, validated and well understood then we can start to consider extracting new insights from the data.

3.4.1 Using population health data to investigate inequalities in Diabetes

The value of linking healthcare activity and population health data is clear but often underused. This work demonstrated how a suite of techniques could be applied to a linked population health data set to investigate inequalities in diabetes pathways. A key learning point was the need for primary care data, this tool a while to obtain in the project and essentially meant our intern had to start again when it finally arrived. There are many quick wins around geospatial visualisations of data and network analysis of disease as well as single objective prediction models.

3.4.2 Forecasting

In order to support of forecasting offer, we have developed a base forecasting tool which can be applied quickly without much development. This is designed to be applied at the beginning of a forecasting project to assess if there is worth in continuing the forecasting on the data and to highlight any data quality issues which will dramatically impact a larger forecasting project.

3.4.3 Simulation

Our SynPath diabetes project investigated how to apply the simulation model to a diabetes use case and started to explore the different optimisation models that could be applied to patient pathways. Q-learning/Q-Networks stood out as the main avenue to investigate going forwards.

Q-Learning

Reinforcement learning allows an algorithm to quickly reach an optimal state usually by rewarding successful actions or better outcomes. This requires a policy to define what success looks like.
Q-learning is a form of reinforcement learning that learns from the current state of an agent in the model by predicting the outcome of a random next step (instead of using a reward system). This allows it to learn when no defined success policy is available.

3.4.4 Text analysis

Through our Internship project investigating applying structural topic modelling to survey data we have developed a reusable topic modelling code and made this open. Through this development we have learnt the importance of incorporating metadata alongside the text contents in order to support the topic modelling produced.

3.4.5 Geospatial

The currently available SHAPE tool [12] developed by Public health England is excellent and should be well considered for these tasks. However, this tool does not enable the user to upload their own data alongside the tools data which limits the use significantly. We have explored developing python code for calculating travel times for a single and multi point scenario (e.g. 30 min driving distance from any GP in an local authority). From this the efficiency of the underlying geospatial structure has been highlighted as needed significant consideration otherwise any tool developed is difficult and expensive to scale.

3.4.6 Open Coding

This year we have increased our push to open coding and collaborative analysis through creating a open code and RAP oversight group in order to address chapter 2 of the Goldacre report. We have also learnt the importance of framing this

area along reproducible analytical pipelines in order to bring in the excellent current work from NHS Digital and bring the whole area closer to business as usual analysis. Significant challenges still exist in the closed working culture and IT software to enable sharing across the NHS.

3.4.7 Next Steps

The value of the work undertaken in the population health work with ESNEFT appears to be high. We therefore aim to continue this work into a second project as well as starting to think through how to produce starter code for applying a similar approach to other data sets.

To generate usable patient pathways an approach is required that can create pathways from assumptions (rather than data) to preserve privacy but this needs to be at a high enough complexity to represent the ground truth data sufficiently. An agent based approach using pre-defined pathways but with the choices dependent on the individual simulated patients rather than through the traditional approach of sampling from probabilities. Our work in this area has demonstrated many of the required components and investigated how to include intelligence in these models for what if scenario testing but the tooling remains computationally inefficient at present. We are now working on an engine that would be able to run these simulations quickly for multiple interacting patients whilst also acting as a standardised open framework for building simulations from. This would then support sharing and bench marking of these models.

Synthea™[13][14] project gives a good marker to be built off for how pathways could be developed transparently and with clinicians leading the process. However, there is a significant amount of work to develop an efficient, scalable and reuseable interaction with an engine which could simulate standardised agents (patients) flowing through these defined pathways within a healthcare environment include options for optimisation and learning within the model. Our current work is investigating building the base engine in the GO language. We are also investigating the right intelligence approach with a current lean towards Q-learning over Monte-Carlo or reinforcement learning WHY??

