The Importance of Anonymised Unstructured Data in Advancing Medical Research and Patient Outcomes

Unstructured free text in electronic health records (EHR) can provide an invaluable source of information beyond structured (coded) information for medical research. Free text within EHRs may contain diagnoses, investigation results, medication side-effects, symptoms, social issues, reasons for switches in therapy and cause of death. It may also contain third-party letter entries, for example, secondary or private care and laboratory results.

If health records are sufficiently anonymised and the NHS Health Research Authority (HRA) provides approval, the data can be used for research without requiring patient consent. The Health Improvement Network (THIN) is a research database containing GP records which has overarching HRA approval.

However, free text in health records may contain information that could potentially identify a person. Therefore use of this information for research generally requires individual patient consent, which is a major barrier for large population-based studies. Data protection regulations such as GDPR and the Common Law Duty of Confidentiality prohibit the use of identifiable patient information for research without consent, unless specific legal exemption is granted. In England and Wales, such an exemption (known as Section 251) can be granted by the HRA Confidentiality Advisory Group, but these exemptions are usually granted for one-off use of identifiers to link patient datasets, rather than ongoing analysis of identifiable free text.

Some large NHS trusts have developed in-house capabilities to analyse the text within their patient records without the data leaving the trust. However, for GP data, records from a large number of practices need to be combined to create a large enough research database for statistical power. To be successful in harnessing this data, we need a system that is secure, and has enough trust built into it, to enable the data to be analysed at scale whilst maintaining patient confidentiality.

Samir Dhalla, Head of THIN and Dr. Anoop Shah, Clinical Lecturer at UCL Institute of Health Informatics highlight how the ability to re-harness the use of free text is paramount in the medical research arena, providing access to millions of patient records with a higher quality of data and information. Therefore, it was vital to overcome the challenge of handling identifiable free text in a safe, legal and privacy- preserving way that has patient and public support.

Benefiting Patient Outcomes with Free Text

Data remains critical in informing healthcare pathways and provision, but the way it is recorded can significantly impact its usability. Clinical data stored in electronic health records can be either classified as structured or unstructured. Structured data is highly organised and follows a prescribed data model and value set which is easily searchable, including numerical and categorical values. However, using structured data alone can limit the data that is available to research topics, such as patient experiences or clinical reasoning, because this information is rarely recorded in a structured way.

In contrast, unstructured data has no predefined format, offering more flexibility and freedom when recording an entry, which can make it more complex to collect, process and analyse. This type of data is typically found in GP notes, for example, and can often require manual interpretation.

The free text may contain information that provides insight into the consultation and decision-making process. GPs may record the final diagnosis in a structured format, but during the consultation process will typically make notes, which are recorded as unstructured data. It's this narrative that sits between the initial consultation and the ultimate diagnosis that is going to help inform future patient consultations, pathways – and outcomes.

Previous research conducted around heart failure, for example, has found that symptoms such as shortness of breath, tiredness and leg swelling may arise months before a formal diagnosis is made. Research using the free text can help to understand the reason for any delays and help to improve patient pathways to reduce these delays in the future. It may also enable GPs to make better, more informed decisions to investigate patients at risk earlier.

Another potential benefit of using free text is to study the side-effects of medication. Medication tested in clinical trials only involves a few thousand patients, compared to tens of thousands of patients in practice. There may be rare side-effects that are only found out when a larger number of patients or a more diverse group of patients is exposed to specific medication. Some of these symptoms or adverse effects of medication might not be coded, but will be recorded in the free text.

The Technology Behind the Solution

With the future of healthcare acknowledged as being data-driven, technology continues to reshape the way data is collected, coded and used. Natural language processing tools can be used to automatically anonymise, classify or extract coded information from unstructured free text

