

Title page:

**Identifying Hepatocellular Carcinoma from imaging reports using natural language processing to facilitate data extraction from electronic patient records**

**Authors:**

Tingyan Wang<sup>1,2,#</sup>, Ben Glampson<sup>3,4</sup>, Luca Mercuri<sup>3,4</sup>, Dimitri Papadimitriou<sup>3,4</sup>, Christopher R Jones<sup>4,5</sup>, David A Smith<sup>2,6</sup>, Hizni Salih<sup>1,7</sup>, Cori Campbell<sup>1,2</sup>, Oliver Freeman<sup>1,7</sup>, Steve Harris<sup>1,8</sup>, Kinga A Várnai<sup>1,6</sup>, Gail Roadknight<sup>1</sup>, Stephanie Little<sup>1</sup>, Theresa Noble<sup>1,6</sup>, Kerrie Woods<sup>1,6</sup>, Philippa C Matthews<sup>2,6,9,10,11</sup>, NIHR Health Informatics Collaborative Viral Hepatitis Theme Consortium<sup>§</sup>, Jim Davies<sup>1,8,\*</sup>, Graham S Cooke<sup>3,4,12,\*</sup>, Eleanor Barnes<sup>2,6,\*,#</sup>

**Affiliations:**

1. NIHR Oxford Biomedical Research Centre, Oxford, UK
2. Nuffield Department of Medicine, University of Oxford, Oxford, UK
3. NIHR Health Informatics Collaborative, Imperial College Healthcare NHS Trust, London, UK
4. NIHR Imperial Biomedical Research Centre, London, UK
5. Department of Infectious Disease, Imperial College London, London, UK
6. NIHR Health Informatics Collaborative, Oxford University Hospitals NHS Foundation Trust, Oxford, UK
7. Nuffield Department of Population Health, University of Oxford, Oxford, UK
8. Department of Computer Science, University of Oxford, Oxford, UK
9. The Francis Crick Institute, London, UK
10. Division of Infection and Immunity, University College London, London, UK
11. Department of Infectious Diseases, University College London Hospital, London, UK
12. Faculty of Medicine, Department of Infectious Disease, Imperial College London, UK

# Corresponding authors. The Peter Medawar Building for Pathogen Research, South Parks Road, Oxford, OX1 3SY, UK. Email: [ellie.barnes@ndm.ox.ac.uk](mailto:ellie.barnes@ndm.ox.ac.uk); [tingyan.wang@ndm.ox.ac.uk](mailto:tingyan.wang@ndm.ox.ac.uk)

\*Joint last authors, represents equal contribution.

§ NIHR Health Informatics Collaborative (HIC) Viral Hepatitis Theme Consortium:

Eleanor Barnes, Jim Davies, Kerrie Woods, Philippa C Matthews, William Gelson, Graham S Cooke, Salim I Khakoo, Eleni Nastouli, Alexander J. Stockdale, Stephen Ryder, Ahmed Elsharkawy, Douglas Macdonald, Kinga A Várnai, Tingyan Wang, Cori Campbell, Gail Roadknight, Stephanie Little, Ben Glampson, Dimitri Papadimitriou, Luca Mercuri, Christopher R Jones, Jakub Jaworski, Afzal Chaudhry, Vince Taylor, Cai Davis, Josune Olza, Florina Borca, Hang Phan, Frazer Warricker, Louise English, Luis Romão, David Ramlakhan, Stacy Todd, Heather Rogers, Karl McIntyre, Andrew Frankland, Hizni Salih, Steve Harris, Theresa Noble, David A Smith, Oliver Freeman, Sarah Montague, Irene Juurlink, Ruth Norris, George Tilston.

Word count: 4618

## 1 **Abstract**

2 **Background:** The National Institute for Health Research Health Informatics Collaborative  
3 (NIHR HIC) viral hepatitis theme is working to overcome governance and data challenges to  
4 collate routine clinical data from electronic patients records from multiple UK hospital sites  
5 for translational research. The development of hepatocellular carcinoma (HCC) is a critical  
6 outcome for patients with viral hepatitis with the drivers of cancer transformation poorly  
7 understood.

8  
9 **Objective:** This study aims to develop a natural language processing (NLP) algorithm for  
10 automatic HCC identification from imaging reports to facilitate studies into HCC.

11  
12 **Methods:** 1140 imaging reports were retrieved from the NIHR HIC viral hepatitis research  
13 database v1.0. These reports were from two sites, one used for method development (site 1)  
14 and the other for validation (site 2). Reports were initially manually annotated as binary  
15 classes (HCC vs. non-HCC). We designed inference rules for recognising HCC presence,  
16 wherein medical terms for eligibility criteria of HCC were determined by domain experts. A  
17 rule-based NLP algorithm with five submodules (regular expressions of medical terms, terms  
18 recognition, negation detection, sentence tagging, and report label generation) was  
19 developed and iteratively tuned.

20  
21 **Results:** Our rule-based algorithm achieves an accuracy of 99.85% (sensitivity: 90%,  
22 specificity: 100%) for identifying HCC on the development set and 99.59% (sensitivity: 100%,  
23 specificity: 99.58%) on the validation set. This method outperforms several off-the-shelf  
24 models on HCC identification including “machine learning based” and “deep learning based”  
25 text classifiers in achieving significantly higher sensitivity.

26  
27 **Conclusion:** Our rule-based NLP method gives high sensitivity and high specificity for HCC  
28 identification, even from imbalanced datasets with a small number positive cases, and can  
29 be used to rapidly screen imaging reports, at large-scale to facilitate epidemiological and  
30 clinical studies into HCC.

31  
32 **Keywords:** Hepatocellular Carcinoma (HCC), Imaging Reports, Free-text, Natural Language  
33 Processing (NLP), Rule-based, HBV/HCV

## 34 **Statement of Significance**

35

36 **Problem:** Establishing a cohort of hepatocellular carcinoma (HCC) from imaging reports via  
37 manual review requires advanced clinical knowledge and is costly, time consuming,  
38 impractical when performed on a large scale.

39

40 **What is Already Known:** Although some studies have applied natural language processing  
41 (NLP) techniques to facilitate identifying HCC information from narrative medical data, the  
42 proposed methods based on a pre-selection by diagnosis codes, or subject to certain  
43 standard templates, have limitations in application.

