BMJ Health & Care Informatics

Analysis of 'One in a Million' primary care consultation conversations using natural language processing

Yvette Pyne [1], Yik Ming Wong,[2] Haishuo Fang,[2] Edwin Simpson[2]

To cite: Pyne Y, Wong YM, Fang H, *et al*. Analysis of 'One in a Million' primary care consultation conversations using natural language processing. *BMJ Health Care Inform* 2023;**30**:e100659. doi:10.1136/bmjhci-2022-100659

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2022-100659).

Received 05 August 2022
Accepted 30 March 2023

ABSTRACT

Background Modern patient electronic health records form a core part of primary care; they contain both clinical codes and free text entered by the clinician. Natural language processing (NLP) could be employed to generate these records through 'listening' to a consultation conversation.

Objectives This study develops and assesses several text classifiers for identifying clinical codes for primary care consultations based on the doctor–patient conversation. We evaluate the possibility of training classifiers using medical code descriptions, and the benefits of processing transcribed speech from patients as well as doctors. The study also highlights steps for improving future classifiers.

Methods Using verbatim transcripts of 239 primary care consultation conversations (the 'One in a Million' dataset) and novel additional datasets for distant supervision, we trained NLP classifiers (naïve Bayes, support vector machine, nearest centroid, a conventional BERT classifier and few-shot BERT approaches) to identify the International Classification of Primary Care-2 clinical codes associated with each consultation.

Results Of all models tested, a fine-tuned BERT classifier was the best performer. Distant supervision improved the model's performance (F1 score over 16 classes) from 0.45 with conventional supervision with 191 labelled transcripts to 0.51. Incorporating patients' speech in addition to clinician's speech increased the BERT classifier's performance from 0.45 to 0.55 F1 (p=0.01, paired bootstrap test).

Conclusions Our findings demonstrate that NLP classifiers can be trained to identify clinical area(s) being discussed in a primary care consultation from audio transcriptions; this could represent an important step towards a smart digital assistant in the consultation room.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Natural language processing (NLP) has the potential to revolutionise clinical specialties that rely on free text such as primary care which extensively uses electronic health records.
⇒ Existing NLP tools are focused on classifying free text created by health professionals or generating free text from predefined clinical data.
⇒ The creation of a tool to classify a clinical consultation based on the conversation that occurs in it could have a significant positive effect on clinician workload and could form part of the tools used in an 'augmented consultation'.

WHAT THIS STUDY ADDS

⇒ This study is the first to analyse and classify primary care consultations from the conversations that took place between doctors and patients.
⇒ This study develops and assesses the efficacy of several NLP classifiers, including recent pretrained deep neural networks, for classifying verbatim medical conversation transcripts, which use very different language to clinical notes, and for which extremely limited training data are available.
⇒ This study identifies limitations of the existing healthcare datasets and tools containing primary care free text and makes recommendations for further avenues of research and appropriate data sources.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study highlights the importance of building datasets of clinical conversations and other healthcare-based natural language sources for use in clinical research.
⇒ This study suggests several further research topics combining the fields of clinical primary care and machine learning.

INTRODUCTION

Technology is becoming increasingly pervasive in primary care[1] and a significant proportion of a clinician's day is spent interacting with the patient electronic health record (EHR). EHRs are a form of 'handover', either to another health professional, or to the same clinician when they meet the patient again; the records also provide key evidence in legal cases and are used for performance targets (such as the UK National Health Service Quality and Outcomes Framework) and billing (in the USA). EHRs incorporate free text and clinical codes such as SNOMED-CT, ICD (International Classification of Diseases) or Read codes. Historically, EHRs have been for clinicians only, but incoming UK legislation will open these records to be viewed by patients as well. For all these reasons, it is vital that clinical notes and their associated codes are accurate and complete.

Check for updates

[1]Bristol Medical School, University of Bristol Centre for Academic Primary Care, Bristol, UK
[2]Intelligent Systems Labs, University of Bristol, Bristol, UK

Correspondence to
Dr Yvette Pyne;
yvette@digitalgp.net

Modern EHR systems used in UK primary care (such as EMIS, SystmOne and Vision) can also direct clinicians to clinically relevant local or national guidelines if the clinician enters an appropriate clinical code. However, clinical codes are often associated with the diagnosis rather than the presenting complaint so may only be entered at the conclusion of the consultation or even after the patient has left. Writing EHR notes or entering clinical codes during a consultation can be disruptive as the clinician has to focus on data capture rather than the patient.[2 3] Motivated by this, we investigated the first steps towards a natural language processing (NLP)[4] application that can 'listen' to a conversation between general practitioner (GP) and patient and automatically recommend clinical codes.

NLP has previously been applied to healthcare in a wide range of applications; for example, to process and analyse patient feedback,[5] identify risk factors,[6] symptoms and treatments,[7] or suspected disease[8] from clinical notes, or even to generate notes automatically from structured hospital data.[9] The technology to transcribe speech to text already exists in tools such as 'Otter.ai',[10] which could enable text processing of clinical conversations. However, the systematic evaluation of the use of NLP for interpreting conversations between clinicians and patients is lacking.[11–14]

We treated the task of assigning clinical codes to transcripts as text classification, which can be addressed using supervised learning. However, training data are in short supply, and recent NLP approaches based on deep learning are data hungry. This research assessed a series of text classifiers trained with small datasets to identify clinical codes associated with real-life GP–patient consultations. Our objectives were to evaluate: (1) the performance of different kinds of text classifiers; (2) the effect of training classifiers using existing medical code descriptions rather than example consultations; (3) the contribution of patients' speech to correct classifications in addition to the clinician's speech and (4) opportunities for improving the classifiers in future.

METHODS

Data sources

'One in a Million' dataset

The 'One in a Million' (OIAM) dataset[15] contains 300 video and audio recordings and verbatim transcripts of real clinical consultations conducted in 12 GP practices around Bristol in English with adult patients with permission in place for reuse. These consultations are associated with one or more International Classification of Primary Care (ICPC-2) clinical problem codes assigned by human coders. Both anonymised transcripts and ICPC-2 codes were available for 239 consultations.[16] A fictional but representative part of a consultation transcript is shown in online supplemental appendix A.

ICPC-2: ICPC-2 code descriptions

This is a primary care focused set of approximately 1300 low-level codes related to clinical problems that are grouped into 17 high level chapters or codes associated with clinical problem areas such as 'urinary' or 'circulatory'.[17] The ICPC-2e-V.7.0 comma separate values file[18] was used to create a data dictionary of high-level codes associated with relevant words for that group of conditions.

National Institute for Health and Care Clinical Knowledge Summaries

We created a National Institute for Health and Care Clinical Knowledge Summaries (NICE CKS) 'Health Topics' dataset using the 'Web Scraper.io' Google Chrome extension on 29 July 2021 from the publicly available web resource covering over 370 clinical topics.[19] For each health topic, we considered text from sections: 'Causes', 'Definition', 'Diagnosis', 'Clinical features', 'History', 'Presentation', 'Signs and symptoms' and 'When to suspect'. The clinical author mapped each NICE CKS topic to one or more ICPC-2 codes (see online supplemental appendix B: ICPC-2 codes and consultations). Then, for each ICPC-2 code, all the related CKS health topics were concatenated into a single document corresponding to that ICPC-2 code. While the ICPC-2 descriptions contain lists of relevant keywords, CKS health topics contain complete sentences that may convey additional information such as descriptions of symptoms.

