ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA241059

GIF: An Automated Genomic Information Finder to Extract Data from Reports

Pavithra Rajendran[a,*], Alexandros Zenonos[b], Sebin Sabu[a], Anastasia Spiridou[a], Helena Spiridou Goncalves[a], Nastazja Laskowski[b], Daniel Key[a], Shiren Patel[a], Rebecca Pope[b] and Neil Sebire[a]

[a]Great Ormond Street Hospital NHS Foundation Trust
[b]Roche Products Ltd

Abstract. Genomic testing is becoming increasingly used within the UK National Health Service (NHS) in an effort to deliver precision medicine that can directly benefit patient management and care. Typically, in hospital settings clinical scientists interpret genomic results and store the output in reports (unstructured format). Recent advances in the field of Natural Language Processing (NLP) techniques have significant advantage over the common time-consuming and laborious manual extraction of the relevant information from the reports. There are a plethora of open-source NLP models available, but limited evidence of their performance on real-world healthcare tasks, specifically for paediatric data. In this paper, we describe the development of an automated pipeline that uses a hybrid approach of combining rules and pretrained NLP models to extract gene variants, related information of these variants and any gene panel information present within the genomic test reports. We evaluated the performance of the pipeline against a manually-curated, expert-annotated, in-house data containing 372 reports. Our results and evaluation highlights the advantages and limitations of using existing pretrained models in a real-world setting, and in particular, when there are computation and resource constraints within the hospital setting.

1 Introduction

Genomic medicine has the potential to offer a greater understanding of how our genetic makeup impacts on our health and to change the way disease is managed and treated[1]. In England, genomic tests are sequenced by Genomic Laboratory Hubs (GLHs) who receive samples from multiple healthcare organisations (HCOs) across the country. Genomic test reports contain information about a patient's health status and genetic variants, as well as implicit information about close family members. The reports support clinicians in making better decisions on diagnoses, risk stratification, treatments, and patients' suitability for clinical trials. In the future, the information embedded within genomic reports may inform disease prognosis, pharmacogenomics, population health studies, drug discovery and significant advance research efforts. The results are interpreted by clinical scientist, who generate a genomic report that is stored as a PDF. Currently, there is an ongoing effort to harmonise to a small number of genomic report types. However, a lot of heterogeneity still exists across the GLHs due to local practices, variation in testing methodology and reporting of variant types. In addition, historic genomic reports vary significantly in both structure and content.

In this paper, we describe an automated pipeline developed for extracting clinically relevant data from the reports (see Figure. 1) and provision the data in a structured format along with our standardised structured data (curated structured data from different Electronic Patient Record (EPR) systems within the hospital) for researchers and clinicians. The outcome of the pipeline will enhance our current standard structured data extraction with additional genomic information, which will aid clinical staff and researchers to access pertinent data at scale for research, better care and planning purposes.

Our contributions are given below:

1. We evidence the potential capability of the developed automated pipeline on the extraction of clinically relevant genomic information in real-world paediatric healthcare settings. This is illustrated through our transition from a Proof-of-Concept with limited compute and data to a deployed solution within our secure on-premise environment. The deployed solution was run on a large volume of 26,464 documents.
2. Our results and evaluation of the automated pipeline was compared with a manually-curated expert annotated data and show the overall potential advantage of using existing pre-trained NLP models for significantly improving a rule-based approach.
3. We also highlight on the limitations of pretrained NLP models trained using open-source datasets on our in-house data, especially in certain diseases areas.

2 Related Work

Extracting information from unstructured data is a time-consuming task when performed manually within the clinical setting. An alternative to manual extraction would be the use of automated approaches underpinned by artificial intelligence and machine learning methods, specifically, natural language processing (NLP). The essence of NLP is in programming a computer to understand the morphology, syntax, semantics and pragmatics of written human language. In the context of genomic reports and this paper, this would be applied for the purposes of information extraction (IE). Developing NLP tools for genomic information represents a historical challenge and prior work has focused on more traditional linguistic approaches by leveraging n-grams from the text [4]. Specifically, an algorithmic, programmatic

* Corresponding Author. Email: pavithra.rajendran@gosh.nhs.uk
[1] https://www.england.nhs.uk/wp-content/uploads/2022/10/B1627-Accelerating-Genomic-Medicine-October-2022.pdf

automated approach that takes unstructured (or semi-structured) text as an input and converts it into structured data as an output, with high levels of accuracy, remains elusive in clinical practice. There are three main approaches that could be applied to tackling automatic extraction of structured information from unstructured reports, and in our case, the genomic information. This can follow in isolation or combination of the following approaches: (1) Rule-based, (2) supervised machine learning (ML) and (3) unsupervised ML [1]. Rule-based NLP is interpretable, domain-specific and allows fine tuning of the output, but it requires manual time-consuming rule creation and has difficulty handling ambiguity. Supervised machine learning (ML) methods are more adaptable, but depend on the availability of labelled data that is often lacking in a clinical setting. Although unsupervised models do not require labelled data, they have a blackbox nature, which is not desirable in a clinical setting.