We are now looking at developing a suite of statistical tests for comparing and bench marking different forecasting methods in order to test the value obtained by any forecasting project

Our structural modelling work needs to be compared with other clustering or hierarchical topic modelling techniques in an ensemble approach in order to get the most robust generalised code. An additional avenue of discovery would be to investigate an ensemble approach to sentiment analysis although we feel that the largest opportunity for sentiment analysis is in the underlying training library rather than the applied techniques.

Working with Google Data Engineers we have started to explore issues around the granular and hierarchical structures in any geospatial work as well as the LSOA mapping to different useful datasets, the best algorithm for calculating multiple travel routes simultaneously, and a user interface wrapper to go around the developed library.

A standardised approach to system dynamics in the NHS is required to support whole system modelling and understanding of the wider implications of interventions. Our work has started to identify possible solutions and partners to build interoperable modules using an underlying standard from OASISopen [15] alongside a range of open source and proprietary software to support functionality.

3.5 Code Plan

The main outputs from the work are technical reports and established code bases. These code bases can become stagnant if not maintained overtime and so it's important to have a strategy and plan for how to maintain and use each piece of code.

We consider our coded outputs in three levels:

- **Prototypes** - these pieces of code are delivered as working examples of a method or tool set but then not continually maintained. Their main purpose is to backup the technical report with shared code and clear examples.
- **Standalone** - these pieces of code are designed for reuse by a developer/analyst. They will need tweaking for the local situation and should only be applied with domain/data specific knowledge to ensure they are not misused. These pieces would be used by any ICS/trust project as a suite of possible starting points to apply a range of data science techniques. These code bases need a code owner to monitor their status and ensure they remain active and updated over time.
- **Integrated** - these pieces of code have an eventual aim to turn into a tool that can be integrated alongside NHS England infrastructure and data to create new capabilities for our analysts. These will need maintenance to keep them active but more importantly a full software engineering cycle including requirements, full refactor to reach an Alpha point and then a full testing programme to move through Beta and into release. This is beyond the current capabilities of the team and so requires support from DMIS or additional resourcing.