Training the NLP classifiers

We initially used the OIAM dataset to train and test a series of classifiers using standard supervised learning (objective (1)). We held out a stratified sample of 20% (48 transcripts) of OIAM as a test set, using the remainder (191 transcripts) for training. Hyperparameter tuning was performed using fivefold cross-validation on the training split (see online supplemental appendix D).

Supervised learning requires a training dataset containing sufficiently representative examples for each class label, yet our training set contains only a small number of example consultations per code. We, therefore, introduced a second approach, 'distant supervision', that used the ICPC-2 code descriptions and NICE CKS datasets as training examples and tested the classifiers on the OIAM dataset (objective 2). We also tested excluding the 'A: General' classification as it includes a wide spectrum of clinical conditions from 'pain general/ multiple sites' to 'viral disease other', and thus assigning the code was unlikely to aid GPs and may confuse the classifiers. Finally, we analysed distant supervision performance considering only the GP's half of the conversation to determine whether transcribing patient's speech is beneficial (objective 3).

To assess classifier performance, we used the macroaverage precision (equivalent to positive predicted value; the fraction of labels assigned by the classifier that were correct), recall (also called 'sensitivity'; the fraction of true labels predicted by the classifier) and F1 score (the harmonic mean of precision and recall).

As a baseline, we assigned labels at random, allowing multiple labels per transcript. We tested shallow, data-efficient classifiers: naïve Bayes (NB), as a linear classifier; support vector machine (SVM) with RBF kernel, a non-linear classifier that performs well with high dimensional feature vectors, such as those used to represent text (see below); and nearest centroid with Euclidean distance as a lightweight clustering-based classifier. While there are many other alternatives, the chosen methods represent broad types of classifier and allowed us to determine the suitability of classifiers with increasing complexity (part of objective (1)). The NB and SVM classifiers were run in 'multilabel' and 'multiclass' classification modes:

- Multilabel: for each possible ICPC-2 code, we train a binary classifier to assign either 'yes' or 'no' per consultation, so that more than one code can be assigned to the consultation.
- Multiclass: we train one classifier to assign the single most likely ICPC-2 code to the consultation. In training, we select the first code for each consultation, with codes sorted alphabetically.

In both modes, classifiers were evaluated on the complete test dataset using the same metrics. For consultations with more than one ICPC-2 code, the correct set of labels must be predicted to achieve perfect recall. As there are 110 consultations with more than one label, this puts a ceiling on the recall of the multiclass approach. However, the training data are more balanced, which may lead to better recall than the multilabel setup, where the training data for each binary classifier contains only a small minority of positive examples. Precision could also be higher as the multiclass mode directly compares classes that are easy to confuse.

For the shallow classifiers, we removed stopwords from the consultation transcripts before processing them. Considering the choice of stopwords as part of objective (1), we tested 3 sets: 318 'English' stopwords (from sklearn's default ENGLISH_STOP_WORDS); 203 'medical' stopwords[20] and 61 'custom' stopwords (see online supplemental appendix C: custom stopword dictionary). We encoded each transcript, ICPC-2 code description and CKS health topic as a feature vector containing the counts of the 5000 most frequent unigrams (individual words) and bigrams (consecutive pairs of words).

We also trialled recent deep learning classifiers that leverage a pretrained transformer, PubMedBERT,[21] a variant of BERT[22] that was pretrained on biomedical text (objective (1)). PubMedBERT encodes text into dense vector representations that take word order into account and include medical terms not present in our training examples. We tested a 'conventional BERT' classifier, in which we fine-tuned a classification head on top of PubMedBERT (multiclass mode). For distant supervision, we compared this to two BERT setups designed for training with very few examples: using next sentence prediction (NSP) to compare the text to a prompt containing the name of each class (multilabel mode); and using masked language modelling (MLM) to predict the category name by filling in the blank word in a prompt (multiclass)[23] ; both used 'this is a problem of ___' as a prompt. We hypothesised that the BERT approaches would outperform shallow classifiers thanks to their pretrained language representations, and that MLM would perform best as it reuses the pretraining task, so does not need to learn new classifier layers from scratch. Since BERT has a length limit of 512 tokens, transcripts and CKS topics were broken into multiple documents consisting of complete sentences. For training, all chunks were assigned the corresponding ICPC-2 training label. For prediction, we took the union of labels predicted for each of the chunks.

RESULTS

The consultation and patient demographics for the OIAM dataset are given in table 1, and the number of transcripts with multiple labels is shown in figure 1.

Objective (1): types of NLP classifiers

Table 2 shows the results for classifiers trained on OIAM transcript texts, with best performances highlighted in bold. As the held-out test set is small, we include the results of fivefold cross-validation over the larger training set. Nearest centroid is the best shallow classifier. Multiclass NB clearly outperforms SVM, while BERT provides substantial improvements all round. Compared with multilabel mode, multiclass classifiers have higher precision. However, recall and F1 are lower for multiclass SVM, while they are higher for multiclass NB, despite being unable to assign multiple codes to a single transcript. The baseline slightly outperforms multilabel NB on the test set and is competitive with some other shallow methods.

A comparison of F1 scores with different stopwords is shown in table 3, with the best choice for each classifier in bold, corresponding to the results in Table 2. Removing English or medical stopwords is helpful, while removing the words in all three stopword lists is most effective.

Objective (2): distant supervision

Table 4 compares F1 scores for different stopword lists with distant supervision. With CKS, the combined list is again most effective, but medical stopword removal is detrimental with ICPC-2 descriptions. Since ICPC-2 descriptions contain keywords rather than prose, any medical stopwords included by the authors of the descriptions may be part of informative key phrases that should not be removed.

Table 5 compares performance on the OIAM training set using distant supervision with the ICPC-2 code descriptions and NICE CKS topics. NB performs best with ICPC-2 supervision, in this case outperforming nearest centroid. BERT does not match the performance of NB multiclass on this small training set and conventional BERT fails to learn at all. BERT variants perform better with CKS than ICPC-2 as PubMedBERT was pretrained to process prose, rather than keywords. Combining both distant

Table 1 Details of the OIAM dataset used in this work, with patient information for the complete dataset

ICPC-2 code	No of transcripts	%
A: General	14	5.9
B: Blood, blood forming	8	3.3
D: Digestive	44	18.4
F: Eye	5	2.1
H: Ear	11	4.6
K: Circulatory	32	13.4
L: Musculoskeletal	65	27.2
N: Neurological	20	8.4
P: Psychological	50	20.9
R: Respiratory	37	15.5
S: Skin	32	13.4
T: Metabolic, endocrine, nutritional	24	10.0
U: Urinary	18	7.5
W: Pregnancy, family planning	11	4.6
X: Female genital	14	5.9
Y: Male genital	7	2.9
Total ICPC-2 code labels	392	164
Total unique consultations	239	100
No of ICPC-2 codes assigned to a consultation (see figure 1)		
0	2	1
1	128	53
2	62	26
3	40	17
4+	8	3
Duration (minutes)		
<5	13	5.4
5–10	79	33.1
10–15	82	34.3
15–20	52	21.8
20–35	13	5.4
Dataset statistics below are for the original patient sample of N=334.[16] This information was not available to compute for the N=239 subset in our experiments	No of patients	%
Sex		
Female	212	63.5
Male	122	36.5
Age		
18–34	91	27.2
35–54	94	28.1
55–74	99	29.6
≥75	36	10.8
Not reported	14	4.2
Ethnic group		
White	291	87.1
Other	43	12.9
IMD (Indices of Multiple Deprivation) quintile		
1st (least deprived)	106	31.7
2nd	54	16.2
3rd	35	10.5
4th	53	15.9
5th (most deprived)	84	25.1
Data unavailable	2	0.6

ICPC-2, International Classification of Primary Care-2; OIAM, One in a Million.

supervision sources does not improve performance for any of the methods (table 6).

