Using Bidirectional Encoder Representations from Transformers (BERT) to remove personal data from healthcare records.

Ewart J Sheldon[1,3]*, Anthony Shek[2], Mohammed Al-Agil[1], Vlad Dinu[2], Sophie E Maxey[1,3], Clodagh H McGuire[1,3], Phil Davidson[1,3], James TH Teo[1]

•1: King's College Hospital, Denmark Hill, Brixton, London SE5 9RS, 2: King's College London, Strand, London WC2R 2LS, 3: The University of Manchester, Oxford Rd, Manchester M13 9PL *ewart.sheldon@nhs.net

Introduction

- Patient data is stored in electronic healthcare records (EPR) which contain large amount of data that is valuable for research
- Before this personal data can be used for secondary purposes it must be de-identified
- This is a legal requirement and a patient safety concern
- De-identification is currently done using rule-based approaches but this will miss novel situations like shortened nicknames, misspelt names, fragmented addresses
- A natural language processing method such as BERT[(1)] could help this as it looks at the context of words to identify personal data terms

De-identification concept database

- UK government guidelines[(2)] were used to specify key types of data that needed to be annotated (these comply with the NHS England confidentiality policy[(3)], and are similar to US Safe Harbour legislation):
  - Dates of Birth
  - Contact details
  - Names
  - Identifiers
  - Healthcare identifiers
- Data was associated with either:
  - Patients
  - Relatives
  - Healthcare professionals (anyone involved in the care pathway)

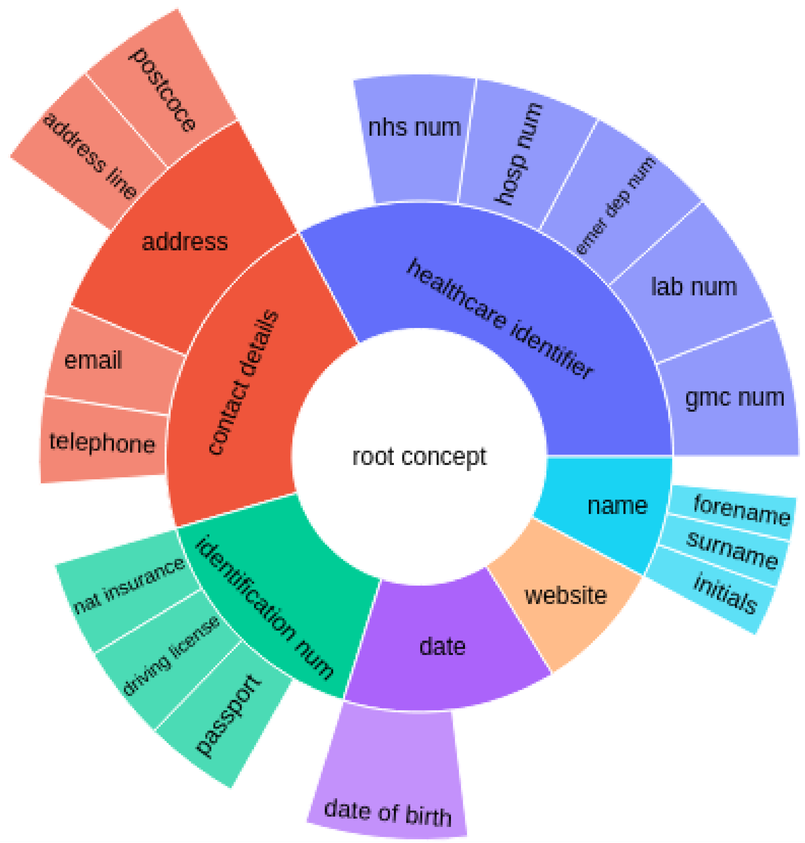
Methods

- A concept database of personal data terms was developed (Fig 1)
- MedCAT[(4)] and MedCATtrainer[(4)] was used to assist in annotating identifiable data in unstructured clinical text documents (inpatient clinical notes & outpatient letters).
- This project is Governed by the patient-led KERRI committee at Kings College Hospital which audits adequate de-identification of clinical text.

Dataset description

- 2667 Documents extracted from Cogstack[(5)]
- GP A&E Letters/Clinical note sections from 01/2016 – 12/2020 were uploaded into MedCATtrainer and annotated
- 56,128 annotations were generated
- 11,551 unique terms
- These terms related mainly to the healthcare professionals in the notes (53.95%), patients (45.88%) and the rest were relatives or careers (0.17%)
- All terms >1000 will be used to train BERT

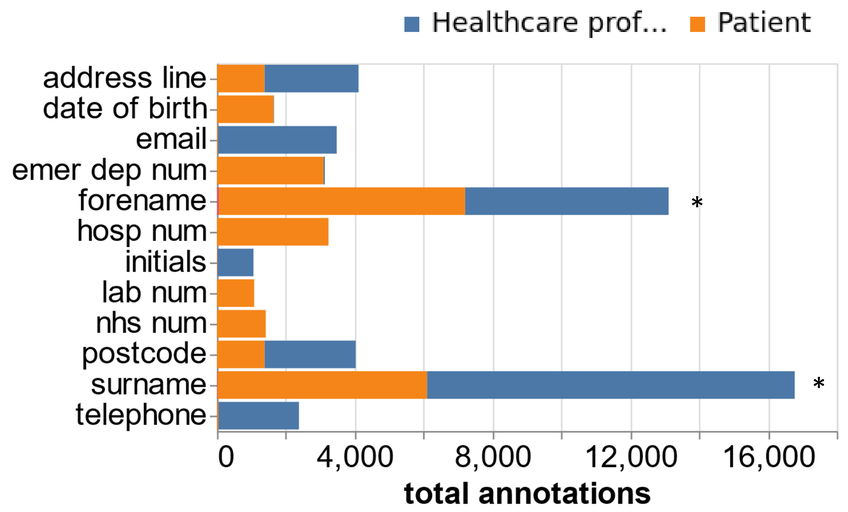
Fig 1. Diagram of the de-identification hierarchical ontology structure. Five broad terms serve as the primary child nodes to which more specific terms child nodes, e.g. postcode, were connected. The terminal nodes were used to annotate documents with patient data terms.



Current/Future work

- The annotation output has been converted into BERT NER input and BERT will be trained on the data
- When BERT has annotated documents the patient data terms will be replaced by database concept name in patient documents
- Will be used with existing rules based systems to create a multi-layer anonymization system

Fig 2. Personal data terms annotated from patient documents. The red line indicates 1000 annotations for a term. * = relative annotations in this group; there were only 88 relative annotations (20 surnames, and 68 forenames)



Conclusion

- We've been able to build a patient data concept database
- We are now training BERT to replace these terms, intext, with these concepts. To retain the contextual information of the document for future downstream information extraction tasks.

References

1. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv181004805 Cs. 2019 May 24; Available from: http://arxiv.org/abs/1810.04805
2. Information Commissioner's Office. Draft anonymisation, pseudonymisation and privacy enhancing technologies guidance. 2021 May.
3. Corporate Information Governance. NHS England » Confidentiality Policy. NHS; 2019. Available from: https://www.england.nhs.uk/publication/confidentiality-policy/
4. Kraljevic Z, Searle T, Shek A, Roguski L, Noor K, Bean D, et al. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. Artif Intell Med. 2021 Jul 1;117:102083.
5. Jackson R, Kartoglu I, Stringer C, Gorrell G, Roberts A, Song X, et al. CogStack - experiences of deploying integrated information retrieval and extraction services in a large National Health Service Foundation Trust hospital. BMC Med Inform Decis Mak. 2018 Jun 25;18(1):47.