44

45 **What This Paper Adds:** We have developed a hierarchical rule-based NLP method for  
46 automatic identification of HCC that uses diagnostic concepts and tumour feature  
47 representations that suggest an HCC diagnosis to form reference rules, accounts for  
48 differing linguistic styles within reports, and embeds a data pre-processing module that can  
49 be configured and customised for different reporting formats. In doing so we have overcome  
50 major challenges including the analysis of imbalanced data (inherent in clinical records) and  
51 lack of existing unified reporting standards.

## 52 Background

53 Primary liver cancer (of which the vast majority are hepatocellular carcinoma; HCC) is the  
54 sixth most common cancer and the fourth leading cause of cancer-related mortality globally  
55 (1). There were approximately 0.9 million incident cases of HCC in 2020 and this will rise to  
56 an estimated ~1.4 million in 2040 (2). HCC typically develops in the setting of chronic liver  
57 disease; chronic infection with hepatitis B or C virus (HBV or HCV), alcohol and metabolic  
58 syndromes are the most frequent HCC risk factors (3). However, most patients with HCC are  
59 diagnosed in the late stages of the disease when treatment options are limited (4), leading to  
60 poor outcomes (~12% 5-year survival rate for these late-stage HCC patients (5)).

61  
62 Imaging investigations are currently crucial in HCC surveillance and diagnosis as there are  
63 no sensitive tumour biomarkers for accurate early HCC detection (6). The European  
64 Association for the Study of the Liver (EASL) clinical guidelines (7) strongly recommend that  
65 HCC surveillance using abdominal ultrasound is performed every 6 months in populations at  
66 high-risk of HCC. Imaging modalities including multiphasic computed tomography (CT) or  
67 magnetic resonance imaging (MRI), usually with intravenous contrast injection, are non-  
68 invasive methods currently recommended in clinical guidelines for HCC diagnosis in  
69 patients with liver cirrhosis —usually where liver nodules have first been detected by  
70 ultrasound (7-10). Imaging reports are typically generated, in non-standardised narrative  
71 format by radiologist, after reviewing the scans (11). These reports typically contain vital  
72 information about the presence or absence, and stage of HCC.

73  
74 Despite their potential utility, imaging reports exist mainly as free-text clinical narrative,  
75 which are difficult to query for secondary use. The potential utility of accurate automatic  
76 interpretation of these reports is enormous and includes clinical use in audit, diagnostic  
77 reporting and dissemination, financial management and planning of health service provision.  
78 Additionally these reports could be used for large national research programs, such as NIHR  
79 HIC, that seek to address important research questions for clinical benefit including HCC  
80 epidemiology, risk factors, biomarkers for early detection and clinical outcomes (12).  
81 Establishing an HCC cohort from unstructured imaging report data via manual review is  
82 costly, time consuming, impractical and requires advanced clinical knowledge when  
83 performed on a large scale (13). In the United States, the American College of Radiology  
84 (ACR) published the Liver Imaging Reporting and Data System (LI-RADS) to standardise the  
85 reporting of various imaging modalities for HCC (8), with each report assigned to a category.  
86 This means that these reports are relatively structured for easy for data query (14). However,  
87 this system has not yet been widely adopted in the UK, and it is currently impossible to

88 gather even simple information, like HCC diagnosis from imaging reports in an automated  
89 fashion. There is a clear need to leverage computational approaches for extracting  
90 information such as HCC diagnosis from unstructured imaging reports.

91

92 The classification of HCC using imaging reports requires extracting relevant words and  
93 phrases that define an HCC diagnosis from these free-text reports, for which natural  
94 language processing (NLP) techniques provide promising automatic solutions. NLP is a  
95 range of computational techniques for analysing and representing naturally occurring texts at  
96 different levels of linguistic analysis for the purpose of accomplish human-like language  
97 processing for a range of tasks or applications. Text classification leveraging NLP is an  
98 attractive method that may be used to facilitate the development of precisely phenotyped  
99 clinical cohorts using free-text medical records (13, 19, 20). Over recent years, machine  
100 learning (ML) or deep learning (DL) based NLP models have been shown to be effective in  
101 text classification tasks (15, 16); however, in the medical domain there is often a scarcity of  
102 large amounts of manually labelled data that is required for the robust training of such  
103 models. More recently, advanced NLP techniques, one of which uses pre-trained language  
104 models based on transformers have been developed (17), for which minimal annotated data  
105 is required to fine-tune the model for specific downstream NLP tasks. However, it is still  
106 difficult to leverage the existing pre-trained models for text classification where the data is  
107 highly imbalanced such that there is a significantly lower proportion of imaging reports with  
108 HCC cases compared to those that do not suggest a HCC diagnosis. To overcome the  
109 challenges of small sample size of labelled data and high-degree imbalance of data, in this  
110 study we propose to develop a rule-based NLP method to classify imaging reports for HCC  
111 identification, as diagnosis information extraction based on HCC inference rules does not  
112 require large amounts of manually-annotated data as well as can handle the imbalance  
113 issue by direct keywords recognition and matching.

114

115 Previous studies have applied NLP techniques including ML or rule-based methods or hybrid  
116 approaches into HCC-related medical text to extract standardised descriptions of anatomical  
117 findings, annotate tumour characteristics, or track the tumour statuses based on reports from  
118 patients with HCC (12, 22-24). Other studies have focused on classifying imaging reports to  
119 predict downstream radiology resource utilisation in patients undergoing HCC  
120 surveillance (25). However, there are few studies (13,14) addressing the same task as this  
121 study that focused on how to identify information that defines an HCC diagnosis from  
122 imaging reports. One study conducted text classification for HCC identification using imaging  
123 reports that were pre-selected by international classification of disease (ICD) codes (13),  
124 which inherently does not address the high imbalance issue of imaging reports, nor consider

125 the misclassification issue by diagnosis codes. Another study used a large bank of standard  
126 LI-RADS formatted imaging reports as annotated data to train models (14), which is not  
127 applicable in the region/ countries that do not widely use such template system for reporting  
128 imaging examination findings. Therefore, we set out to develop a rule-based NLP method  
129 which (a) utilises both diagnostic concepts and tumour feature representations that suggest  
130 an HCC diagnosis to form reference rules, (b) takes into account linguistic styles in different  
131 sections of imaging reports and the matching priority of medical terms, and (c) embeds data  
132 preprocessing steps that can be configured and customised for different reporting formats.  
133 Given these, our method will be capable of classifying HCC using imaging reports in real-  
134 world datasets that are being drawn automatically from operational systems (i.e., without  
135 pre-selection by diagnosis codes), as well as can be applied in imaging reports with different  
136 reporting styles, rather than being subject to certain templates e.g., LI-RADS reporting  
137 standard.