In recent years, efforts have begun to focus on applying NLP methods to genomic information within PubMed extracts to unlock valuable scientific insights ([2], [3]). More specifically, Botsis et al. [2] developed a platform that uses a rule-based system with lexical resources from various databases. Although these results show promise, the use of NLP approaches in a clinical setting applied to genomic reports, has not been extensively evaluated. Chen et al. [3] found that a hybrid approach of ML fine-tuned with rule-based extraction worked best in extracting genomic information from Chinese medical records. To date, there is no English equivalent real-world study in genomic reports.

3 Data Description

3.1 Input Data

An automated data extraction process was carried out to extract retrospective documents stored within the hospital's different EPR systems ranging from August 2018 to August 2023 respectively. A SQL Query $\mathcal{Q}_{\mathcal{D}}$ based on a specific set of Internal Classification of Diseases (ICD) codes represented by $\mathcal{D}$ is used for this purpose. Here, $\mathbf{d}_i \in \mathcal{D}$ represents a specific disease area. In our case, we chose the following: *epilepsy*, *spinal muscular atrophy*, *muscular dystrophy*, *cystic fibrosis*, *neuroblastomas* and *tumours* respectively.

This provided us with a total of 26,982 documents, out of which, 2894 were genomic reports. A subset of 372 genomic reports were used for producing a manually expert-curated ground-truth dataset, which is used for evaluation purposes. This dataset contains structured information extracted manually by an expert annotator based on a set of annotation guidelines provided and using the RedCap tool. [2]

Table. 1 provides the total number of reports used per diagnosis for the ground truth dataset.

3.2 Output Data

The structured information is extracted from the reports are as follows:

- Given a variant name, the final classification outcome or the conclusion of the result. This is either positive or negative.
- Given a variant name, it's corresponding information (if applicable) such as Transcript Reference ID, DNA Change ID, Amino Change ID, Inheritance, Zygosity.
- List of genes tested. In a report, the information on the different genes tested is provided as a list of genes or sometimes linked to a panelapp version present in the PanelApp Genomics England database. For example, in Figure. 1, the example report lists out all the genes tested within the test methodology. In such cases, full list of genes is extracted.

Table 1. Total number of reports per diagnosis that is used for preparing the ground truth dataset.

Diagnosis	Count of Reports
Epilepsy	65
Spinal Muscular Atrophy	38
Muscular Dystrophy	43
Cystic Fibrosis	111
Tumours	86
Neuroblastomas	29

4 Proposed Approach

Our proposed approach uses an automated pipeline $\mathcal{P}$, the outcomes of which in combination with a set of expert-curated rules $\mathcal{R}$ is used for extracting the relevant structured information. The main steps are described below.

4.1 Document Filtering

In this step, we automatically filter relevant documents that are specifically genomic reports representing a predefined subset of diagnosis. From these filtered documents, the tables and text content are extracted separately.

4.2 Table and Section Classification

In this step, each of the detected table within a genomic report is classified based on the table header information. This information can either be present horizontally or vertically. We make use of the header titles provided within a predefined knowledge base to identify whether the titles are present horizontally or vertically. A distant-supervision based approach is used for classifying the table headers. Here, within the knowledge base, each table type is mapped to a set of predefined header titles which is used in the classification process.

Similarly, using a knowledge base that maps a set of predefined section headers to a standard section title is used for classifying sentences as section headers.

4.3 Gene Variant and Classification Outcome

In this step, gene variant names are detected automatically using a Named Entity Recognition (NER) tool, using these variant name predictions, corresponding information (see Section. 3.2) and final outcomes or classification information is extracted using a rule-based approach using the rules present in $\mathcal{R}$.