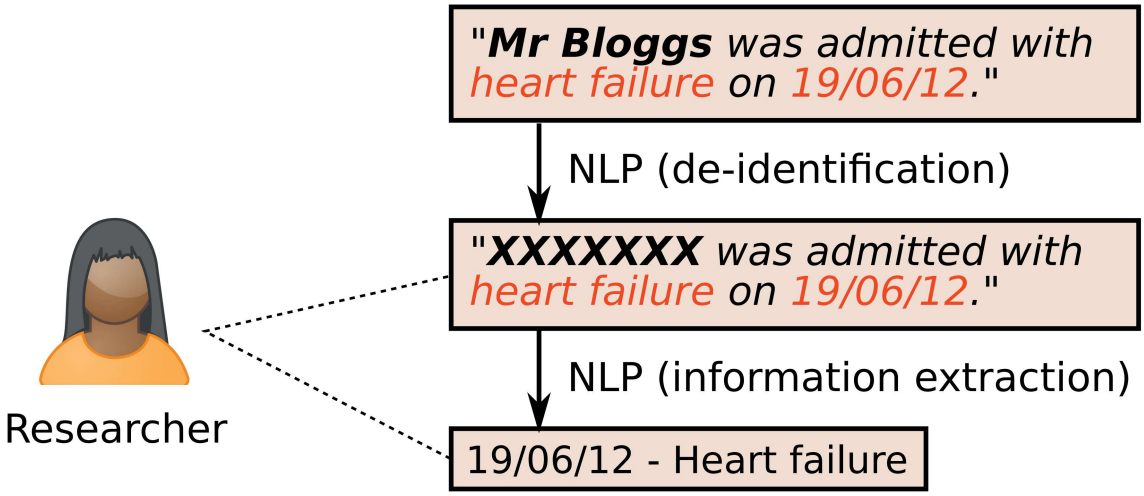
Automated anonymisation software based on rules or a machine learning approach can redact identifying information in the text. Although such automated methods can never be 100% effective, they form a key part in the data minimisation process so that when personal data is analysed, the amount of sensitive and uniquely identifiable information being processed at each stage is minimised.

Machine learning algorithms can 'learn' features from the text and apply them to classification tasks. A possible application might be to classify the purpose of the consultation, such as whether it was an emergency or a routine patient appointment. The algorithm learns from manually annotated samples ('supervised learning'), and applies the learned model to classify new texts.

Information extraction algorithms can be used to extract a range of different items of information from the text. This

To understand what information is in the text of clinical notes and letters, we can create a secure research database. Computers analyse the text (natural language processing, NLP).

NLP can remove names and other identifiers, and extract clinical information. Samples of text are checked by researchers.



can be done using rules-based algorithms, which break down the text into individual words, detect parts of speech and context, and map words and phrases to standard terminology. Such algorithms can then be used to extract a list of diagnoses or symptoms from the text, along with their context, such as negation.

Machine learning algorithms have also been developed for this type of task, and have the advantage that they can learn from additional data that is labelled by manual annotation. The benefits of using an automated approach include saving time and learning on the job from the outcomes – ultimately learning from any mistakes to achieve the correct results.

As discussed, there are many technologies available to extract information from unstructured data. However, there is also the opportunity to use these same technologies to create structured information in real time. A new project in University College London hospitals will use natural language processing (NLP) within electronic health records to create structured data as soon as the free text note is created. The clinician will automatically receive suggestions for structured data, which can be inserted and validated at the point of care. The aim is to create structured data more efficiently from the outset, which will have widespread benefits.

Key Challenges

There remain some fundamental barriers to the wider accessibility of anonymised unstructured data that, if we are to truly advance patient outcomes on both a population health and personalised level, need to be overcome. These include pre- and misconceptions at both a patient and healthcare professional level.

Arguably, singularly the largest challenge to overcome is an understanding of the importance of this data and how it will be used. A second major consideration is the security, privacy and ethical application of this data.

For research to be patient-centred, healthcare practitioners and patients have to understand what their records are being used for, why this is happening and what the benefits of this are. Patients and the public are only likely to consent to their records being accessed, if healthcare practitioners advocate this.

This primary challenge can be overcome by healthcare practitioners educating patients on the application of unstructured data in painting a much wider, long-term picture of healthcare and patient outcomes. With different levels of health literacy, understanding and involvement, there also has to be access to a range of educational activities to meet the needs and levels of different individuals.

With regard to concerns around security, privacy and ethics, it is also critical that we collectively address patients' suspicions around how their health data is being used, especially with the influx of negative news about the UK government selling patient data to US corporations, and fake news being spread around COVID-19 with the track and trace app, amongst others.