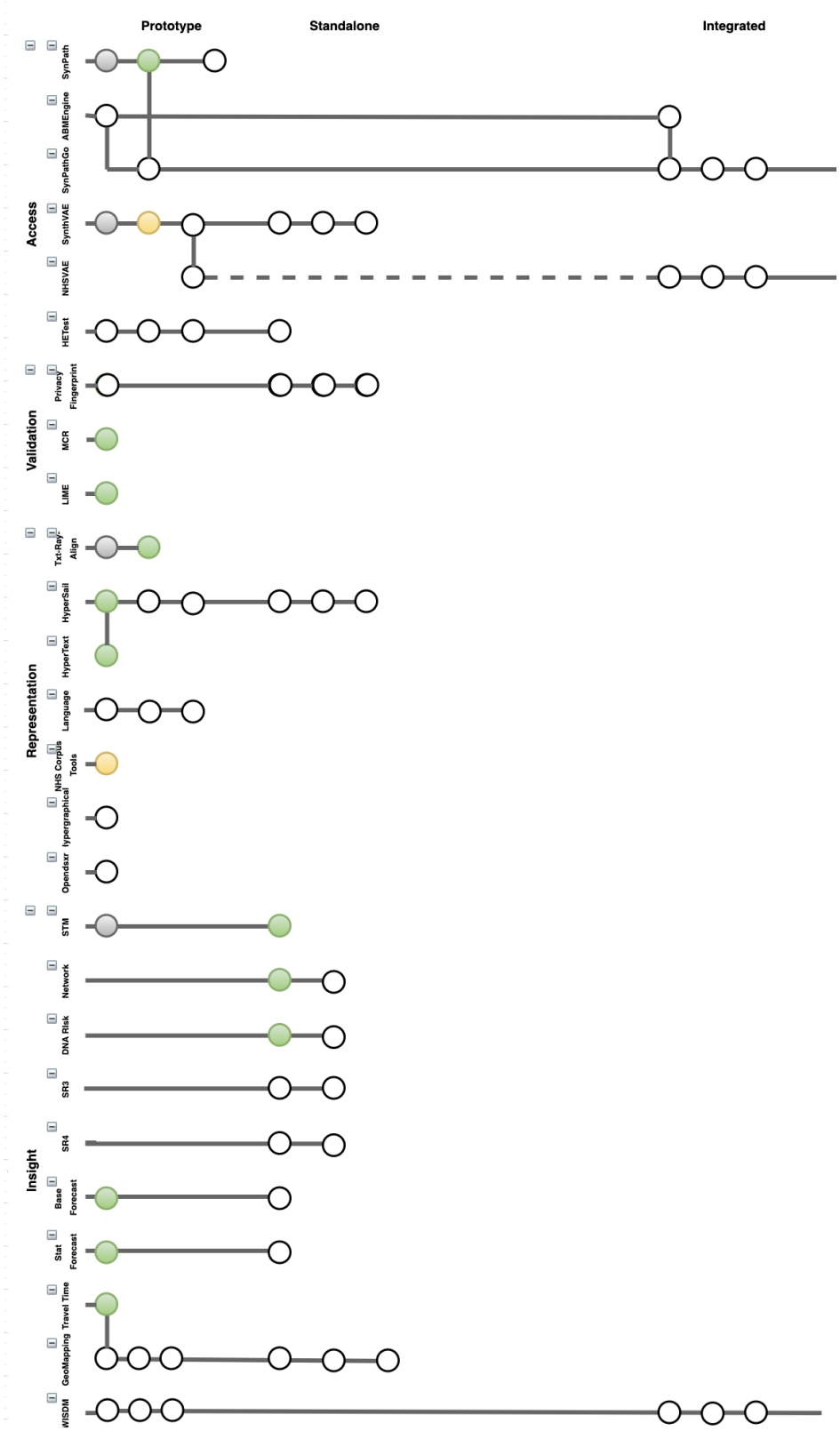

Figure 3: Planned development of current code bases. Green/yellow indicate current status (with yellow highlighting an issue that needs fixing), grey is passed releases, white represents future releases.

4 On-going Opportunities

There are significant benefits to expanding this research function in particular in supporting the route from research to operations/policy and expanding the number of collaborations with trusts and ICSs to develop reproducible analytical pipelines.

Specific suggestions for expanding this work include:

4.1 Option A: Expansion of the internship scheme

An increase to both the number of supervisors (Senior Data Scientists) and interns. Each Senior Data Scientist can manage up to three students per year. Each year we have at least double the number of good applicants than we can accept.

Example:

One more senior data scientist and an additional two intern positions

Pros:

- More high value, high return on investment projects

Cons:

- A large programme may water down the value of the current of each individual project as knowledge becomes to spread out and less focus is given to individual project impact - Increased management burden

4.2 Option B: Support and fund full PhD projects

DART could look to collaborate with University Groups of earlier interns to (partially) fund future PhD projects. This could be extensions of projects that have been carried out during the internship that have enough 'future directions' to call for a lengthier study, or proposed projects by the team or group.

Example:

Extending the SAIL Databank Hypergraphs work into further research which makes up a full three-to-four-year PhD project.

Pros:

- Continued in-depth work on a project of interest to DART

- Closer collaboration with specific University groups

- Further in-depth knowledge transfer into DART

Cons:

- High costs associated with supporting a full PhD project

- Risk of churn within DART which may happen at a higher frequency than the length of a PhD

- Sits significantly outside of the original idea for the PhD Internship program

4.3 Option C: Continue collaboration with internship student on internship project

Description:

DART could look to continue working with an intern when they return to their PhD study if the contents of the internship project had significant alignment to their PhD topic, and there was buy in from the supervisors. This may also be needed in cases where the project leads to a publication that is unlikely to be completed within the initial timeframe of the internship.