138

139

## 140 **Methods and Materials**

### 141 **Data sources**

142 Imaging reports used in this study were collected from patients with HBV, HCV, or HEV and  
143 stored in the central data repository of the NIHR HIC viral hepatitis theme, for which the data  
144 collection process has been detailed previously (15, 16). The imaging reports for this study  
145 cover various imaging modalities, including CT, MRI, and ultrasound, and were originally  
146 from two different sites: Oxford University Hospitals NHS Foundation Trust (site 1) and  
147 Imperial College Healthcare NHS Trust (site 2). Imaging reports were manually labelled (i.e.,  
148 assigning a binary label ('yes' or 'no') to an imaging report to indicate whether there is HCC  
149 presence in that imaging examination). We then selected imaging reports that have been  
150 manually labelled to develop and validate a rule-based method.

151

### 152 **Rule-based NLP model**

153 The framework of the proposed rule-based NLP method for HCC identification from imaging  
154 reports is shown in **Figure 1**, including text preprocessing, HCC inference rules, rule-based  
155 algorithm, and iterative development.

156

#### 157 **Module 1: Text preprocessing using NLP tools**

158 All retrieved raw reports were preprocessed in Python 3.6 using NLP typical techniques,  
159 including Natural Language Toolkit (NLTK) and spaCy. Firstly, we excluded those invalid  
160 reports, e.g., with a “due to image omission or degradation” note, or blank reports. Then for

161 the eligible reports, scripts were made to clean and pre-process the text (section extraction  
162 or removal). In this part, we applied the keywords matching algorithm to extract the  
163 impression part of an imaging report or remove the irrelevant description. Specifically,  
164 different strategies were used to tackle reports based on their characteristics (**Figure S1**):

- 165 • If an “impression” section was available in a report, we used this section for HCC  
166 identification. This was performed using regular expression for the words  
167 “impression”, “opinion”, “conclusion”, “summary”, “comment”, or their corresponding  
168 abbreviations or synonyms.
- 169 • For those reports without an “impression” section but with a “findings” section  
170 available, we extracted the “findings” section. Additionally, as the “findings” section  
171 typically includes descriptions of multiple body systems, we excluded the text  
172 irrelevant to liver system by removing sentences mentioning other body systems. To  
173 perform this process we built a lexicon of body systems that are not relevant to HCC  
174 information extraction, such as gallbladder, spleen, kidney, or their synonyms, and  
175 conditions related to these organs (eg. gallstone, splenomegaly, hydronephrosis)  
176 (**Table S1**).
- 177 • For those reports without an explicit “impression” or “findings” section title, we  
178 considered a removal of irrelevant content from the report. Specifically, as we are  
179 interested in the findings of the current imaging investigation, we excluded the  
180 sections such as “reason for study”, “reason for examination”, “clinical history”, or  
181 “clinical information”, by using keyword matching filters, and used the remaining text  
182 for a report.
- 183 • For those reports without any section titles, we split the report and excluded the text  
184 irrelevant to findings of liver system based on the built lexicon of other body systems  
185 as mentioned above.

186  
187 For the extracted text, we then employed text processing techniques to reduce the  
188 complexity and noise of the data, and prepare the substantive content for further analysis. In  
189 particular, we first split reports into paragraphs using the new line character (as it not  
190 uncommon that imaging reports list relevant findings/impressions line by line) and then  
191 tokenise text into individual sentences. We also used lemmatization technique to  
192 standardise words and removed stop words (e.g., “an”, “of”, and “the”) that are used for  
193 connection and grammar but do not provide informative meaning. However, we did not  
194 consider words such as “nor”, “not”, “none”, “can’t”, “aren’t”, “amount” as stop words in this  
195 study, as they have negation meaning or useful information.

196



## 197 Module 2: Design of HCC inference rules for imaging reports

198 To build inference rules for HCC from imaging reports, key terms for the inclusion and  
199 exclusion criteria of HCC were highlighted by a senior hepatologist (E.B.) to form a list of  
200 “seed terms”. We then expanded the list based on multiple clinical guidelines for diagnosis of  
201 HCC based on imaging examination (7-9, 17). In the development phase, we initially ranked  
202 the priorities of the key terms to form diagnostic rules, with a reference to the LI-RADS (8).  
203 Typically, medical terms presenting in the “impression” section are different from those in the  
204 “findings” section in an imaging study report, as the former tend to be conclusive/diagnostic  
205 terms (e.g., “no evidence of hepatoma”) whilst the latter tend to be descriptive terms of  
206 organ/tumour features (e.g., “*There are two subcentimetre foci of hypervascularity within*  
207 *segment 8 of the liver. There is no washout demonstrated*”). Therefore, to identify HCC  
208 information from different sections we further classified key terms into two levels, one  
209 concerning concepts of tumour features and the other concerning the variations of tumour  
210 diagnostic concepts. The medical terms for HCC identification are provided in **Table 1**.