Table 2 also shows that removing the option of assigning class A causes a collapse in performance with BERT NSP and MLM with ICPC-2 descriptions, and nearest centroid with either supervision source, while NB is improved slightly.

Table 2 shows test set performance with the most successful distant supervision source for each classifier. In comparison with standard supervision, the performance improves substantially for most classifiers, validating the use of external sources for distant supervision.

The NB model allows direct interpretation of the important features for classification. The wordclouds in figure 2 show the unigrams and bigrams for each class, weighted by the probability of the class given the feature, as learnt by NB (multiclass) from ICPC-2 descriptions. The informative features correspond well with medical terms in each category, but we do not see colloquial terms that may be used in conversation, or expressions longer than two tokens. Therefore, classifiers may benefit from augmenting ICPC-2 descriptions with alternative terms and phrases (objective 4, future improvements).

Objective (3): contribution of patient speech transcripts

Table 6 shows that using only the GP's part of the transcript reduces performance of most classifiers, indicating that patients provide useful information that is not contained in the GP's speech. The CIs only indicate strong evidence of a performance difference for BERT conventional, hence the finding may require investigation with a larger dataset.

DISCUSSION

We evaluated a range of text classifiers, achieving the highest F1 score on the test set of 0.51 for conventional

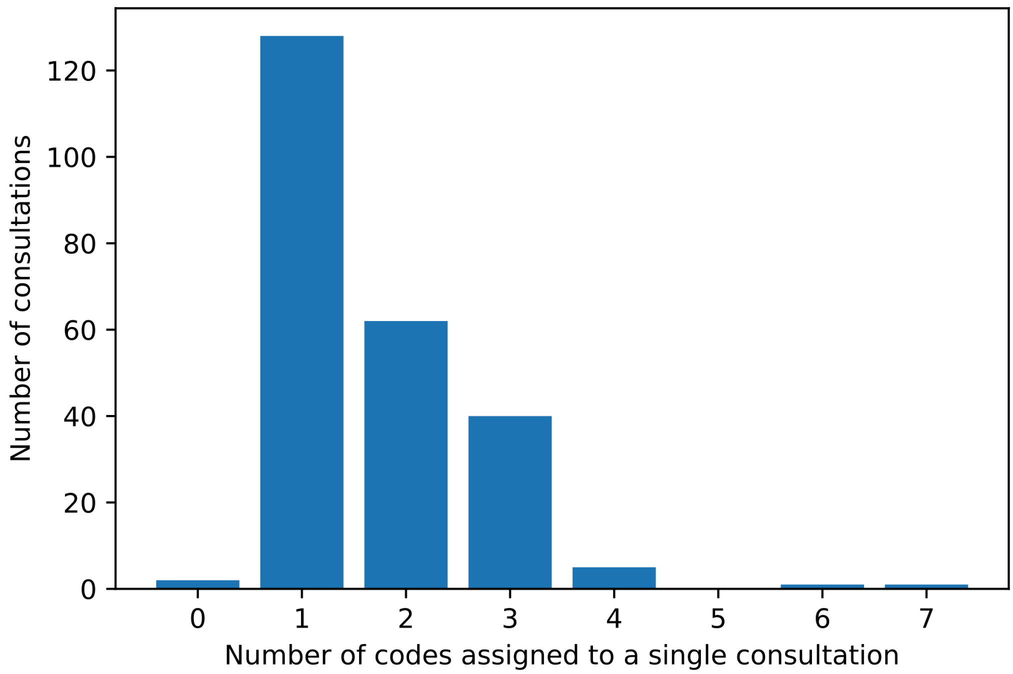

Figure 1 Distribution of consultations with multiple labels.

BERT, with recall at 56% and precision at 55%, substantially better than n-gram-based classifiers (objective 1). This classifier was trained on medical code descriptions, which outperformed standard supervision with a training set of 191 transcripts (those with no missing data such as codes, transcripts or notes) with F1=0.45 (objective 2). When patients' speech transcripts were excluded, the performance also dropped from F1=0.55 to 0.45 showing that is beneficial to capture the complete conversation (objective 3). Below, we identify specific ways to further improve the classifiers (objective 4).

More work is required to determine whether classifiers with this level of performance could usefully assist clinicians. Our scores are at the lower end of results for comparable multiclass text categorisation tasks,[24] which achieved between 53% and 86% average accuracy using a RoBERTa classifier with 100 training examples, and substantially lower than BERT for intent classification on dialogue benchmarks,[25] which achieves almost 93% accuracy with 10 training examples. Future work could, therefore, draw on these related tasks to identify improvements to the classifiers.

NB was competitive with BERT suggesting that unigrams and bigrams provide strong signals about health topics, and that datasets on the scale of OIAM may be insufficient to make full use of deep models. Against our expectations, conventional BERT was marginally the strongest, outperforming BERT MLM on the test set. The BERT models are costly to run (several hours GPU training for all BERT variants vs a few seconds with NB; testing takes in around 100 times longer), although this may not be an issue if training is performed only once before deploying the model. Future work could investigate replacing PubMedBERT with other domain-specific pretrained models (such as BioBERT[26] and ClinicalBERT[27]). Extremely large language models (LLMs) may also offer improved few-shot learning, although extensive prompt engineering is required and computational costs are huge. These LLMs could potentially generate explanations of their decisions that could bring relevant parts of the conversation to a doctor's attention.

The multilabel classifiers did less well than the multiclass classifiers, possibly because their training data was highly imbalanced (harming recall) or because multiple labels were assigned in cases where only one of the labels should have been chosen (hurting precision). However, given the complexity and breadth of primary care consultations, any effective classifier would need to be able to suggest multiple medical areas, so multilabel methods must be a focus for future research.

Given the low numbers of examples of some codes (eg, only five consultations were coded as 'F: eye'), overfitting was an issue for supervised learning, with higher performance on the training set than the validation and test sets. Distant supervision with the NICE CKS Health Topics and ICPC-2 Code descriptions demonstrated clear improvements. The key phrases in the ICPC-2 descriptions are a natural fit for NB: these features are individually informative, which allows linear models such as NB to perform well. The imperfect mapping between CKS topics and ICPC-2 codes may reduce the performance of NB on CKS topics. Improving the mapping would require

Table 2 Performance with standard supervised learning, 95% CIs shown in parentheses