Example:

Extending the work on Patient Safety Language Modelling due to similarities with work being carried out in Niall's PhD project (mental health free text note focus with CRIS data)

Pros:

- If supported by the department, this could give continued knowledge sharing in both directions

- Possible route to extend projects without the added cost of partially funding a full PhD program

Cons:

- Ambiguity on how this would work contractually (honorary contract with NHS England or the University needed?)

- IP considerations could be challenging

- Would require support of supervisors and would need to fit alongside agreed timelines on both sides

4.4 Option D: Continue collaboration with university group on related project

Like earlier choice but instead the collaboration is on a related problem to the PhD internship but is not a direct extension or spin-off. This would have to be agreed with the group and would need to align to DART team member's current knowledge or interests.

Example:

Continued support on Niall's PhD through informal/formal agreement with the Chronosig project (or other)

Pros:

- Learning for DART team member on an area of research outside of usual projects

- May lead to future collaborations beyond the options described here

Cons:

- As above, would require stable commitment from the DART team members if more formally agreed (otherwise would need a significant handover)

- IP issues again could prove problematic

4.5 Option E: Temporary/fixed-term role with NHS England

Interns can continue to work on a DART internship project after they complete their PhD studies (or through another pause of studies) if they secure funding from another route, or we can fund further work. This would be supported in an equivalent way to the internship with a given DART supervisor for the project time.

Example:

One of our previous students has returned to continue work on SynPath, although here she has brought her own funding, and thus might not fit the usual model of requiring a (fixed term) role to be available in NHS England

Pros:

- Allow us to continue to gain the benefits of working with interns that have developed an in-depth knowledge of the project area

- Interns will bring further learning from their continued studies

Cons:

- In the current climate, it seems unlikely that these types of roles will be easy to get sign-off for

- Academic funding routes which allow students to continue their studies but work with industry are not always clear and internal extra funding may be challenging to arrange

4.6 Option F: Creation of a internship scheme focused on Master's students

To increase our support to trust/ICS analytics, a small group (2 or 3) of Master students aiming to apply established techniques to different locations and contexts (i.e. applying using the code generated by the PhD internship scheme in different geographies). This year we had 50+ Master students applying to our PhD Data science internship.

Example:

One Data Scientist could manage one group of up to three master students to apply a small suite of code bases to trust/ICS data

Pros:

- Front line impact and wider knowledge share of our main research on local data - Increased offer of training and advertising data science in the NHS to Master students.

Cons:

- Additional management burden to ensure the work is being applied appropriately and the outputs are useful

4.7 Option G: Do nothing

Once an internship has ended the project is frozen until picked up by a future intern, and work with the student stops when they leave.

Example:

Some projects will have a natural end point if work has not been successful, or needs/interests change

Pros:

- Can develop a useful framework so projects can be 'mothballed' in a state which would allow for sharing and transparency, but means they don't need any further work (and may be useful across DART)

- DART team members will have time to explore other projects or attend to other work

Cons:

- Possibility of lost knowledge on both the intern and DART side

- Too many 'dead-end' projects may look like a negative on the scheme

4.8 Recommendations

Our recommendation is based on balancing opportunities and the additional burden of admin, funding and quality control.

Option F - Increasing our focus on reusing analysis and research rather than creating more to aim for more impact from each piece of work.

and

Option E (aiming to address the Pros of options C and D) - Establish an offer to work with internship students after the internship has finished including a honorary role with NHS England (to enable them to continue their PhD work). This role may need to be funded. The role would need to be fixed term and establish a deliverable. The supervision and support would need to be minimal with NHS England becoming a customer rather than supervisor. Standard data sharing agreement and interlectual property agreements would needed

References

[1] NHS Digital. Artificial data generator. https://github.com/NHSDigital/artificial-data-generator, November 2022.

[2] CRPD. Uk data driving real-world evidence. https://cprd.com/synthetic-data, November 2022.