Corporate and organisational reputations are critical. The Health Improvement Network (THIN) is a large database of anonymised electronic medical records collected at primary care practices throughout the UK and mainland Europe. As part of its organisational structure, THIN has an advisory committee with patient representatives which takes into account patient views, ultimately supporting the THIN application for use of unstructured data, and initiating projects with patient input throughout.

Fundamentally, there remains a lack of awareness around the process of clinical research, how patient data is used and the positive impact this has on patient outcomes. There must be an emphasis on educating patients about how their data can be used to benefit the healthcare system as a whole, as well as individual patients.

Using Anonymised Free Text in Healthcare
The use of anonymised free text in the healthcare and life science industries has the potential to make a vast impact, improving pathways, outcomes and access to medicines, and reducing the cost of healthcare.

For example, King's College Hospital's database of unstructured data enables searches to identify patients on a particular drug or with a particular diagnosis. This information can be used for managing cohorts of patients or monitoring the quality of care.

Mental health is another area that may particularly benefit, because much of the information in mental health records is unstructured. The South London and Maudsley NHS Trust has a secure environment which enables researchers to use unstructured text in their health record database, which has been particularly beneficial for mental health research.

Specifically in the pharmaceutical industry, free text may present an opportunity to bring faster and more effective treatment to market. With the use of unstructured data providing a greater evidence base, proposals for new treatments can be pushed forward. Additionally, suitable patients can be identified earlier for treatments based on this smarter use of data, as well as being much easier to carry out research and develop therapies.

Similarly, the use of free text will impact how patients are recruited for clinical trials, with the ability to completely transform this pathway. By looking at patients' records and clinical criteria, the process of identifying and matching eligible patients with trial requirements is much easier. This change in thinking becomes more inclusive, particularly to clinical trials from the point of view of the patient. There are the additional benefits of time-saving, allowing trials to progress more quickly and efficiently the first time around by having the right person in the right place.

The use of free text in electronic health records is largely discussed, but there are other potential uses of free text within social media and patient reports. There have been multiple research studies which have looked into identifying symptoms that people report on social media and detecting trends using social media and patient reports. This can help clinical teams become aware of anything new and innovative that might be developing, or perhaps, linking to our current challenging times, have the potential to spot signs and symptoms of a pandemic earlier.

Conclusion
Unstructured data is an invaluable source of information which has the potential to reveal a deeper understanding of the 'why' behind healthcare pathways and clinical decision-making. By combining the 'what' from structured data with the 'why' from unstructured data, healthcare professionals, researchers, and policy-makers will gain a much more cohesive and holistic picture of patients' diagnoses, pathways and, ultimately, help inform personalised healthcare.

With the increasing range of artificial intelligence and machine learning solutions available, these healthcare stakeholders stand to benefit more than ever by re-harnessing the use of free text. Improved access to data allows for better clinical decisions and informed patients, which as a result, has a positive impact on patient pathways now and into the future. By recognising the value of unstructured data, overcoming these existing challenges and joining databases such as THIN together, both patients and healthcare professionals can leverage the benefits of this enhanced visibility, and put patients first.

Samir Dhalla

Samir has a wealth of experience across the healthcare industry having started his early career as a pharmacist. He has worked with some of the UK's most influential Hospitals, creating efficiencies as well as new entities/ departments all with a view to improving patient care and ensuring the NHS organisations were able to be sustainable as part of a long term strategy. Today Samir sets the strategy and innovation for Cegedim Health Data in the UK as well as leading Cegedim's world-renowned healthcare database THIN; The Health Improvement Network.

Dr. Anoop Shah

Dr. Anoop Shah is a clinical academic at the Institute of Health Informatics, University College London, and a consultant in clinical pharmacology and general (internal) medicine at University College London Hospitals. He has 15 years' experience in using electronic health record databases for epidemiological research, including the use of natural language processing to extract information from the free text. He is currently undertaking a postdoctoral fellowship funded by THIS Institute, to improve the recording of problems and diagnoses in electronic health records. He is a Fellow of the Faculty of Clinical Informatics (FCI) and I lead the FCI Diagnosis Recording Special Interest Group.