	Validation			Train	Test		
Model	**Precision**	**Recall**	**F1**	**F1**	**Precision**	**Recall**	**F1**
Random baseline	0.499	0.104	0.161	0.161	0.501	0.102	0.158
Conventional supervision							
Naïve Bayes (multilabel)	0.284 (0.237 to 0.325)	0.140 (0.135 to 0.192)	0.175 (0.161 to 0.222)	**0.999**	0.234 (0.169 to 0.276)	0.113 (0.087 to 0.158)	0.139 (0.106 to 0.185)
Naïve Bayes (multiclass)	0.372 (0.298 to 0.399)	0.327 (0.314 to 0.398)	0.300 (0.266 to 0.342)	0.696	0.178 (0.146 to 0.232)	0.238 (0.213 to 0.294)	0.181 (0.154 to 0.226)
SVM (multilabel)	0.107 (0.112 to 0.132)	**1.000** (1.000 to 1.000)	0.184 (0.192 to 0.223)	0.181	0.102 (0.095 to 0.124)	1.000 (1.000 to 1.000)	0.177 (0.166 to 0.211)
SVM (multiclass)	0.200 (0.171 to 0.244)	0.159 (0.157 to 0.211)	0.154 (0.142 to 0.196)	0.696	0.217 (0.145 to 0.263)	0.169 (0.14 to 0.227)	0.164 (0.129 to 0.213)
Nearest centroid (multiclass)	0.349 (0.297 to 0.395)	0.270 (0.254 to 0.327)	0.278 (0.247 to 0.325)	0.694	0.307 (0.18 to 0.355)	0.205 (0.15 to 0.276)	0.219 (0.151 to 0.278)
BERT conventional (multiclass)	**0.467** (0.434 to 0.549)	0.577 (0.546 to 0.654)	**0.480** (0.447 to 0.550)	0.696	**0.484** (0.414 to 0.575)	**0.509** (0.434 to 0.610)	**0.452** (0.390 to 0.525)
Distant supervision							
Naïve Bayes (multilabel), ICPC-2	0.626 (0.515 to 0.687)	0.234 (0.196 to 0.278)	0.323 (0.268 to 0.362)	0.979	0.590 (0.427 to 0.656)	0.285 (0.206 to 0.384)	0.378 (0.274 to 0.456)
Naïve Bayes (multiclass), ICPC-2	0.516 (0.466 to 0.569)	0.590 (0.541 to 0.639)	0.512 (0.462 to 0.549)	1.00	0.511 (0.412 to 0.611)	0.524 (0.449 to 0.628)	0.481 (0.404 to 0.567)
Nearest centroid, ICPC-2	**0.718** (0.565 to 0.765)	0.416 (0.373 to 0.463)	0.444 (0.384 to 0.489)	1.00	0.520 (0.400 to 0.615)	0.362 (0.298 to 0.448)	0.386 (0.303 to 0.467)
Conventional BERT, CKS	0.603 (0.553 to 0.653)	0.584 (0.53 to 0.64)	0.550 (0.494 to 0.593)	0.927	**0.551** (0.477 to 0.649)	0.562 (0.483 to 0.691)	**0.508** (0.429 to 0.594)
BERT NSP, CKS	0.364 (0.333 to 0.394)	**0.816** (0.767 to 0.865)	0.462 (0.424 to 0.488)	0.291	0.257 (0.215 to 0.331)	**0.598** (0.525 to 0.711)	0.306 (0.257 to 0.371)
BERT MLM, CKS	0.600 (0.547 to 0.64)	0.615 (0.566 to 0.673)	**0.567** (0.512 to 0.604)	0.792	0.481 (0.409 to 0.574)	0.536 (0.469 to 0.639)	0.467 (0.397 to 0.548)

For conventional supervision, 'train' and 'test' results are for classifiers trained on the whole 80% training split, and validation was performed using 5-fold cross-validation over the training set.
For distant supervision, the OIAM training set was repurposed as a validation set, as it was not used to train the models with this setup.
CKS, Clinical Knowledge Summaries; ICPC-2, International Classification of Primary Care-2; MLM, masked language modelling; NSP, next sentence prediction; OIAM, One in a Million; SVM, support vector machine.

Table 3 F1 scores for fivefold cross-validation performance on the OIAM training set with different sets of stopwords

Model	No removal	English	Medical	Custom	Medical+custom	English+custom	English+medical +custom
Naïve Bayes (multilabel)	0.157	0.159	0.154	0.143	0.166	0.170	**0.175**
Naïve Bayes (multiclass)	0.225	0.266	0.243	0.228	0.245	0.272	**0.300**
SVM (multilabel)	0.184	0.184	0.184	0.184	0.184	0.184	0.184
SVM (multiclass)	0.141	0.151	0.141	0.142	0.142	0.150	**0.154**
Nearest centroid	0.234	0.256	0.239	0.234	0.247	0.252	**0.278**

OIAM, One in a Million; SVM, support vector machine.

Table 4 F1 scores for distant supervision performance, evaluated on the OIAM training set, with different sets of stopwords, and training on either CKS topics or ICPC-2 descriptions

Model	No removal	English	Medical	Custom	Medical+custom	English+custom	English+medical+custom
NB (multilabel), ICPC-2	0.139	0.170	0.136	0.253	0.297	**0.323**	0.297
NB (multilabel), CKS	0.096	0.160	0.119	0.126	0.191	0.207	**0.234**
NB (multiclass), ICPC-2	0.324	0.354	0.307	0.461	0.471	**0.512**	0.470
NB (multiclass), CKS	0.245	0.274	0.249	0.275	0.340	0.368	**0.375**
Nearest centroid, ICPC-2	0.312	0.354	0.317	0.432	0.437	**0.445**	0.437
Nearest centroid, CKS	0.326	0.349	0.344	0.349	0.353	0.357	**0.365**

CKS, Clinical Knowledge Summaries; ICPC-2, International Classification of Primary Care-2; NB, Naïve Bayes; OIAM, One in a Million.

Table 5 F1 scores for different sources of distant supervision, and the effect of removing class A, evaluated on the OIAM training set

Model	ICPC-2	ICPC-2 without A	CKS	CKS without A	ICPC-2 and CKS combined	Combined without A
Naïve Bayes (multilabel)	0.323 (0.268, 0.362)	**0.345** (0.286, 0.389)	0.234 (0.196, 0.262)	0.249 (0.207, 0.285)	0.254 (0.21, 0.287)	0.271 (0.225, 0.308)
Naïve Bayes (multiclass)	**0.512** (0.462, 0.549)	0.508 (0.458, 0.546)	0.375 (0.325, 0.411)	0.391 (0.34, 0.428)	0.378 (0.33, 0.416)	0.385 (0.338, 0.421)
Nearest centroid	0.444 (0.384, 0.489)	0.093 (0.063, 0.12)	0.365 (0.312, 0.401)	0.086 (0.057, 0.107)	0.367 (0.315, 0.403)	0.090 (0.063, 0.113)
BERT conventional	0.057 (0.049, 0.065)	0.027 (0.02, 0.037)	0.550 (0.494, 0.593)	**0.521** (0.459, 0.565)	**0.540** (0.476, 0.576)	**0.545** (0.483, 0.590)
BERT NSP	0.285 (0.232, 0.324)	0.347 (0.309, 0.371)	0.462 (0.424, 0.488)	0.434 (0.392, 0.466)	0.445 (0.402, 0.476)	0.467 (0.425, 0.498)
BERT MLM	0.505 (0.444, 0.544)	0.486 (0.425, 0.528)	**0.567** (0.512, 0.604)	0.497 (0.441, 0.535)	0.532 (0.472, 0.571)	0.475 (0.424, 0.512)

Highest F1 scores in bold.
CKS, Clinical Knowledge Summaries; ICPC-2, International Classification of Primary Care-2; MLM, masked language modelling; NSP, next sentence prediction; OIAM, One in a Million.

costly manual editing of the scraped CKS health topics, as some CKS topics lack a one-to-one mapping to an ICPC-2 code. Still, CKS topics produce competitive performance with BERT, which was pretrained with complete sentences, suggesting that the health topics do include useful training signals. Future work could, therefore, investigate ensembles that stack[28] models trained with different sources of data.