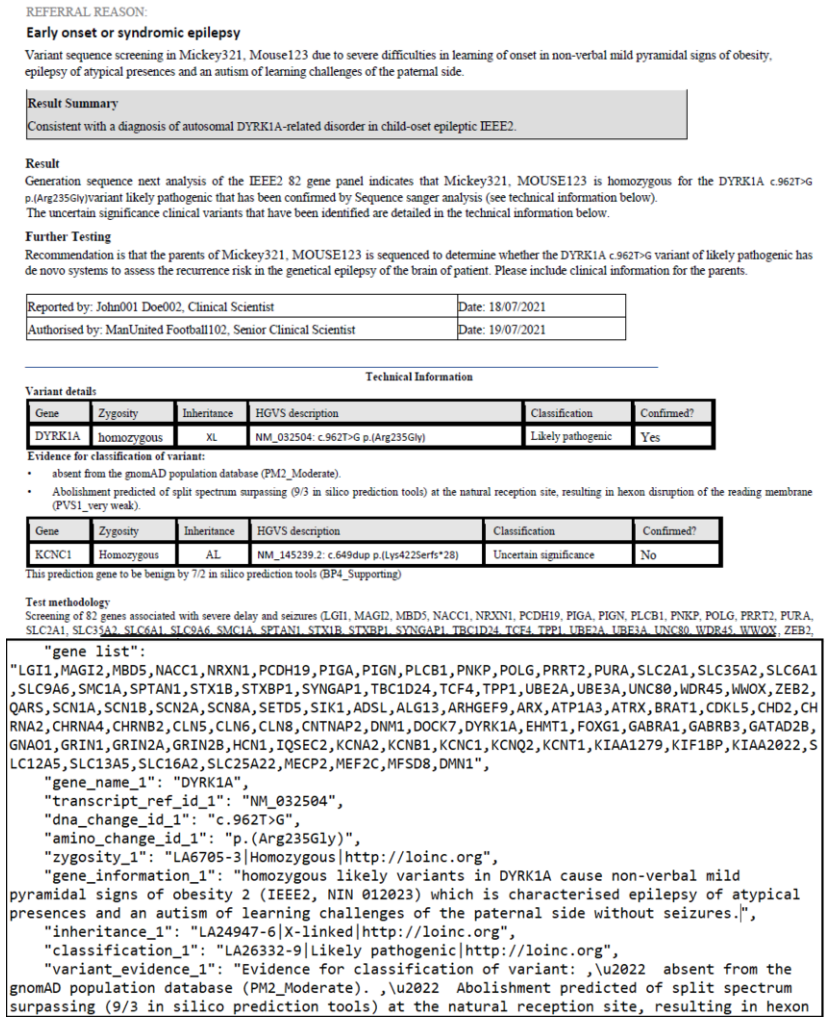
The implementation details for the different steps are described in detail below.

5 Implementation Details

5.1 Document Filtering

Figure. 2 shows a detailed flow diagram of the different steps performed for document filtering. Documents are automatically filtered

[2] www.project-redcap.org



REFERRAL REASON:

Early onset or syndromic epilepsy

Variant sequence screening in Mickey321, Mouse123 due to severe difficulties in learning of onset in non-verbal mild pyramidal signs of obesity, epilepsy of atypical presences and an autism of learning challenges of the paternal side.

Result Summary

Consistent with a diagnosis of autosomal DYRK1A-related disorder in child-oset epileptic IEEE2.

Result

Generation sequence next analysis of the IEEE2 82 gene panel indicates that Mickey321, MOUSE123 is homozygous for the DYRK1A c.962T>G p.(Arg235Gly)variant likely pathogenic that has been confirmed by Sequence sanger analysis (see technical information below).
The uncertain significance clinical variants that have been identified are detailed in the technical information below.

Further Testing

Recommendation is that the parents of Mickey321, MOUSE123 is sequenced to determine whether the DYRK1A c.962T>G variant of likely pathogenic has de novo systems to assess the recurrence risk in the genetical epilepsy of the brain of patient. Please include clinical information for the parents.

Reported by: John001 Doe002, Clinical Scientist	Date: 18/07/2021
Authorised by: ManUnited Football102, Senior Clinical Scientist	Date: 19/07/2021

Technical Information

Variant details

Gene	Zygosity	Inheritance	HGVS description	Classification	Confirmed?
DYRK1A	homozygous	XL	NM_032504: c.962T>G p.(Arg235Gly)	Likely pathogenic	Yes

Evidence for classification of variant:

- absent from the gnomAD population database (PM2_Moderate).
- Abolishment predicted of split spectrum surpassing (9/3 in silico prediction tools) at the natural reception site, resulting in hexon disruption of the reading membrane (PVS1_very weak).