211

## 212 Module 3: Development of rule-based NLP algorithm for HCC identification

213 We used the initially-developed inference rules to design a rule-based NLP algorithm  
214 prototype. The logical process of the algorithm is to tag each sentence within a report and  
215 then aggregating all the tags to generate a final label for the report based on the inference  
216 rules set out (**Figure S2**). The algorithm module contains five submodules as follows:

- 217 • *Regular expressions of medical terms.* We iteratively built and tuned the regular  
218 expressions of target medical terms based on the development set (**Table S2**).
- 219 • *Relevant medical terms recognition.* We matched and extracted terms from imaging  
220 reports by the built regular expressions. To facilitate the following step of negation  
221 detection we replaced a matched term with a corresponding standard terminology  
222 that can be accurately recognised by a named entity recognition (NER) model.
- 223 • *Negation detection.* For detecting if a target term is negated in imaging reports, we  
224 built a negation detection module based on negSpacy (18) and Stanza. negSpacy is  
225 a pipeline developed based on NegEx (19) for identifying negations in text. Stanza is  
226 a collection of state-of-the-art NLP tools, for syntactic analysis and NER in both  
227 general and biomedical domains (20, 21). Specifically, our negation detection module  
228 firstly wraps Stanza, using the pre-trained syntactic analysis model based on Medical  
229 Information Mart for Intensive Care III (MIMIC-III) database and the pre-trained NER  
230 model based on MIMIC-III imaging reports. Then we added negSpacy components  
231 into our negation detection module. For negation patterns in negSpacy, we used the



232 clinical sensitive version designed for clinical domain (18), and also added new  
233 customised negation patterns for imaging reports based on our development set.

234 • *Sentence tagging based on priority (sentence-level)*. With identifying medical terms  
235 and detecting negation, we sequentially tagged a sentence based on the priority  
236 ranking to obtain a list of candidate tags for each sentence. We then assigned the  
237 tag with the highest priority among the candidate tags to each sentence. As a result,  
238 a set of sentence tags will be generated for each report.

239 • *Inferring report label (document-level)*. To infer a final label (HCC or non-HCC) for  
240 each report, we sorted the sentence tags by the priority ranking, and generated  
241 report label using the sentence tag with highest priority.

242

243 **Module 4: Iterative tuning process for the rule-based NLP algorithm**

244 We used the development set to tune the designed rule-based NLP algorithm (**Figure 1**,  
245 *Module 4*). Specifically, we first utilised the algorithm to label reports, then checked the  
246 reason for errors so as to adjust the key terms lists and their priority ranking accordingly.  
247 The tuning process was iteratively performed until the accuracy could not be further  
248 improved.

249

#### 250 **Evaluation metrics**

251 To evaluate the algorithm performance, a confusion matrix was generated. The true positive  
252 (TP) value indicates the number of HCC-labelled reports that are correctly recognised by the  
253 algorithm as HCC, whilst the true negative (TN) value indicates the number of reports  
254 labelled with non-HCC and correctly recognised by the algorithm as non-HCC. The false  
255 positive (FP) value represents the number of reports labelled with non-HCC but wrongly  
256 recognised by the algorithm as HCC, whilst the false negative (FN) value represents the  
257 number of reports labelled with HCC but wrongly recognised by the algorithm as non-HCC.

258 We used the following metrics to evaluate algorithm performance: sensitivity (aka, recall or  
259 true positive rate),  $TP / (TP + FN)$ ; specificity (aka, true negative rate),  $TN / (TN + FP)$ ;  
260 precision (aka, positive predictive value (PPV)),  $TP / (TP + FP)$ ; accuracy,  $(TP + TN) / (TP +$   
261  $FP + TN + FN)$ , and; F1 score,  $TP / (TP + 1/2(FP + FN))$ . As the original labels of each report  
262 were binary, “probably positive” or “definitely positive” from our algorithm were collapsed as  
263 “positive”, and “probably negative” or “definitely negative” were collapsed as “negative”.

264

#### 265 **ML and DL models for comparison**

266 We compared the performance of our rule-based NLP algorithm against the following three  
267 categories of off-the-shelf supervised learning models: (a) conventional simple yet

268 competitive ML based text classifiers, including Naive Bayes (NB), Logistic Regression (LR),  
269 and Support Vector Machine (SVM), with bag-of-words (BOW) or term frequency-inverse  
270 document frequency (TF-IDF) as the text representation methods (22). (b) DL based text  
271 classification architectures, including text convolutional neural network (CNN) (23) and  
272 hierarchical attention networks (HAN) (24), with word2vec as the representation method of  
273 text (25, 26). To obtain word embedding related to target domain of imaging reports, we  
274 trained a word2vec model on all reports of CT, MRI, or ultrasound scans from MIMIC-III  
275 database. (c) Fine-tuning pre-trained DL models with contextual representation of text,  
276 where we used Bidirectional Encoder Representations from Transformers (BERT) (27) and  
277 its variant versions in clinical domain including ClinicalBERT (28) and BlueBERT (29). For all  
278 the supervised learning models, we used data from site 1 (the development set) to train  
279 classifiers and evaluate the performance on data from site 2 (the test set) with the metrics  
280 above. As data are highly imbalanced, we applied cost-sensitive learning scheme during  
281 training to take into account the skewed distribution (30).

282

## 283 **RESULTS**

### 284 **Descriptive characteristics of imaging reports**

285 We included 1140 imaging reports with labels available (655 from site 1 and 485 from site 2)  
286 for this study. The earliest reports are sourced from the year 2003, while the latest reports  
287 are from the year 2019. Due to the long timespan (~15 years) and reports collected from  
288 different sites, these imaging reports have variable reporting and linguistic styles. For both  
289 sites, the imaging modality of most reports was ultrasound, with CT and MRI examinations  
290 performed less frequently, in accordance with clinical guidelines (**Table 2**). For site 1 vs. site  
291 2, the prevalence of HCC presence in ultrasound reports was 0.84% (5/598) vs. 1.27%  
292 (4/314), whilst 18.18% (6/33) vs. 5.43% (5/92) for CT reports and 12.50% (3/24) vs. 3.80%  
293 (3/79) for MRI reports.

294

295 Text characteristics of reports across the two sites are summarised in **Table 3**. The  
296 proportion of reports that had “impression” sections was varied between sites. Only 18%  
297 (117/655) of reports from site 1 included an “impression” section, while the remaining 82% of  
298 reports did not, making HCC identification more challenging. In contrast, 75% (366/485) of  
299 the reports from site 2 had “impression” sections. The number of sentences or words per  
300 report was significantly different between the two sites, both before and after text pre-  
301 processing (all p values <0.001), which reflects significant heterogeneity between sites. The  
302 information of running time for data preprocessing are provided in **Table S3**.