To identify common classifier mistakes, the clinician on the research team reviewed individual consultation transcripts and their human and predicted codes and noted several distinct types of errors. First, shallow classifiers demonstrated simple linguistic errors such as misunderstanding idioms. In one consultation, the GP repeatedly mentioned 'keeping an eye on it' and the NB classifier incorrectly coded it as an ophthalmology-related consultation; BERT overcame this by avoiding reliance on isolated words as features.[29] Second, perusing specific consultations where the NLP classifier appeared to get the coding significantly wrong highlighted errors by the original human labelling team. Third, the 'A: General' category was often selected erroneously, as the class is non-specific (precision=0.154 for NB multiclass, trained on ICPC-2 descriptions), although excluding this class often hurt performance. Finally, there were examples where a lack of clinical knowledge caused errors such as the NLP classifier assuming that a consultation discussing someone's wrist was a musculoskeletal rather than a neurological issue (such as in carpal tunnel syndrome).

Many of these specific types of error relate to limitations of the dataset: its scale, labelling quality and labelling scheme; we consider its small size to be the most significant issue. When scaling up the dataset, further limitations to address include the dataset being only in English and all the consultations taking place in one part of the UK. The current areas where clinical machine learning is excelling are radiology and pathology due to their large and accessible (anonymised) datasets, and the creation of a large, anonymised, free text dataset related to primary care would be hugely valuable for research. The COVID-19 pandemic accelerated the use of online consultations producing potential sources of patient-entered free text (eg, AskMyGP[30]) and recorded audio/video consultations for examination (eg, by FourteenFish[31]). We advocate for routinely incorporating consent to use digitally recorded clinical consultations for research and providing robust anonymisation of them, so that researchers can conduct valuable and translational research in this area.

Further directions for future research include processing the consultations in 'real-time' and assigning them to the more fine-grained NICE CKS health topics rather than ICPC-2 codes, which would allow the system to link a doctor automatically to the corresponding health topic guidelines. Performance may also be improved by combining text with other data from electronic medical records.

Table 6 F1 scores when patients' transcribed speech is excluded

Model	Including GP and patient speech	Only GP speech
Naïve Bayes (multilabel) ICPC-2	0.323 (0.268, 0.362)	**0.372** (0.3, 0.417)
Naïve Bayes (multiclass) ICPC-2	**0.512** (0.462, 0.549)	0.484 (0.429, 0.521)
Nearest centroid ICPC-2	**0.444** (0.384, 0.489)	0.425 (0.361, 0.47)
BERT conventional, CKS	**0.550** (0.494, 0.593)	0.445 (0.384, 0.465)
BERT NSP, CKS	**0.462** (0.424, 0.488)	0.436 (0.398, 0.464)
BERT MLM, CKS	**0.567** (0.512, 0.604)	0.500 (0.434, 0.539)

The classifiers were trained using their most effective distant supervision source and evaluated on the OIAM training set (repurposed as a validation set). Bold indicates best performance in a comparison between including and excluding patients' speech with the same classifier.
CKS, Clinical Knowledge Summaries; GP, general practitioner; ICPC-2, International Classification of Primary Care-2; MLM, masked language modelling; NSP, next sentence prediction; OIAM, One in a Million.

CONCLUSION

This paper offers a promising avenue of research using NLP to extract information from the conversation between a patient and their doctor in a primary care consultation and demonstrates a successful collaboration between clinical and computing disciplines. Previous projects using NLP in a clinical setting have focused on classifying free text created by health professionals (such as radiology reports) or generating free text from codes and defined data (such as investigation results). To our knowledge, this is the first time that the original conversation between a doctor and their patient has been analysed using NLP. Our comparison of text classifiers showed modest gains from deep learning approaches, that the models can be trained using health topics scraped from web pages, and that patients' speech contains valuable signals for assigning medical codes. We identified potential improvements, including adding colloquial vocabulary to health topic descriptions, increasing the dataset size and domain-specific pretraining of language models. Our ultimate goal would be to provide a smart digital assistant that can create effective consultation notes and suggest questions or guidelines to the clinician[32]; this is likely to require significant advances both in NLP and in our understanding of what makes good clinical notes. While this goal is still a long way off, our work represents one small step towards that reality.

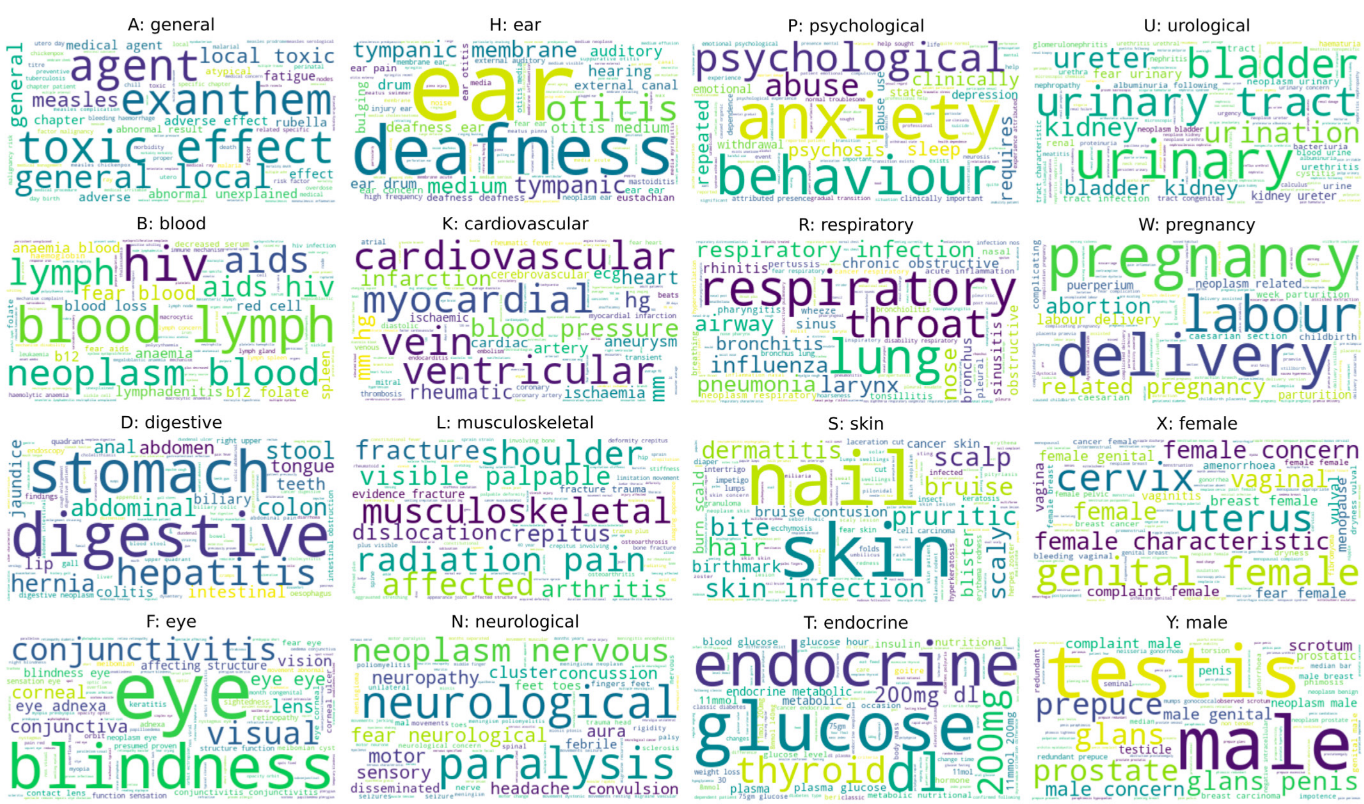

Figure 2 Wordclouds for each ICPC-2 category, with unigrams and bigrams weighted by the probability of the class label given the feature. ICPC-2, International Classification of Primary Care.

