Gloucestershire Safety & Quality Improvement Academy

Gloucestershire Hospitals NHS
NHS Foundation Trust

# Identification of IBD cohorts from linked endoscopy and histology reports using natural language processing

Jonathan L Brown
Consultant Gastroenterologist

## 1. Summary of the purpose

This service evaluation project (8622) investigated the potential for natural language processing (NLP) algorithms to comprehend free text reports within electronic patient records (EPR) to characterise a countywide inflammatory bowel disease (IBD) cohort.

## 2. Problem

Patients with IBD are likely to undergo multiple lifetime endoscopic procedures which generate histopathological reports. Managing these patients requires clinicians to derive a phenotypic overview from numerous episodes and diverse sources which can be time consuming, incomplete and subjective.

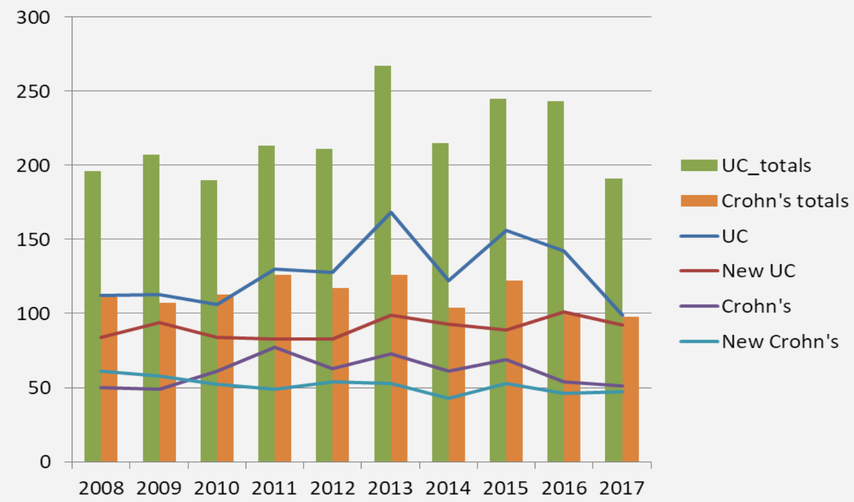
## 3. Methods

118,108 lower GI endoscopic procedure reports (2002-2017) and 62,051 lower GI histology reports (2008-2017) from GRH were pseudo-anonymised with hexadecimal GUIDs and imported into an SQL database. Text processing was undertaken in Python pandas dataframes and involved 3a) conversion to lower case, key word spelling correction, sentence tokenization and 3b) regular expression identification of diagnoses with supporting or negating text. Cycles of random sample analysis 3c) allowed for the iterative development of the regular expressions.

## 4. Results

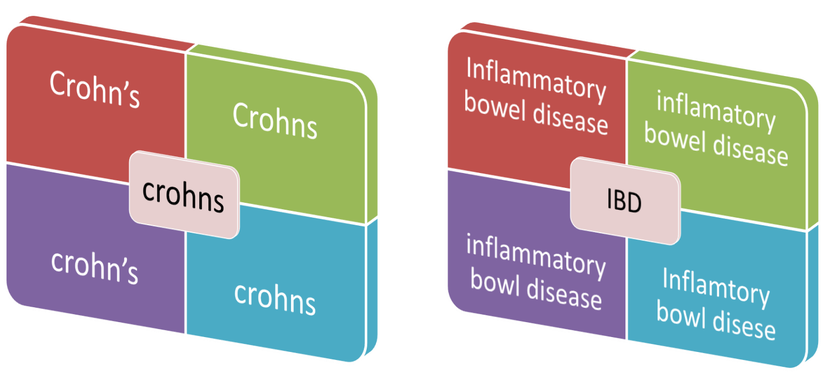
A 64 bit desktop computer took 11 minutes to identify 2119 colitis, 1166 Crohn's and 231 IBD unclassified patients. The algorithms were100% sensitive and specific at distinguishing index cases from follow up procedures, 100% sensitive at identifying IBD in linked histology reports and 98% specific at rejecting diagnoses other than IBD.