303

304 **Performance on the development set**

305 After iteratively tuning and amending the algorithm using the development set, the priority  
306 ranking of terms in diagnostic rules were finalised (**Table 1**). Compared to the original labels  
307 in the development set, our algorithm obtained an overall consistency of 99.38%, with 71.43%  
308 consistency on HCC cases and 100% on non-HCC cases, with a short duration of running  
309 time (**Table S3**). The confusion matrix is shown in **Figure 2A**. However, among the five  
310 HCC-labelled reports being identified as non-HCC by the algorithm, four were actually non-  
311 HCC reports after further manual checking, and the remaining one was incorrectly identified  
312 by the algorithm (**Table S4**). To further verify the correctness of our algorithm to label non-  
313 HCC cases, we randomly selected 10% (64/641) of original non-HCC reports by stratified  
314 sampling (59 for ultrasound, 3 for CT, 2 for MRI), and manually checked the ground truth  
315 labels of these reports. We found that all the 64 reports were accurately identified by the  
316 algorithm. We then generated a new confusion matrix after data correction (**Figure 2B**), and  
317 the overall accuracy is 99.85%, with 90% (9/10) of sensitivity and 100% (645/645) of  
318 specificity.

319

320 We summarised the frequency of terms used for the identification process in the  
321 development set (**Figure 2C**). HCC presence identification was derived mostly based on  
322 “(probable) HCC” or its synonyms (“hepatoma/ hepatocellular carcinoma”), followed by  
323 “enlarged liver lesion” or “hypervascular lesion”. In contrast, non-HCC (HCC absence)  
324 identification was inferred mostly based on normal examination or negation of liver lesion,  
325 followed by benign tumour or negation of HCC.

326

327 **Performance on the validation set**

328 Compared to the originally collected labels in the test set, the algorithm achieved an overall  
329 consistency of 99.18% (481/485), with 91.67% consistency (11/12) on HCC cases, and  
330 99.37% (470/473) on non-HCC cases; the confusion matrix is shown in **Figure 3A**. However,  
331 for the one HCC-labelled report being identified as non-HCC, the ground-truth label of the  
332 report is in fact “non-HCC” (**Table S5**). For the three non-HCC-labelled reports being  
333 identified as HCC, one report was in fact highly suspicious of HCC, two reports were  
334 equivocal or indeterminate (**Table S5**). To further verify the correctness of our proposed  
335 algorithm on non-HCC cases in the test set, 47 reports (10%) of the original 468 non-HCC  
336 reports were randomly selected, for which ground truth labels were manually checked. We  
337 found that all these 47 reports were accurately identified by our proposed NLP algorithm.

338

339 After data correction, the overall accuracy of the rule-based NLP method on the test set is  
340 99.59% (483/485), with 100% sensitivity and 99.58% (471/473) specificity (**Figure 3B**).

341 Similar to that of the development set, we also summarised the frequency of terms used for  
342 identifying the process in the test set (**Figure 3C**). HCC presence identification was derived  
343 mostly based on “(probable) HCC” or its synonyms (“hepatoma/ hepatocellular carcinoma”),  
344 followed by “liver lesion” and “hypervascular lesion”. In contrast, non-HCC (HCC absence)  
345 identification was derived mostly based on normal examination or negation of liver lesion,  
346 then followed by negation of HCC.

347

### 348 **Comparison of performance between our rule-based method and ML/DL models**

349 We used the development set (after data correction) to train ML and DL based models. The  
350 performance comparisons between the proposed rule-based NLP method and supervised  
351 learning models in the test set are summarised in **Table 4**. The rule-based method was  
352 superior to conventional ML models regardless of the feature representation method (BOW  
353 or TF-IDF) as well as outperformed DL models (CNN, HAN, BERT, ClinicalBERT, or  
354 BlueBERT). Particularly, the rule-based method achieved significantly higher sensitivity  
355 compared to other models, although the rule-based method demonstrated only a marginal  
356 improvement in overall accuracy due to a small number of HCC cases. HAN with word2vec  
357 obtained a higher sensitivity (75%) than other supervised learning models; however, it is still  
358 significantly inferior to the rule-based method (100%). Although the pre-trained transformers  
359 models can achieve comparable specificity (97.67% to 99.79%), they had low sensitivity  
360 (33.33% to 58.33%) due to small sample size of data for the class of HCC. The rule-based  
361 method was capable of handling skewed data distributions by achieving a much higher F1  
362 score and precision (F1 score = 0.9231; Precision = 99.59%) compared to supervised  
363 learning models regardless of model architecture and feature representation method (F1  
364 score = 0.11 to 0.47; Precision = 6.3% to 80.0%).

365

## 366 **DISCUSSION**

367 In this study, we designed a rule-based NLP algorithm to identify the presence of HCC from  
368 free-text imaging reports. With a reference to ground-truth labels, the proposed rule-based  
369 NLP method achieve superior performance, especially higher sensitivity in highly-  
370 imbalanced datasets with small numbers of positive cases, compared to existing off-the-  
371 shelf models. The inner-working and sensitivity of the proposed method are not related to  
372 imaging modality, because we considered the major features used for HCC identification in  
373 different imaging modalities. For those labelled reports, the proposed method can be used to  
374 correct data, as we found some reports were incorrectly labelled during data collection,  
375 especially for HCC presence. For those unlabelled reports, the proposed method is capable

376 of rapidly identifying the presence of HCC from the imaging reports, so as to provide labels  
377 for further cohort studies to advance HCC-related research.

378

379 Rule-based methods, ML or DL based models, or hybrid methods have been used to  
380 develop NLP pipelines in imaging reports (31-36), but few studies have focused on HCC  
381 identification (13, 14). A previous study has applied a hybrid NLP algorithm (Automated  
382 Retrieval Console, ARC), that was originally designed for other cancers, in HCC  
383 identification from free-text medical notes (13), wherein HCC were firstly identified by ICD-9  
384 codes and then further verified by the hybrid NLP algorithm. With a pre-selection of  
385 candidate reports by ICD codes, this method can avoid the imbalanced data for text  
386 classification, whilst the issue of HCC misclassification by ICD codes is not considered. As  
387 LI-RADS has been well applied in US, another study (14) trained ML based models for HCC  
388 identification, which used imaging reports created from a LI-RADS template as labelled data.  
389 However, it is difficult to transfer such methods to those regions where LI-RADS standard  
390 are not widely used. Our study differs from previous studies and is novel in that we designed  
391 a rule-based NLP which utilised both tumour feature and diagnostic concept representations  
392 of HCC, along with the priority ranking of these terms, for which we accounted for linguistic  
393 styles in different sections of reports. Therefore, our algorithm has high specificity and  
394 perfect sensitivity in the datasets we have interrogated, and is capable of accurately  
395 identifying the presence of HCC, even in a small and highly-imbalanced dataset.