Acknowledgements This work was carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol—http://www.bristol.ac.uk/acrc/. The authors would like to thank Professor John Macleod for his support during this project.

Contributors YP and ES contributed to the conception and design of the work, the sourcing of the data and the analysis of the results. YP and ES wrote the paper with review and comments by YMW and HF. YMW, HF and ES wrote the code associated with the project. YP acts as guarantor for the work and conduct of the study, had access to the data, and controlled the decision to publish.

Funding YP is funded via an NIHR Academic Clinical Fellowship in Primary Care—Award Identifier: ACF-2018-25-007. ES is employed by the University of Bristol as a Lecturer. He has no other funding streams. YMW and HF were funded by an Elizabeth Blackwell Institute (EBI) Wellcome Grant: 204813/Z/16/Z. The 'One in a Million' study was funded by the National Institute for Health Research (NIHR) School for Primary Care Research (208) and the South West GP Trust.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The project operated under South West—Central Bristol Research Ethics Committee (Reference: 21/SW/0069) approval granted on 3 June 2021.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement The 'One in a Million' dataset is available for research use following valid ethics approval.ICPC-2 Codes and descriptions are freely downloadable from the web.The NICE CKS Health Topics dataset is freely downloadable from their website using freely available 'web-scraping' tools.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD

Yvette Pyne http://orcid.org/0000-0001-5920-484X

REFERENCES

1. Topol EJ. The topol review: preparing the healthcare workforce to deliver the digital future. 2019.
2. Young RA, Burge SK, Kumar KA, *et al*. A time-motion study of primary care physicians' work in the electronic health record era. *Fam Med* 2018;50:91–9.
3. Sinsky C, Colligan L, Li L, *et al*. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Ann Intern Med* 2016;165:753–60.
4. Yim W-W, Yetisgen M, Harris WP, *et al*. Natural language processing in oncology: A review. *JAMA Oncol* 2016;2:797–804.
5. Khanbhai M, Anyadi P, Symons J, *et al*. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. *BMJ Health Care Inform* 2021;28:e100262.
6. Khalifa A, Meystre S. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. *J Biomed Inform* 2015;58 Suppl(Suppl):S128–32.
7. Koleck TA, Dreisbach C, Bourne PE, *et al*. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J Am Med Inform Assoc* 2019;26:364–79.
8. Doan S, Maehara CK, Chaparro JD, *et al*. Building a natural language processing tool to identify patients with high clinical suspicion for kawasaki disease from emergency department notes. *Acad Emerg Med* 2016;23:628–36.
9. Moen H, Peltonen L-M, Heimonen J, *et al*. Comparison of automatic summarisation methods for clinical free text notes. *Artif Intell Med* 2016;67:25–37.
10. Corrente M, Bourgeault I. *Innovation in transcribing data: meet otter. ai*. 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom, 2022.
11. How robin works. Available: https://www.robinhealthcare.com/how-robin-works [Accessed 14 Mar 2023].
12. Quiroz JC, Laranjo L, Kocaballi AB, *et al*. Challenges of developing a digital scribe to reduce clinical documentation burden. *NPJ Digit Med* 2019;2:114.
13. van Buchem MM, Boosman H, Bauer MP, *et al*. The digital scribe in clinical practice: a scoping review and research agenda. *NPJ Digit Med* 2021;4:57.
14. Krishna K, Pavel A, Schloss B, *et al*. Extracting structured data from physician-patient conversations by predicting noteworthy utterances. 2020.
15. Barnes R. One in A million: A study of primary care consultations. 2017.
16. Jepson M, Salisbury C, Ridd MJ, *et al*. The "one in a million" study: creating a database of uk primary care consultations. *Br J Gen Pract* 2017;67:e345–51.
17. World Organization of National Colleges A and Academic Associations of General Practitioners, Family Physicians, Classification Committee. *International classification of primary care: ICPC-2*. Oxford: Oxford Univ. Press, 1998.
18. ICPC-2e – english version. ehelse. Available: https://www.ehelse.no/kodeverk/icpc-2e--english-version [Accessed 5 May 2022].
19. NICE. Health topics A to Z | CKS | NICE. Available: https://cks.nice.org.uk/topics/ [Accessed 11 Feb 2022].
20. Bobo WV, Pathak J, Kremers HM, *et al*. An electronic health record driven algorithm to identify incident antidepressant medication users. *J Am Med Inform Assoc* 2014;21:785–91.
21. Gu Y, Tinn R, Cheng H, *et al*. Domain-Specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthcare* 2022;3:1–23.
22. Devlin J, Chang M-W, Lee K, *et al*. BERT: pre-training of deep bidirectional transformers for language understanding. 2021. Available: http://arxiv.org/abs/1810.04805
23. Radford A, Wu J, Child R, *et al*. n.d. Language models are unsupervised multitask learners. ;24.
24. Schick T, Schütze H. Exploiting cloze-questions for few-shot text classification and natural language inference. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 255–69.
25. Qu J, Hashimoto K, Liu W, *et al*. Few-shot intent classification by gauging entailment relationship between utterance and semantic label. Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI; Stroudsburg, PA, USA: Association for Computational Linguistics, 2021:8–15
26. Lee J, Yoon W, Kim S, *et al*. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;36:1234–40.
27. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. 2020. Available: http://arxiv.org/abs/1904.05342
28. Wolpert DH. Stacked generalization. *Neural Networks* 1992;5:241–59.
29. Arora S, May A, Zhang J, *et al*. Contextual embeddings: when are they worth it? Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020:2650–63
30. AskmyGP | the most effective online triage and consultation tool for gps. askmyGP. Available: https://askmygp.uk/ [Accessed 8 Jul 2022].
31. Appraisal Toolkit, AKT, RCA, Education, Trainee Portfolio - FourteenFish. Trainee portfolio - fourteenfish. Available: https://www.fourteenfish.com/ [Accessed 8 Jul 2022].
32. Stewart S, Pyne Y, McMillan B. Augmented consulting: the future of primary care? *BJGP Open* 2021;5:BJGPO.2020.0177.

Analysis of "One in a Million" Primary Care Consultation Conversations using Natural Language Processing

Appendix A: Fictional Representative Partial Transcript Example

GP:	Hello, I'm Doctor [REDACTED]; thank you for waiting. I'm just going to read out that number you have on that piece of paper if that's okay? 123456 Great, thank you for that, what is troubling you today?
Patient:	My headache is carrying on and on doctor. It's been weeks and nothing seems to be making it any better. There were some blood tests done, what happened about them?
GP:	Oh, I'm sorry to hear about your headache, let me look on the system for the results.
Patient:	Okay.
GP:	Oh, that's not loading up at the moment, tell me more about your pain.
Patient:	Yes, I need some extra medicine because it's very painful.
GP:	Okay are you taking anything at the moment?
Patient:	Paracetamol and Ibuprofen It's not helping at all.
GP:	You say it's not helping at all?
Patient:	No.
GP:	Okay, so where are the headaches happening when they come on?