Gene	Zygosity	Inheritance	HGVS description	Classification	Confirmed?
KCNC1	Homozygous	AL	NM_145239.2: c.649dup p.(Lys422Serfs*28)	Uncertain significance	No

This prediction gene to be benign by 7/2 in silico prediction tools (BP4_Supporting)

Test methodology

Screening of 82 genes associated with severe delay and seizures (LGI1, MAGI2, MBD5, NACC1, NRXN1, PCDH19, PIGA, PIGN, PLCB1, PNKP, POLG, PRRT2, PURA, SLC2A1, SLC35A2, SLC6A1, SLC9A6, SMC1A, SPTAN1, STX1B, STXBP1, SYNGAP1, TBC1D24, TCF4, TPP1, UBE2A, UBE3A, UNC80, WDR45, WWOX, ZEB2,


```
    "gene list":
"LGI1,MAGI2,MBD5,NACC1,NRXN1,PCDH19,PIGA,PIGN,PLCB1,PNKP,POLG,PRRT2,PURA,SLC2A1,SLC35A2,SLC6A1
,SLC9A6,SMC1A,SPTAN1,STX1B,STXBP1,SYNGAP1,TBC1D24,TCF4,TPP1,UBE2A,UBE3A,UNC80,WDR45,WWOX,ZEB2,
QARS,SCN1A,SCN1B,SCN2A,SCN8A,SETD5,SIK1,ADSL,ALG13,ARHGEF9,ARX,ATP1A3,ATRX,BRAT1,CDKL5,CHD2,CH
RNA2,CHRNA4,CHRNB2,CLN5,CLN6,CLN8,CNTNAP2,DNM1,DOCK7,DYRK1A,EHMT1,FOXG1,GABRA1,GABRB3,GATAD2B,
GNAO1,GRIN1,GRIN2A,GRIN2B,HCN1,IQSEC2,KCNA2,KCNB1,KCNC1,KCNQ2,KCNT1,KIAA1279,KIF1BP,KIAA2022,S
LC12A5,SLC13A5,SLC16A2,SLC25A22,MECP2,MEF2C,MFSD8,DMN1",
    "gene_name_1": "DYRK1A",
    "transcript_ref_id_1": "NM_032504",
    "dna_change_id_1": "c.962T>G",
    "amino_change_id_1": "p.(Arg235Gly)",
    "zygosity_1": "LA6705-3|Homozygous|http://loinc.org",
    "gene_information_1": "homozygous likely variants in DYRK1A cause non-verbal mild
pyramidal signs of obesity 2 (IEEE2, NIN 012023) which is characterised epilepsy of atypical
presences and an autism of learning challenges of the paternal side without seizures.",
    "inheritance_1": "LA24947-6|X-linked|http://loinc.org",
    "classification_1": "LA26332-9|Likely pathogenic|http://loinc.org",
    "variant_evidence_1": "Evidence for classification of variant: ,\u2022  absent from the
gnomAD population database (PM2_Moderate). ,\u2022  Abolishment predicted of split spectrum
surpassing (9/3 in silico prediction tools) at the natural reception site, resulting in hexon
```


Figure 1. Example dummy report created using our hospital template and a sample output from the proposed automated pipeline.

as *machine-readable* or not, and whether they are *genomic reports* or not. Here, since most of our prospective reports are *machine-readable* documents, we do not focus on scanned reports. To classify whether a report is *machine-readable* or not, we use the open-source Python package, PdfPlumber [3]. Next, we create a knowledge base, mapping unique keyphrases relevant to a predefined set of diagnosis and a set of unique keyphrases from genomic reports. We use a rule-based NLP classification algorithm that identifies these predefined keyphrases present in documents to distinguish between the genomic reports and other documents (e.g., referral letters).

These keyphrases identified for this purpose were extracted using the following steps:

1. Extracting a random set of 68,508 text entries including RTF-formatted texts from lab documents including genomic reports such that it does not overlap with our ground-truth data,
2. using an open-sourced pretrained NER model (HunFlair) trained for detecting gene variants to detect gene variants within the text entries,
3. filtering entries with gene variants mentioned more than once and,
4. a manual approach where an annotator picked potential key phrases, mainly subtitles and titles from the filtered entries.

[3] https://pypi.org/project/pdfplumber/

Due to computation limitations, we chose to use a rule-based NLP algorithm using the keyphrases picked in (4) for the classification purpose.

5.2 *Table and Section Classification*

The open-source Python package, PDFPlumber, is used for extracting the raw text and tables separately. Given a document $\mathbf{g}_i \in \mathcal{G}$ such that $\mathbf{g}_i$ represents the $i-th$ genomic report within a set of genomic reports, $\mathcal{G}$, a rule-based NLP algorithm is used for classifying k number of tables in a report represented as $\mathbf{t}_k \in \mathcal{T}_{g_i}$. This classification is based on the header information and the outputs of the classification is from the following classes: *patient information*, *variant information*, *report and authorisation information*, *panel and coverage information* or *not relevant*.