[3] Centre for Data Ethic and Innovation. Pets adoption guide. https://cdeiuk.github.io/pets-adoption-guide/, November 2022.

[4] ICO. Chapter 5: Privacy-enhancing technologies (pets). https://ico.org.uk/media/about-the-ico/consultations/4021464/chapter-5-anonymisation-pets.pdf, November 2022.

[5] OpenMinded. Differential privacy. https://blog.openmined.org/tag/differential-privacy/, November 2022.

[6] Joseph Rocca. Understanding variational autoencoders (vaes). https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73, September 2019.

[7] OpenMinded. Build pate differential privacy in pytorch. https://blog.openmined.org/build-pate-differential-privacy-in-pytorch/, November 2022.

[8] Centre for Data Ethics and Innovation. Pets adoption guide. https://cdeiuk.github.io/pets-adoption-guide/, September 2021.

[9] Gov.uk Press Release. Winners announced in first phase of uk-u.s. privacy-enhancing technologies prize challenges. https://www.gov.uk/government/news/winners-announced-in-first-phase-of-uk-us-privacy-enhancing-technologies-prize-challenges, November 2022.

[10] NHS England and Butterfly Data. Privacy of unstructured data. https://github.com/nhsx/PrivacyFingerprint, March 2022.

[11] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019. URL `http://jmlr.org/papers/v20/18-760.html`.

[12] PHE. Strategic health asset planning and evaluation. https://shapeatlas.net/, November 2022.

[13] Jason Walonoski. Synthea. `https://github.com/synthetichealth/synthea`, November 2022.

[14] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, 2017. ISSN 1527-974X. doi:10.1093/jamia/ocx079. URL `https://doi.org/10.1093/jamia/ocx079`.

[15] Oasis. Xml interchange language (xmile) for system dynamics tc. https://www.oasis-open.org, November 2022.

Appendix: Wave 1

P11: SynPath Diabetes

Exploration work into incorporating learning into a pathway simulator for diabetes. This work has fed our current SynPathGo project to create synthetic patient pathways and a foundation for agent based modelling in the NHS.

Intern: Tiyi Morris

Supervisor: Jonny Pearson

Start date: Jun 2021

End Data: Aug 2021

Code: github.com/nhsx/SynPath_Diabetes

Report: github.com/nhsx/.../Technical%20Report%20(SynPath%20Diabetes)%20v1.pdf

P12: SynthVAE

The initial creation of a variational autoencoder with differential privacy for generating single table tabular gaussian data. This work demonstrated the feasibility of this approach for healthcare and fed into further interactions of the code base.

Intern: Dom Danks

Supervisor: Jonny Pearson

Start date: Jul 2021

End Data: Sep 2021

Code: github.com/nhsx/SynthVAE

Report: github.com/nhsx/SynthVAE/blob/main/reports/report.pdf

P13: NHS Text Data Exploration

Work completed as part of a Master's project placement arrangement with Lancaster University. Using a pre-defined toolset this project looked to understand how to ingest NHS.UK text data into a curated form. This work has fed into our current push on privacy of unstructured data as well as defining the needs around the technical tool stack to collate, curate and share large bodies of healthcare text data.

Intern: Beth Rushton-Woods

Supervisor: Dan Schofield

Start date: Jul 2021

End Data: Sep 2021

Code: Available on request

Report: Available on request

P14: Value of Commercial Data (Work Experience Arrangement)

A work experience project investigating the impact of including high-street drug sales data into a prediction model for respiratory deaths. The technique used to establish the value of these data was model class reliance. The project demonstrated the methodological approach and it's use for understanding the impact of specific variables on predictive models. It also showed the value of including these data for the task. The work continues through the CIVIC-II programme.

Intern: Elizabeth Dolan

Supervisor: Jonny Pearson

Start date: Oct 2021

End Data: Dec 2021

Code: github.com/nhsx/commercial-data-healthcare-predictions

Report: github.com/nhsx/.../NHSX%20Report_ValueofCommercialProductSalesDatainHealthcarePrediction_V2.pdf

Appendix: Wave 2

P21: SynthVAE - cont.