396

397 Two types of methods are typically used to handle the issue of highly imbalanced classes in  
398 text classification: a) resampling method and b) cost-sensitive learning scheme (30). In our  
399 study, the sample size of the minority class (HCC) was too small to apply resampling  
400 methods (specifically oversampling techniques such as SMOTE), which may produce un-  
401 representative synthetic examples for the minority class (30, 37). Therefore, we used a cost-  
402 sensitive learning scheme, putting greater weights on the minority class (in this case, HCC)  
403 to tackle the issue of imbalance. However, even in our study the ML or DL models trained  
404 with this imbalanced learning scheme, were unable to achieve comparable performance to  
405 the proposed rule-based method, due to limited data on positive cases for supervised  
406 learning.

407

408 Compared to ML or DL models, the advantages of a rule-based method are that a large  
409 amount of labelled data is not required for development and validation, and that the method  
410 is useful as a cold-start to automatically label reports. It is therefore sensible to choose a  
411 rule-based method at the initial phase, instead of ML or DL models as large volume of

412 labelled data is needed for training such models to achieve high performance, particularly  
413 when a large imbalance between positive and negative cases exists.

414

415 Our rule-based NLP method achieved a high accuracy on both HCC and non-HCC cases  
416 and it is capable of rapidly labelling data on a large-scale. Our findings provide an  
417 opportunity for DL to be applied in HCC identification in future by weak supervision schemes  
418 (38, 39). Weak supervision uses noisy labelled data that might include incorrect labels to  
419 train supervised learning models. DL has been demonstrated to be valuable and robust in  
420 medical text mining tasks, even in noisy data. However, the lack of labelled data is a major  
421 bottleneck to applying such technique for disease information extraction (40). Recently DL-  
422 based NLP models trained with rule-based method labelled data (weak supervision learning)  
423 has been successfully applied into medical free-text mining tasks such as hip fracture  
424 classification (38) and Alzheimer's disease risk factor characterisation(41), achieving  
425 competitive performance compared to models trained on human-annotated data.  
426 Promisingly, our proposed rule-based NLP method provides a foundation for application of  
427 these techniques to identify HCC from free-text imaging reports. Data collection within the  
428 NIHR HIC viral hepatitis theme is still ongoing (15), and therefore it is expected that a large  
429 amount of unlabelled imaging reports will be collected from more collaborators across  
430 England in the near future. Reports can be automatically labelled by our rule-based NLP  
431 pipeline and accordingly can be fed into a DL model for training.

432

433 It is well recognised that a rule-based method typically has limited flexibility and robustness.  
434 Misspellings and numerous non-standardised ways of describing the same concept are  
435 common in imaging reports. Rule-based NLP systems may not work well if vocabularies and  
436 linguistic styles are significantly different. In the future, with more data accumulated, a more  
437 robust and flexible NLP pipeline using DL will be developed on the basis of the proposed  
438 rule-based method as discussed above. In addition, staging HCC is important to determine  
439 outcomes and planning of optimal therapy (17). Future work will also extend the proposed  
440 methodology for HCC staging.

441

## 442 **Conclusion**

443 A rule-based NLP method to classify free-text imaging reports is developed to accurately  
444 identify HCC information for highly-imbalanced small-size dataset. When labelled data are  
445 not sufficiently available, a rule-based NLP algorithm with an integration of the domain  
446 knowledge can achieve a desirable accuracy, compared to supervised learning algorithms.  
447 Such a rule-based NLP algorithm not only can be utilised to rapidly screen patient free-text

448 imaging reports on a large scale to facilitate epidemiological and clinical studies but also can  
449 serve as a tool for data correction.  
450



451 **Ethics approval and consent to participate:** The research database for the NIHR HIC  
452 viral hepatitis theme was approved by South Central - Oxford C Research Ethics Committee  
453 (REF Number: 15/SC/0523). All methods in this study were carried out in according to  
454 relevant guidelines and regulations. The requirement for written informed consent was  
455 waived by South Central - Oxford C Research Ethics Committee, because data have been  
456 anonymised before its use and the study is retrospective.

457

458 **Funding:** This research has been conducted using National Institute for Health Research  
459 (NIHR) Health Informatics Collaborative (HIC) data resources and funded by the NIHR HIC,  
460 and has been supported by NIHR Biomedical Research Centres at Oxford and Imperial. EB  
461 is an NIHR senior investigator. GSC is supported in part by the Imperial NIHR Biomedical  
462 Research Centre and NIHR Research Professorship. PCM is funded by the Wellcome Trust  
463 (ref. 110110/Z/15/Z), the Francis Crick Institute, and UCL NIHR BRC. CC is a doctoral  
464 student who receives partial doctoral funding from GlaxoSmithKline. The views expressed in  
465 this article are those of the authors and not necessarily those of the National Health Service,  
466 the NIHR, or the Department of Health.

467

468 **Acknowledgements:** The authors would like to thank all the research nurses and research  
469 admin staff at the contributing site for their help in data collection and submission.

470

471 **Conflicts of Interest:** GC reports personal fees from Gilead and Merck Sharp & Dohme  
472 outside the submitted work. EB and PCM have academic collaborative partnerships with  
473 GSK. Other authors have no conflict of interest.

474

#### 475 **Author contributions**

476 TW, EB, GSC, and JD conceived the study. BG, LM, DP, CRJ, TW, DAS, HS, CC, OF, SH, KAV,  
477 GR, SL, TN, and KW contributed to the acquisition, processing, interpretation, or management of  
478 data from study sites. TW designed the analytical strategy and conducted the analysis,  
479 supervised by EB, GSC, and JD. EB, GSC, and PCM helped interpret the findings. TW drafted  
480 the original manuscript with input from EB. All authors reviewed and revised the manuscript  
481 critically and approved the final version for publication.