Auto-mated compu-tation of IBD endo-scopic work-load:



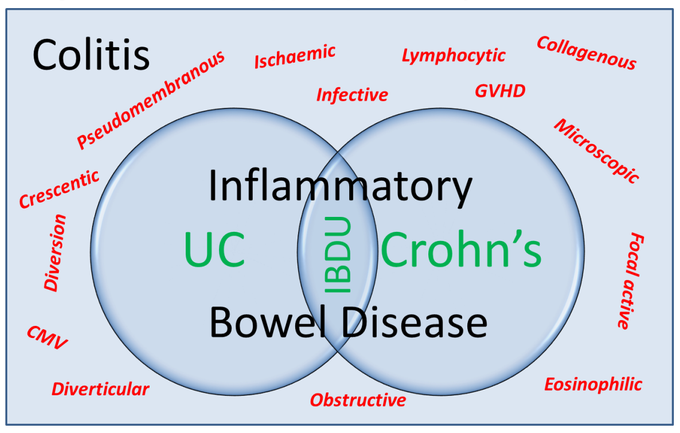

## 3a. Data cleaning

Pre-processing algorithms developed including keyword Levenshtein thresholding

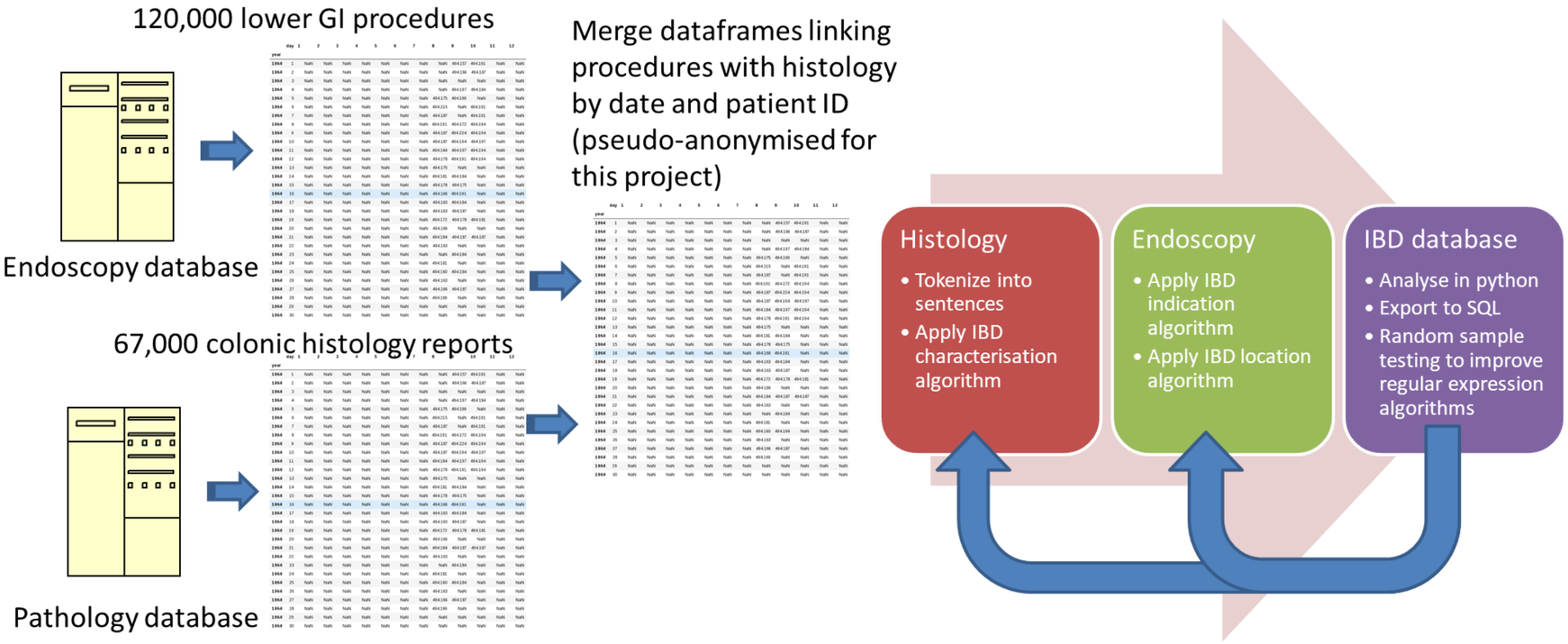



## 3b. Textual hierarchy of colitis

Regular expression algorithms developed specifically to find IBD




## 3c. Recursive regular expression evolution from random sample feedback




Random sample feedback loop comprised

- 100 UC
- 100 Crohns
- 100 IBDU
- 400 IBD excluded cases

Multiple developmental iterations required in order to optimise and derive sensitivity and specificity of algorithms

## 5. Conclusion

NLP offers a powerful tool for the automated characterisation of IBD cohorts from text in semi-structured endoscopy and histology reports. The potential for the scheduling of surveillance and linkage to other systems, such as primary care prescribing, are obvious. The technology in the context of an EPR could be applicable to many other chronic disease cohorts.

## 6. Further work

This project demonstrated the potential for NLP to integrate with an EPR and characterise disease cohorts. An evaluation within a clinical environment would be a logical progression, with linkage to other systems. The phenotypic characterisation of patients from semi-structured text complements the ambitious NHS program for treatment optimisation and prognostic determination using artificial intelligence.