The raw text content $\mathcal{C}_{g_i}$ for a corresponding genomic report g_i is tokenized into a list of sentences $\mathcal{S}_{g_i}$. Each sentence $\mathbf{s}_j \in \mathcal{S}_{g_i}$, if is a potential section header, is then mapped to a standardised section header, using a distant supervision approach that fuzzy-matches it with a list of section headers within a predefined knowledge base. The knowledge base is populated with section titles and sub-titles collected from different genomic report templates that are provided by our hospital organisation, as well as other organisations. An

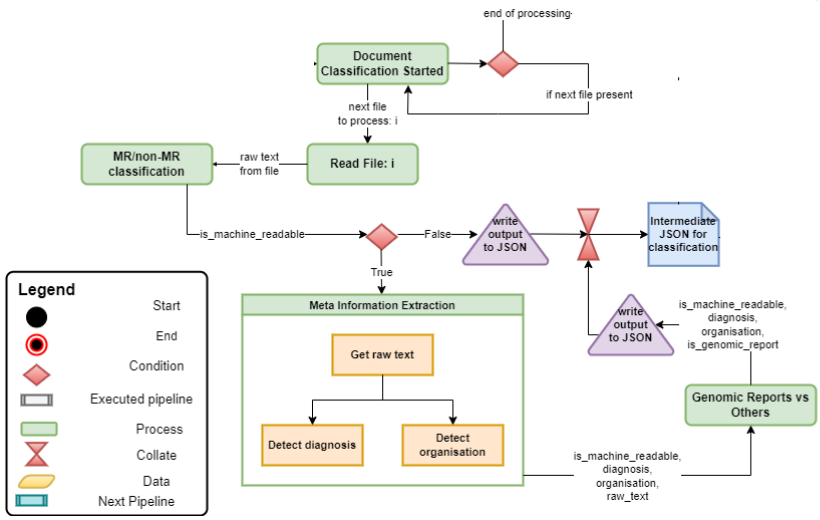


Figure 2. Flowchart diagram representing the different steps described in Section. 5.1 that includes classification of each document as *machine-readable* (MR) or not, and whether the document is a *genomic report* or not. Additional metadata information such as the *diagnosis* present and *organisation* title is also extracted. The final output is a JSON.

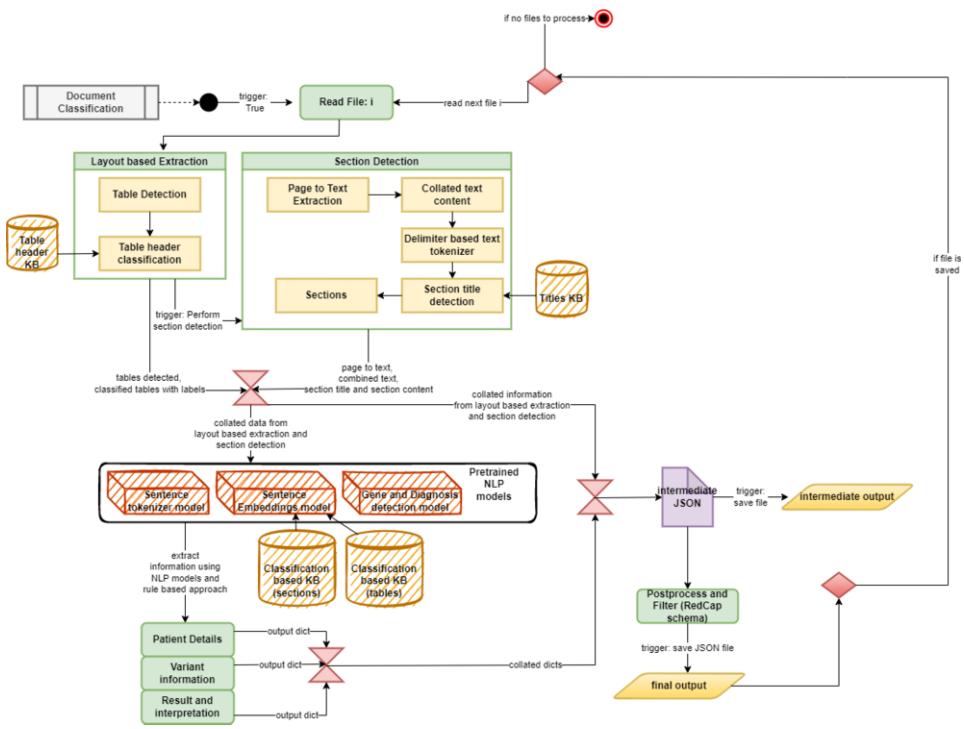


Figure 3. Flowchart diagram representing the different steps described in Sections. 5.2 and 5.3. This includes processing and classifying tables, sections, and identifying variant details and the final classification outcome. Different outputs from the different steps shown are stored as an intermediate output in a JSON and the final output is a JSON with the final structured results.

overview of this approach specific to the genomic reports is as follows and outlined in Figure, 3:

- Identify whether there is any section title related to *referral reason* present within the introductory section
- Ensure that the section title related to *result* is identified before retrieving the title related to *testing or follow-up*.
- The final search is to identify any section titles related to *test information* and in some cases, *gene information.*

5.3 Gene Variant and Classification Outcome

This subsection explains the use of different pretrained NLP models for (1) identifying gene variants and, (2) the classification outcome. Both these information are present either in tables or in the text content.