A continuation of the development of our variational autoencoder, this project focused on non-Gaussian input data, hypertuning the code and starting to consider how fairness in the created data can be assessed and implemented. This project has fed into future development work of turning this code into a production level tool.

Intern: David Brind

Supervisor: Jonny Pearson

Start date: Jan 2022

End Data: May 2022

Code: github.com/nhsx/SynthVAE

Report: github.com/nhsx/SynthVAE/blob/main/reports/NHSXSynthVAE%20(2).pdf

P22: Automated Text Images

An investigation of extracting insight from multi-modal data using contrastive learning. This work aimed to demonstrate how to go about creating text from images for chest X-rays. This project was directly followed by a second in this area.

Intern: Dekai Zhang

Supervisor: Dan Schofield

Start date: Jan 2022

End Data: May 2022

Code: github.com/nhsx/txt-ray-align

Report: github.com/nhsx/txt-ray-align/blob/main/report/TxtRayAlign_Report_DZ.pdf

P23: STM for NHS Survey Data

the development of an R code for investigating the topics found in free text survey data using a technique that monitors both the content of the responses but also the metadata (e.g. when the response was made, which organisation the response relates to) in order to support the construction of these topics. This is now an established reusable code.

Intern: Anna Linton

Supervisor: Paul Carroll

Start date: Feb 2022

End Data: Apr 2022

Code: github.com/nhsx/stm-survey-text

Report: github.com/nhsx/stm-survey-text/blob/main/reports/report_stm.pdf

P24: LIME for XAI in Facial Disease Classification

A work experience project investigating the application of a generalised explanation technique to an image classification task around identifying Rosacea. The work highlighted the considerations required around balancing granularity of explanations against a fast an intuitive method.

Intern: Anwesha Mohanty

Supervisor: Jonny Pearson | Dan Schofield

Start date: Jan 2022

End Data: May 2022

Code: github.com/nhsx/LIME-XAI-Facial-Disease-Classification

Report: github.com/nhsx/LIME-XAI-Facial-Disease-Classification/blob/main/reports/report_AM.pdf

Appendix: Wave 3

P31: Automated Text Images - cont.

A continuation of previous work in this area, this project sought to identify the clinical application, pipeline and validation metrics for this work. This work established how to take a interesting method and apply it in healthcare in a valid way.

Intern: Sarah Hickman

Supervisor: Dan Schofield

Start date: June 2022

End Data: August 2022

Code: github.com/nhsx/txt-ray-align

Report: github.com/nhsx/txt-ray-align/blob/main/report/TxtRayAlign_Report2_SH.pdf

P32: Predicting the Impact of Health Inequalities - Diabetes

A collaborations with East Suffolk and North Essex foundation trust to apply a suite of data science techniques to a large population health data including both primary and secondary care data. The aim of the project was to identify inequalities in diabetes care whilst making reusable code bases which can now be applied for different conditions and in different organisations.

Intern: Stephen Richer

Supervisor: Paul Carroll

Start date: Jun 2022

End Data: Dec 2022

Code:

1. github.com/nhsx/dna-risk-predict
2. github.com/nhsx/p24-pvt-diabetes-inequal
3. github.com/nhsx/morbidity_network_analysis

Report: Coming in Jan-23

P33: Exploring Large-scale Language Models with NHS Incident Data

In collaboration with the NHS England patient safety data team, an investigation into how to produce a useful and valid representation space when training a language model for a healthcare task. This work highlighted the need for down stream tasks in order to validate the application of the language model.