Appendix B: ICPC-2 Codes & Consultations

ICPC-2 Code	CKS Health Topic
A General	Adverse drug reactions, AIDS and HIV infection, Analgesia - mild-to-moderate pain, Angio-oedema and anaphylaxis, Childhood cancers - recognition and referral, Chronic pain, Corticosteroids - oral, DMARDs, Drugs - adverse reactions, End of life care in children, Falls - risk assessment, Faltering growth, Feverish children - management, Feverish children - risk assessment, Glandular fever (infectious mononucleosis), Healthcare-associated infections, Hiccups, Immunizations - childhood, Immunizations - travel, Infectious mononucleosis - Glandular fever, Lyme disease, Malaria, Malaria prophylaxis, Measles, Multimorbidity, Mumps, NSAIDs - prescribing issues, Opioid dependence, Palliative cancer care - pain, Palliative care - general issues, Parvovirus B19 infection, Poisoning or overdose, Roundworm, Rubella, Scarlet fever, Sleep disorders - shift work and jet lag, Tamoxifen - managing adverse effects, Tiredness/fatigue in adults, Travel immunizations
B Blood, blood-forming	Anaemia - B12 and folate deficiency, Anaemia - iron deficiency, Anticoagulation - oral, Antiplatelet treatment, B12 and folate deficiency anaemia, Bruising, Erythrocytosis/polycythaemia, Folate and B12 deficiency anaemia, Gout, Haematological cancers - recognition and referral, HIV infection and AIDS, Hypercholesterolaemia - familial, Iron deficiency anaemia, Multiple myeloma, Neutropenic sepsis, Nosebleeds (epistaxis), Platelets - abnormal counts and cancer, Polycythaemia/erythrocytosis, Sepsis, Sickle cell disease, Sinusitis
D Digestive	Anal fissure, Aphthous ulcer, Appendicitis, Bowel screening, Candida - oral, Coeliac disease, Colic - infantile, Constipation, Constipation in children, Cow's milk allergy in children, Crohn's disease, Dental abscess, Diarrhoea - adult's assessment, Diarrhoea - antibiotic associated, Diarrhoea - prevention and advice for travellers, Diverticular disease, Dyspepsia - pregnancy-associated, Dyspepsia - proven functional, Dyspepsia - proven GORD, Dyspepsia - proven peptic ulcer, Dyspepsia - unidentified cause, Gallstones, Gastroenteritis, Gastrointestinal tract (lower) cancers - recognition and referral, Gastrointestinal tract (upper) cancers - recognition and referral, Gingivitis and periodontitis, GORD in children, Halitosis, Head and neck cancers - recognition and referral, Herpes simplex - oral, Irritable bowel syndrome, Nausea/vomiting in pregnancy, Palliative care - constipation, Palliative care - nausea and vomiting, Palliative care - oral, Periodontitis and gingivitis, Pilonidal sinus disease, Pruritus ani, Teething, Threadworm, Ulcerative colitis, Ulcers - aphthous
F Eye	Age-related macular degeneration, Blepharitis, Cataracts, Chalazion (meibomian cyst), Conjunctivitis - allergic, Conjunctivitis - infective, Corneal superficial injury, Dry eye syndrome, Glaucoma, Herpes simplex - ocular, Hordeola (styes), Macular degeneration - age-related, Meibomian cyst (chalazion), Red eye, Retinal detachment, Squint in children, Styes (hordeola), Uveitis
H Ear	Cholesteatoma, Earwax, Hearing loss in adults, Otitis externa, Otitis media - acute, Otitis media - chronic suppurative, Otitis media with effusion
K Circulatory	Angina, Atrial fibrillation, Cardiac arrest - out of hospital care, Chest pain, Chilblains, Compression stockings, CVD prevention - lipid modification, CVD risk assessment and management, Deep vein thrombosis, DVT prevention for travellers, Giant cell arteritis, Haemorrhoids, Heart failure - chronic, Hypertension, Hypertension in pregnancy, Leg ulcer - venous, Lipid modification - CVD prevention, Lipodermatosclerosis and venous eczema, MI - secondary prevention, Palpitations, Peripheral arterial disease, Raynaud's phenomenon, Superficial vein thrombosis (superficial thrombophlebitis), Varicose veins, Venous eczema and lipodermatosclerosis
L Musculoskeletal	Achilles tendinopathy, Acute childhood limp, Ankylosing spondylitis, Back pain - low (without radiculopathy), Baker's cyst, Bone and soft tissue sarcoma - recognition and referral, Bunions, Bursitis - pre-patellar, Carpal tunnel syndrome, Cervical radiculopathy - neck pain, Chest pain, Childhood limp - acute, Developmental rheumatology in children, Dupuytren's disease, Greater trochanteric pain syndrome, Knee pain - assessment, Leg cramps, Limp (childhood) - acute, Low

	back pain (without radiculopathy), Morton's neuroma, Neck lump, Neck pain - acute torticollis, Neck pain - cervical radiculopathy, Neck pain - non-specific, Neck pain - whiplash injury, Olecranon bursitis, Osgood-Schlatter disease, Osteoarthritis, Osteoporosis - prevention of fragility fractures, Plantar fasciitis, Polymyalgia rheumatica, Pre-patellar bursitis, Restless legs syndrome, Rheumatoid arthritis, Sarcoma (bone and soft tissue) - recognition and referral, Sciatica (lumbar radiculopathy), Shoulder pain, Sprains and strains, Temporomandibular disorders (TMDs), Tennis elbow, Torticollis (acute) - neck pain, Whiplash injury - neck pain
N Neurological	Bacterial meningitis and meningococcal disease, Bell's palsy, Benign paroxysmal positional vertigo, Blackouts, Brain and central nervous system cancers - recognition and referral, Carbon monoxide poisoning, Central nervous system and brain cancers - recognition and referral, Cerebral palsy, Cervical radiculopathy - neck pain, Delirium, Dementia, Epilepsy, Febrile seizure, Head injury, Headache - assessment, Headache - cluster, Headache - medication overuse, Headache - tension-type, Hearing loss in adults, Learning disabilities, Meniere's disease, Meningitis - bacterial meningitis and meningococcal disease, Migraine, Morton's neuroma, Multiple sclerosis, Neuralgia - post-herpetic, Neuropathic pain - drug treatment, Parkinson's disease, Post-herpetic neuralgia, Radiculopathy (cervical) - neck pain, Radiculopathy (lumbar) - sciatica, Sciatica (lumbar radiculopathy), Stroke and TIA, Tinnitus, Trigeminal neuralgia, Vertigo, Vertigo - benign paroxysmal positional, Vestibular neuronitis
P Psychological	Attention deficit hyperactivity disorder, Autism in adults, Autism in children, Benzodiazepine and z-drug withdrawal, Bipolar disorder, Delirium, Depression, Depression - antenatal and postnatal, Depression in children, Dyspepsia - proven functional, Eating disorders, Generalized anxiety disorder, Insomnia, Irritable bowel syndrome, Learning disabilities, Mental health in students, Obsessive-compulsive disorder, Post-traumatic stress disorder, Postnatal and antenatal depression, Problem drinking - alcohol, Psychosis and schizophrenia, Schizophrenia and psychosis, Self-harm
R Respiratory	Allergic rhinitis, Asthma, Breathlessness, Bronchiectasis, Chest infections - adult, Chest pain, Chronic obstructive pulmonary disease, Common cold, Coronavirus - COVID 19, Corticosteroids - inhaled, Cough, Cough - acute with chest signs in children, Croup, Epistaxis (nosebleeds), Immunizations - pneumococcal, Immunizations - seasonal influenza, Influenza - seasonal, Influenza (seasonal) - immunizations, Lung and pleural cancers - recognition and referral, Nosebleeds (epistaxis), Obstructive sleep apnoea syndrome, Palliative care - cough, Palliative care - dyspnoea, Palliative care - secretions, Pneumococcal immunizations, Pulmonary embolism, Sore throat - acute, Tuberculosis, Whooping cough
S Skin	Acne vulgaris, Alopecia areata, Alopecia, androgenetic - female, Alopecia, androgenetic - male, Animal and human bites, Bites - human and animal, Bites and stings - insect, Boils, carbuncles, and staphylococcal carriage, Burns and scalds, Candida - skin, Carbuncles, boils and staphylococcal carriage, Cellulitis - acute, Chickenpox, Chilblains, Corticosteroids - topical (skin), nose, and eyes, Dermatitis - contact, Eczema - atopic, Fungal nail infection, Fungal skin infection - body and groin, Fungal skin infection - foot, Fungal skin infection - scalp, Hand foot and mouth disease, Head lice, Herpetic whitlow - and staphylococcal, Hirsutism, human and animal bites, Hyperhidrosis, Impetigo, Insect bites and stings, Itch - widespread, Itch in pregnancy, Lacerations, Melanoma and pigmented lesions, Molluscum contagiosum, MRSA in primary care, Nappy rash, Neck lump, Palliative care - malignant skin ulcer, Paronychia - acute, Pigmented lesions and melanoma, Pityriasis rosea, Pityriasis versicolor, Psoriasis, Rosacea, Scabies, Scalds and burns, Seborrhoeic dermatitis, Shingles, Skin cancers - recognition and referral, Staphylococcal carriage, boils and carbuncles, Staphylococcal whitlow - and herpetic, Urticaria, Venous eczema and lipodermatosclerosis, Veruccae and warts, Vitiligo, Warts - anogenital, Warts and verrucae, Whitlow (staphylococcal and herpetic)
T Metabolic, endocrine, nutritional	Addison's disease, Alcohol - problem drinking, Cholecystitis - acute, Cirrhosis, Diabetes - type 1, Diabetes - type 2, Diabetes type 1 - insulin therapy, Diabetes type 2 - insulin therapy, Food allergy, Gilbert's syndrome, Hepatitis A, Hepatitis B, Hepatitis C, Herpes simplex - genital, Hypercalcaemia, Hyperthyroidism, Hyponatraemia, Hypothyroidism, Insulin therapy in type 1 diabetes, Insulin therapy in type 2 diabetes, Jaundice in adults, Jaundice in the newborn, Neck