For identifying any gene variants present in a report, we use the **pretrained state-of-the-art NER tagger, HunFlair** [8], a neural model trained on a character-level language model with around 24 million biomedical abstracts and roughly around 3 million biomedical texts. The model further integrates 31 biomedical NER datasets for identifying different entity types. Any information relevant to the corresponding gene variant are then picked using the sentences in the which the gene variant occurs and expert-curated rules $\mathcal{R}$. For example, we can extract its corresponding DNA change ID and Protein change ID. In addition, based on the gene variant name, the gene information, if present, in the test information section is extracted. This can also help us to extract the transcript ID if present.

There are two ways to identify the classification outcome: (1) from a table containing the gene variant results and (2) from the text content present in the *results and interpretation* section. For a given set of classification labels, a knowledge base containing sentences for each corresponding classification label is collected with the help of an expert annotator.

A set of pretrained sentence-transformers [5] NLP models were investigated as part of an Exploratory Data Analysis (EDA) to chose the best retriever model. The retriever model is used for identifying whether there exists a sentence within the table or within the text content which is semantically closest to one of the sentences within the knowledge base.

A knowledge base with few example sentences mapped to a variant classification outcome is created with the help of an expert annotator. Here, we are interested in two different outcomes namely, whether the reported variant is positive or negative. Since we do not have a large volume of such labelled data, we use this in a few-shot setting [6], where we use a retriever model to pick the best sentence from the knowledge base that is closest to the information present within the report. This is done by computing the embedding vectors for both (a) sentences present in the knowledge base and (b) sentences present in the *result and interpretation section* and comparing the cosine similarity between the embedding vectors. The best sentence chosen in the knowledge base based on the cosine similarity score is then used and its corresponding label is considered as the final classification outcome. In order to reduce errors, we use a threshold value of *0.50* after conducting various experiments to ignore the classification outcome, if the cosine similarity measure computed is less than the threshold value.

6 Real-World Deployment

Our in-house, secure, on-premise environment (GRID) has two servers – (1) development server with internet access and no access to patient data and, (2) staging server with access to secure data and no internet access and limited to restricted users. Our automated pipeline is developed, tested and containerised using the open-source solution, Podman [4] within our development server which is a secure on-premise Linux environment. Here, all the pretrained models and required packages are downloaded into the container image since the secure environment does not have access to internet. The container image is then deployed moved into the staging server and run on patient identifiable 26,464 documents. The outputs are stored in a secured shared drive that is accessible to restricted users. This deployment was run as a pilot study and the results were evaluated based on a subset of data, i.e. ground-truth data. All the source code [5] are version-controlled and managed within our internal GitLab repository.

[4] https://podman.io/
[5] https://github.com/gosh-dre/genomics_nlp_pipeline

7 Results and Evaluation

In this section, we present the overall results and evaluation on the experiments carried out for several steps within the pipeline as discussed below.

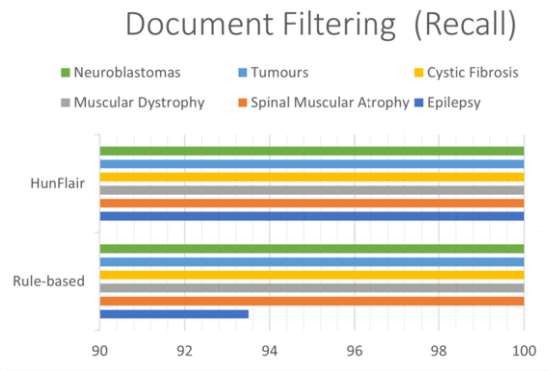


Figure 4. Comparison of the outputs from the document filtering approach and using HunFlair [8] model with the ground-truth data. Here, we present the recall rate.

7.1 Document Filtering

Given the large volume of documents, we were interested in understanding whether the distant-supervision based approach provided significant filtering. Thus, we compared the document filtering approach on the ground-truth and based on using the HunFlair model. Figure 4 provides the recall rate based on each diagnosis. Upon investigation, there are a number of older templates of epilepsy that do not have any proper sub-sections and hence, a lower performance based on the distant-supervision approach.