Intern: Niall Taylor

Supervisor: Dan Schofield

Start date: Jun 2022

End Data: Nov 2022

Code: Coming in Jan-23

Report: Coming in Jan-23

P34: Transforming Healthcare Data with Graph-based Techniques Using SAIL DataBank

In collaboration with Swansea university and the SAIL databank, this work focused on the generation of hypergraphs for investigating the individual and joint impact of comorbidities on a patient pathway. This work will feed into two future projects to continue the creation of directed hypergraphs and then apply graph neural networks to demonstrate the process of extracting useful insights from these data.

Intern: Jamie Burke

Supervisor: Dan Schofield | Kevin Fasusi

Start date: Jul 2022

End Data: Dec 2022

Code: github.com/nhsx/hypergraph-testing

Report: Coming in Jan-23

Appendix: Wave 4 and 5

P41: Hypergraphs applied to multimorbidities

A continuation of the previous work on hypergraphs than can extract the impact of predecessor and successor diseases on disease progression pathways. This work in envisaged to include an implicit relationship to demongraphics and consider the impact of mortality.

Intern: TBA

Supervisor: Dan Schofield

Start date: Jan 2023

End Data: May 2023

Code: TBA

Report: TBA

P42: SynthVAE Production

Further development of the SynthVAE code base to include a causal model and turn the code into production level ready to be used for generating data in NHS England

Intern: TBA

Supervisor: Jonny Pearson

Start date: Jan 2023

End Data: May 2023

Code: TBA

Report: TBA

P43: Enriching Clinical Coding for Neurology Pathways using MedCAT

In collaboration with Lancaster teaching hospital and the University of Lancaster we aim to apply MedCat (an automated named entity recognition with linkage algorithm) to neurology letters to identify related SNOMED CT coding.

Intern: TBA

Supervisor: Paul Carroll

Start date: Jan 2023

End Data: May 2023

Code: TBA

Report: TBA

P44: Using Population health data to identify health inequalities in CVD

A continued collaboration with East Suffolk and North Essex foundation trust we aim to apply and expand our previous learning to a new disease area demonstrating the reusable nature of the code.

Intern: TBA

Supervisor: Paul Carroll

Start date: Apr 2023

End Data: Sep 2023

Code: TBA

Report: TBA

P51: Process mining to discover ambulance pathways

In collaboration with East Midlands Ambulance Service a project aiming to demonstrate the currently available methods for process mining to generate automated pathways. This work will focus on the opportunities and limitations in terms of the amount fo variation which can be included and how to extract insight from these data.

Intern: TBA

Supervisor: Paul Carroll / Kevin Fasusi

Start date: Jun 2023

End Data: Oct 2023

Code: TBA

Report: TBA

P52: Privacy Leakage from Language Models

A current unknown is the best practice for sharing langauge models which have been tuned on NHS data. The question of do these models leak privacy about their training data is currently unanswered and hotly debated. This work would investigate what approach's and considerations we need to ahve in place when using and sharing these models.

Intern: TBA

Supervisor: Dan Schofield

Start date: Jun 2023

End Data: Oct 2023

Code: TBA

Report: TBA

P53: Application of Graph Neural Networks

In collaboration with Swansea University and SAIL databank, this work will investigate the application of neural networks to our previously established graph representations to demonstrate the power of combining appropriate data representation with powerful data science algorithms.

Intern: TBA

Supervisor: Dan Schofield

Start date: Jun 2023

End Data: Oct 2023

Code: TBA

Report: TBA

P54: SynthVAE - Longitundinal

An extension to Synthvae investigating how a causal model approach allows for multi-table and/or longitudinal data to be generated.

Intern: TBA

Supervisor: Jonny Pearson

Start date: Jun 2023

End Data: Oct 2023

Code: TBA

Report: TBA

P55: Fairness and Bias

A research project investigating how to describe and implement fairness and bias in our models. This work will not just focus on the technical side but also on the relates standards and legal definitions that fairness measures need to adhere to.

Intern: TBA

Supervisor: Jonny Pearson

Start date: Jun 2023

End Data: Oct 2023

Code: TBA

Report: TBA