	lump, Non-alcoholic fatty liver disease (NAFLD), Obesity, Pancreatitis - acute, Pancreatitis - chronic, Type 1 diabetes, Type 1 diabetes - insulin therapy, Type 2 diabetes, Type 2 diabetes - insulin therapy, Vitamin D deficiency in adults, Vitamin D deficiency in children
U Urinary	Acute kidney injury, Bedwetting (enuresis), Chronic kidney disease, Colic - renal or ureteric (acute), Enuresis - bedwetting, Incontinence - urinary, in women, Kidney disease - chronic, Kidney injury - acute, LUTS in men, Nocturnal enuresis - bedwetting, Pyelonephritis - acute, Renal or ureteric colic - acute, Urinary incontinence in women, Urinary tract infection - children, Urinary tract infection (lower) - men, Urinary tract infection (lower) - women, Urological cancers - recognition and referral
W Pregnancy, family planning	Antenatal and postnatal depression, Antenatal care - uncomplicated pregnancy, Breastfeeding problems, Contraception - assessment, Contraception - barrier methods and spermicides, Contraception - combined hormonal methods, Contraception - emergency, Contraception - IUS/IUD, Contraception - natural family planning, Contraception - progestogen-only methods, Contraception - sterilization, Dyspepsia - pregnancy-associated, Ectopic pregnancy, Hypertension in pregnancy, Infertility, Itch in pregnancy, Menopause, Miscarriage, Nausea/vomiting in pregnancy, Postnatal and antenatal depression, Pre-conception - advice and management, Pregnancy (uncomplicated) - antenatal care
X Female genital	Amenorrhoea, Bacterial vaginosis, Breast abscess and mastitis, Breast cancer - managing FH, Breast cancer - recognition and referral, Breast pain - cyclical, Breast screening, Candida - female genital, Cervical cancer and HPV, Cervical screening, Chlamydia - uncomplicated genital, Dysmenorrhoea, Endometriosis, Fibroids, Gonorrhoea, Gynaecological cancers - recognition and referral, Herpes simplex - genital, HPV and cervical cancer, Mastitis and breast abscess, Menopause, Menorrhagia, Ovarian cancer, Pelvic inflammatory disease, Polycystic ovary syndrome, Premenstrual syndrome, Pruritus vulvae, Pubic lice, Syphilis, Trichomoniasis, Vaginal discharge, Warts - anogenital
Y Male genital	Balanitis, Chlamydia - uncomplicated genital, Erectile dysfunction, Gonorrhoea, Haematospermia, Herpes simplex - genital, Prostate cancer, Prostatitis - acute, Prostatitis - chronic, Pubic lice, Scrotal pain and swelling, Syphilis, Trichomoniasis, Undescended testes, Urethritis - male, Varicocele, Warts - anogenital
Z Social	Child maltreatment - recognition and management, Conduct disorders in children and young people, Domestic violence and abuse, Smoking cessation, Support for adult carers

Appendix C: Custom Stopword Dictionary

The custom stopword list was written by our clinical author.

also	difference	may	past	taken	want
another	effect	mind	people	thinking	water
around	fine	month	question	thought	way
back	four	months	quite	three	week
bad	good	morning	rather	time	weeks
better	help	need	ray	topic	weight
care	high	never	result	try	well
cause	important	new	right	trying	worried
come	information	night	ring	two	worry
cks	left	normal	side	use	year
day	life	often	six	using	
days	like	one	sometimes	usually	

Appendix D: Classifier Hyperparameters

All hyperparameter tuning was performed using cross validation on the OIAM training split. The classifiers were trained using conventional supervised learning with OIAM transcripts as training examples.

The SVM classifier uses a radial basis function kernel and L2 regularisation. We compared regularisation strengths of C=1, 2, 10, and *100,* finding the best performance with *C=2*, in both multiclass and multilabel modes, which we then use to produce all results shown in the main paper.

Our naïve Bayes (NB) classifier uses multinomial distributions with a smoothing parameter of α*=0.001* for both multiclass and multilabel modes. We tested values of *0.001, 0.01, 0.1, 1, 10,* and *100.* For multiclass NB, we found a small improvement when the class probabilities (priors) were fixed to a uniform distribution, and so also use this setting to produce the results in the main paper.

For the BERT classifiers, the maximum number of epochs was set to 15 and batch size to 8. Using the AdamW optimizer, we tuned the learning rate and weight decay rate, finding $5e^{-5}$ for the learning rate and $1e^{-4}$ for weight decay for all the PubMedBERT classifier variants tested here. Documents above 512 tokens were broken into chunks containing whole sentences with up to 490 tokens per chunk. Initially, we split 15% of the training data off as a validation set to perform early stopping. However, since this reduces the small training dataset further, we found that this harmed performance. Therefore, we perform early stopping only if the training set performance has converged. This is a likely contributor to overfitting, which motivates the creation of a larger training dataset in future work.