7.2 Gene Variant detection

Our experiments were carried out using the HunFlair model for detecting variant names and using the expert-curated rules $\mathcal{R}$ in combination with the detected variant names for identifying other related information. Table. 2 provides the F1-scores and the results, specifically, *related information* that reports the performance of using the automatically identified variants in combination with the rules in $\mathcal{R}$, and indicate a clear dependency on the detection of variants. Our error analysis showed the following observations:

- Most of the variants that were not identified were from reports on diagnosis *Muscular Dystrophy* and *Spinal Muscular Atrophy*.
- Variants identified from reports on diagnosis *Cystic Fibrosis* showed that variants were incorrectly identified without distinguishing whether the reported variant was referring to someone else (e.g., patient's partner or parent).
- Information that required more contextual understanding for mapping to the relevant information (e.g., chromosomes, methylations, exons etc.) were not captured.

Table 2. Comparing the automatically detected variants and corresponding related information identified with the ground-truth data. F1 scores are reported.

Diagnosis	Variants Identified	Related Information
Epilepsy	89.23%	93.84%
Spinal Muscular Atrophy	84.21%	78.94%
Muscular Dystrophy	81.39%	76.74%
Cystic Fibrosis	94.59%	93.69%
Tumours	88.09%	88.37%
Neuroblastomas	75.86%	79.30%
Grand Total	**91.12%**	**87.90%**

7.3 Classification Outcome

In this task, we chose different pretrained sentence-transformer models [5], both domain-specific and generic ones, and a pretrained Instructor model [7]. The models used are described below.

BioBERT-NLI Sentence-transformer model fine-tuned on SNLI and mutliNLI datasets.

multi-qa-miniLM-L6-cos-v1 Sentence-transformer model trained to perform well in semantic search tasks, this model was trained on diverse Question-Answering data; output is a 384 dimensional vector.

multi-qa-mpnet-base-cos-v1 Similar to the above model but outputs 768 dimensional vector.

Instructor-Large Instruction fine-tuned text embedding model that provides embedding vectors based on task instruction provided as a prompt.

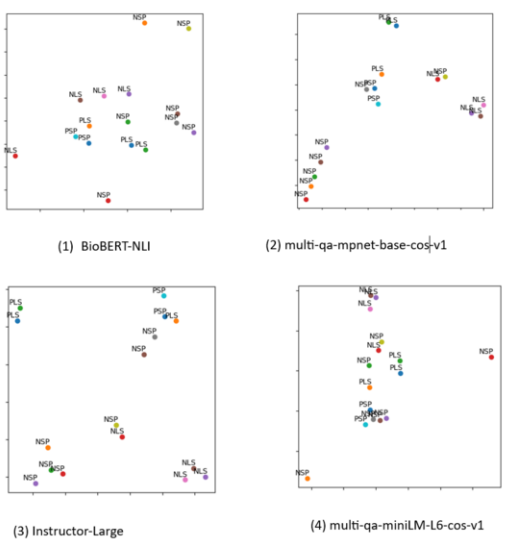


Figure 5. Embedding vectors computed using the different models and mapped to a 2-dimensional vector space using TSNE method. *NSP*, *NLS*, *PSP* and *PLS* represent negative short phrases, negative long sentences, positive short phrases and positive long sentences respectively.

To understand the semantic relatedness among the examples for each classification outcome present within the knowledge base, we computed the embedding vectors for each of the model and used T-Distributed Stochastic Neighbour Embedding (TSNE) dimensional-reduction method for visualising the vectors in a 2-dimensional space. Figure. 5 visualises the vectors obtained using the different models. From the visualisations, we can observe that the *Instructor-Large* model has the best separation between the positive and negative outcome examples.

We experimented on 275 sentences extracted from the ground-truth dataset labelled with the classification outcome to understand whether the models show similar performance. Results are present in Table. 4 which show us that the *Instructor-Large* model performed poorly while the domain-specific model *BioBERT-NLI* was able to retrieve correct labels with the best performance. This indicates the need to have more examples with more variability in the text content as found within the 275 sentences for a better performance.

We use the *BIOBERT-NLI* as our finalised retriever model and compared the results with the ground-truth data (Table. 3). From the results, we can observe that *epilepsy* reports have the best classification outcome results. Overall, using pretrained NLP models provided good results, given that, these models were not trained on paediatric data.

Table 3. Comparing the automatically detected classification outcome with the ground-truth data. F1 score are reported.

Diagnosis	Classification Outcome
epilepsy	**86.15%**
Spinal Muscular Atrophy	78.94%
Muscular Dystrophy	76.31%
Cystic Fibrosis	76.75%
tumours	73.25%
neuroblastomas	48.27%
Grand Total	**74.46%**

Table 4. Comparing the performance of the retriever models with the ground-truth labels on 275 sentences. F1 scores are reported.

Retriever Model	F1-Score
BioBERT-NLI	**81%**
multi-qa-miniLM-L6-cos-v1	64%
multi-qa-mpnet-base-dot-v1	65%
Instructor-Large	68%

8 Error Analysis

Based on our error analysis on the predictions made on the ground truth data, we observed the following:

- *Variants identified*: Spinal Muscular Atrophy and Muscular Dystrophy therapeutic areas posed the biggest challenge regarding missed variants. Incorrect variants were mostly from reports mentioning the partner or parent;
- *Related information*: All information corresponding to a variant was not captured when the missed information required more understanding of the text (e.g. chromosomes, exons, methylation).

9 Limitations and Future Work

The current pipeline was built using existing pretrained models that were not fine-tuned to our data. Further, the pipeline has not been tested across other types of reports (e.g. radiology reports). As an immediate future work, we would like to validate the results of the pipeline on a larger volume of data to understand the limitations of the current pipeline and focus on enhancing it. The current pipeline is also tested on a limited set of diagnosis and requires further work to assess on the different types of genomic reports that may also provide additional variant information.

The continuance of integrating and improving different components and ensuring the components are developed in a customisable

and modularised manner. There are several challenges that are required to be addressed in our future work, after which we would like to deploy and use the pipeline on prospective genomic reports; not just historic that is currently deployed for. For example, improving section detection by automatically classifying sections by finetuning pretrained NLP models with annotated data. We aim to link the structured information to clinical decision support tools, to provide actionable information in a timely manner to Health Care Professionals. As more and more Large Language Models (LLMs) are being developed and open-sourced, we aim to explore their use and effectiveness as part of this pipeline. In particular, we want to explore the use of compact LLMs that can run efficiently on our infrastructure.

10 Conclusion

Genomic reports represent a reservoir of data in the EHR systems of the NHS, which is not yet used to its full potential. Whilst the NHS holds one of the largest resources of genetic information on a population level, this information cannot currently be fed into data analytics tools, as it is not digitally integrated. We have developed an NLP pipeline that classifies genomic reports from the EHR and unlocks meaningful information. In this work, we demonstrated the potential capability of using an NLP-based approach for automatically extracting data from genomic reports. This pipeline is deployed within our on-premise environment and run on a large volume of 26,464 documents. We have also addressed the limitations of the current work and our future plan to enhance this pipeline to be able to use it for prospective reports, and thereby linking the data with our standard structured data for integrating with clinical decision making tools. We make our code open-source and available (https://github.com/gosh-dre/genomics_nlp_pipeline) and interested to engage with other hospitals that may be interested in evaluating it with their genomic reports.

Acknowledgements

This activity is part of a collaborative working agreement between Great Ormond Street Hospital NHS Foundation Trust and Roche Products Ltd. M-GB-00019098 | April 2024

We would also like to acknowledge Natalie Chandler (Principal clinical scientist) and Amy McTague (Consultant neurologist) who provided support in understanding genomic data flows. We would also like to thank the anonymous reviewers for their invaluable feedback.

References

[1] G. T. Berge, O.-C. Granmo, T. O. Tveit, A. L. Ruthjersen, and J. Sharma. Combining unsupervised, supervised and rule-based learning: the case of detecting patient allergies in electronic health records. *BMC Medical Informatics and Decision Making*, 23(1):188, 2023.

[2] T. Botsis, J. Murray, L. Alessandro, D. Palsgrove, W. Wei, J. R. White, V. E. Velculescu, V. Anagnostou, J. H. M. T. B. Investigators, et al. Natural language processing approaches for retrieval of clinically relevant genomic information in cancer. *Studies in health technology and informatics*, 295:350, 2022.

[3] L. Chen, L. Song, Y. Shao, D. Li, and K. Ding. Using natural language processing to extract clinically useful information from chinese electronic medical records. *International journal of medical informatics*, 124:6–12, 2019.

[4] T. Hishiki, N. Collier, C. Nobata, T. Okazaki, N. Ogata, T. Sekimizu, R. Steiner, H. S. Park, and J. Tsujii. Developing nlp tools for genome informatics: An information extraction perspective. *Genome Informatics*, 9:81–90, 1998.

[5] N. Reimers and I. Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020.

[6] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[7] H. Su, W. Shi, J. Kasai, Y. Wang, Y. Hu, M. Ostendorf, W.-t. Yih, N. A. Smith, L. Zettlemoyer, and T. Yu. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, 2023.

[8] L. Weber, M. Sänger, J. Münchmeyer, M. Habibi, U. Leser, and A. Akbik. HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics*, 37:2792–2794, 2021.