482

#### 483 **Code availability**

484 The codes of the rule-based algorithm implementation will be made publicly available at  
485 github repository for this study: <https://github.com/tinaty/NLPHCC>.

486

#### 487 **Data availability**

488 Data from NIHR HIC viral hepatitis theme may be made available to researchers on request  
489 following positive review by the steering committee. Further details are available at  
490 <https://hic.nihr.ac.uk>. Queries regarding data access should be directed to [orh-](mailto:orh-tr.nihrhic@nhs.net)  
491 [tr.nihrhic@nhs.net](mailto:tr.nihrhic@nhs.net). MIMIC is provided through the work of researcher at the MIT Laboratory  
492 for Computational Physiology and the collaborators. Data are available through formally  
493 requesting access with the steps here: <https://mimic.mit.edu/docs/gettingstarted/>.

494

#### 495 **Abbreviations**

496 AASLD, American Association for the Study of Liver Diseases; ACR, American College of  
497 Radiology; AFP, Alpha fetoprotein; APASL, Asian Pacific Association for the Study of the  
498 Liver; BERT, bidirectional encoder representations from transformers; BOW, bag-of-words;  
499 CNN, convolutional neural network; CT, multiphasic computed tomography; DL, deep  
500 learning; EASL, European Association for the Study of the Liver; EHR, electronic health  
501 record; ESMO, European Society for Medical Oncology; FP, false positive; FN, false  
502 negative; HAN, hierarchical attention networks; HBV, hepatitis B virus; HCC, hepatocellular  
503 carcinoma; HCV, hepatitis C virus; HIC, Health Informatics Collaborative; ICHT, Imperial  
504 College Healthcare NHS Trust; LI-RADS, liver imaging reporting and data system; LR,  
505 logistic regression; MRI, magnetic resonance imaging; NB, Naive Bayes; NHS, National  
506 Health Service; NIHR, National Institute for Health Research; NLP, natural language  
507 processing; NLTK, Natural Language Toolkit; OUH, Oxford University Hospitals; SVM,  
508 support vector machine; TF-IDF, term frequency inverse document frequency; TP, true  
509 positive; TN, true negative.

## References

1. International Agency for Research on Cancer WHO. Cancer today <https://gcoiarcfr/today/home>. 2020.
2. International Agency for Research on Cancer WHO. Cancer tomorrow. <https://gcoiarcfr/tomorrow/en>. 2020.
3. Singal AG, El-Serag HB. Hepatocellular Carcinoma From Epidemiology to Prevention: Translating Knowledge into Practice. *Clin Gastroenterol Hepatol*. 2015;13(12):2140-51.
4. Chen B, Garmire L, Calvisi DF, Chua MS, Kelley RK, Chen X. Harnessing big 'omics' data and AI for drug discovery in hepatocellular carcinoma. *Nat Rev Gastroenterol Hepatol*. 2020;17(4):238-51.
5. Krishnamurthy S, Gilot D, Ahn SB, Lam V, Shin JS, Guillemin GJ, et al. Involvement of Kynurenine Pathway in Hepatocellular Carcinoma. *Cancers (Basel)*. 2021;13(20).
6. Villanueva A. Hepatocellular Carcinoma. *N Engl J Med*. 2019;380(15):1450-62.
7. European Association for the Study of the Liver. Electronic address eee, European Association for the Study of the L. EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma. *J Hepatol*. 2018;69(1):182-236.
8. Marrero JA, Kulik LM, Sirlin CB, Zhu AX, Finn RS, Abecassis MM, et al. Diagnosis, Staging, and Management of Hepatocellular Carcinoma: 2018 Practice Guidance by the American Association for the Study of Liver Diseases. *Hepatology*. 2018;68(2):723-50.
9. Omata M, Cheng AL, Kokudo N, Kudo M, Lee JM, Jia J, et al. Asia-Pacific clinical practice guidelines on the management of hepatocellular carcinoma: a 2017 update. *Hepatol Int*. 2017;11(4):317-70.
10. Friedrich Foerster PRG. Comparison of the current international guidelines on the management of HCC. *Jhep Reports*. 2019;1:6.
11. Wilcox JR. The written radiology report. *Applied Radiology*. 2006;35(7):33.
12. Yim W-w, Kwan SW, Yetisgen M. Classifying tumor event attributes in radiology reports. *Journal of the Association for Information Science and Technology*. 2017;68(11):2662-74.
13. Sada Y, Hou J, Richardson P, El-Serag H, Davila J. Validation of Case Finding Algorithms for Hepatocellular Cancer From Administrative Data and Electronic Health Records Using Natural Language Processing. *Med Care*. 2016;54(2):e9-14.
14. Banerjee IC, H.H.; Desser, T.; Rubin, D.L.,. A Scalable Machine Learning Approach for Inferring Probabilistic US-LI-RADS Categorization. In *AMIA Annual Symposium Proceedings*. 2018;2018:215.

15. Smith DA, Wang T, Freeman O, Crichton C, Salih H, Matthews PC, et al. National Institute for Health Research Health Informatics Collaborative (NIHR HIC): Development of a Pipeline to Collate Electronic Clinical Data for Viral Hepatitis Research. medRxiv. 2019.
16. Wang T, Smith DA, Campbell C, Freeman O, Moysova Z, Noble T, et al. Cohort Profile: National Institute for Health Research Health Informatics Collaborative: Hepatitis B Virus (NIHR HIC HBV) Research Dataset. medRxiv. 2021.
17. Vogel A, Cervantes A, Chau I, Daniele B, Llovet J, Meyer T, et al. Hepatocellular carcinoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2018;29(Suppl 4):iv238-iv55.
18. Pizarro J. <https://github.com/jenojp/negspacy>.
19. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001;34(5):301-10.
20. Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD, editors. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. ACL; 2020.
21. Zhang Y, Zhang Y, Qi P, Manning CD, Langlotz CP. Biomedical and clinical English model packages for the Stanza Python NLP library. J Am Med Inform Assoc. 2021;28(9):1892-9.
22. Chapman BE, Lee S, Kang HP, Chapman WW. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. J Biomed Inform. 2011;44(5):728-37.
23. Kim Y, editor Convolutional Neural Networks for Sentence Classification 2014 oct; Doha, Qatar: Association for Computational Linguistics.
24. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E, editors. Hierarchical Attention Networks for Document Classification 2016 jun; San Diego, California: Association for Computational Linguistics.
25. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781. 2013.
26. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems. 2013;26.
27. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018.
28. Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323. 2019.

29. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. arXiv preprint arXiv:190605474. 2019.
30. Padurariu C, Breaban ME. Dealing with Data Imbalance in Text Classification. *Procedia Computer Science*. 2019;159:736-45.
31. Pons EB, L.M.; Hunink, M.M.; Kors, J.A. Natural Language Processing in radiology A systematic review. *Radiology*. 2016;279(2):329-43.
32. D'Avolio LW, Nguyen TM, Farwell WR, Chen Y, Fitzmeyer F, Harris OM, et al. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). *J Am Med Inform Assoc*. 2010;17(4):375-82.
33. Banerjee I, Bozkurt S, Alkim E, Sagreiya H, Kurian AW, Rubin DL. Automatic inference of BI-RADS final assessment categories from narrative mammography report findings. *J Biomed Inform*. 2019;92:103137.
34. Sippo DA, Warden GI, Andriole KP, Lacson R, Ikuta I, Birdwell RL, et al. Automated extraction of BI-RADS final assessment categories from radiology reports with natural language processing. *J Digit Imaging*. 2013;26(5):989-94.
35. Castro SM, Tseytlin E, Medvedeva O, Mitchell K, Visweswaran S, Bekhuis T, et al. Automated annotation and classification of BI-RADS assessment from radiology reports. *J Biomed Inform*. 2017;69:177-87.
36. Casey A, Davidson E, Poon M, Dong H, Duma D, Grivas A, et al. A systematic review of natural language processing applied to radiology reports. *BMC Medical Informatics and Decision Making*. 2021;21(1):179.
37. Li Y, Guo H, Zhang Q, Gu M, Yang J. Imbalanced text sentiment classification using universal and domain-specific knowledge. *Knowledge-Based Systems*. 2018;160:1-15.
38. Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson EJ, et al. A clinical text classification paradigm using weak supervision and deep representation. *BMC Med Inform Decis Mak*. 2019;19(1):1.
39. Dai H, Song Y. Neural Aspect and Opinion Term Extraction with Mined Rules as Weak Supervision. arXiv. 2019;1907.03750.
40. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*. 2018;15(141).
41. Shen Z, Yi Y, Bompelli A, Yu F, Wang Y, Zhang R. Extracting Lifestyle Factors for Alzheimer's Disease from Clinical Notes Using Deep Learning with Weak Supervision. arXiv. 2021;2101.09244.

## Tables

**Table 1.** Medical terms and priority rankings in inference rules for hepatocellular carcinoma (HCC)

Medical terms and their synonyms	types	Ranking Inclusion	Ranking Exclusion	Report section
HCC, hepatocellular carcinoma, hepatoma(s)	diagnostic concepts	1		Impression
haemangioma(s/ta), cyst(s), cystic, cholecystitis	diagnostic concepts		1	Impression/ Finding/ Full report
hepatic h(a)ematoma	diagnostic concepts		2	Impression/ Finding/ Full report
metastasis, metastatic, metastasize	tumour features		3	Impression/ Finding/ Full report
(focal/ enhancing/ mass/ malignant) (liver/ hepatic/ intrahepatic) lesion/ tumour/ mass	tumour features	2		Impression/ Finding/ Full report
arterial (phase) enhancement/ hyperenhancement	tumour features	3a		Finding/ Full report
(non-peripheral) washout	tumour features	3b		Finding/ Full report
enhancing capsule	tumour features	3c		
observation size: ≤9mm, 10-19mm, ≥20mm ( <i>if specified for liver</i> )	tumour features	3d		Finding/ Full report
increase in size ( <i>if specified for liver, rather than other organs</i> )	tumour features	4		Impression/ Finding/ Full report
normal, stable, unremarkable, unchanged, asymptomatic	diagnostic concepts/ tumour features		5	Impression
hypervascularity, hypervascular (lesion/ nodule)	diagnostic concepts/ tumour features	6		Impression/ Finding/ Full report
nodule(s)	tumour features	7		Finding/ Full report

**Table 2.** Distribution of imaging modalities and HCC prevalence (labels annotated by medical students) in each site

Imaging modalities	Site 1 (n=655)		Site 2 (n=485)	
	Total	Labelled with HCC	Total	Labelled with HCC
Ultrasound	598	5 (0.84%)	314	4 (1.27%)
CT	33	6 (18.18%)	92	5 (5.43%)
MRI	24	3 (12.50%)	79	3 (3.80%)

*CT, computed tomography; HCC, hepatocellular carcinoma; ICHT, Imperial College Healthcare NHS Trust; MRI, magnetic resonance imaging.*

**Table 3.** Text characteristics of imaging reports across two sites

		Site 1 (n=655)	Site 2 (n=485)	p-value
Raw reports	Number of reports with an "impression" section <sup>†</sup>	117 (18%)	366 (75%)	<b>&lt;0.001</b>
	Number of sentences per report, median [IQR]	9 [6, 13]	11 [8, 15]	<b>&lt;0.001</b>
	Word count per report, median [IQR]	89 [65, 126]	123 [85, 178]	<b>&lt;0.001</b>
Preprocessed reports	Number of sentences per report, median [IQR]	4 [3, 5]	2 [1, 4]	<b>&lt;0.001</b>
	Word count per report, median [IQR]	38 [28, 51]	23 [11, 51]	<b>&lt;0.001</b>

<sup>†</sup>This also includes an equivalent section of 'impression', such as 'opinion', 'summary', 'comment', etc. *IQR, interquartile range.*

**Table 4.** The performance comparison between the proposed rule-based method and ML- or DL-based models in the test set (Site 2) after label correction

Methods	Sensitivity	Specificity	Precision	Accuracy	F1 score
<b>Rule-based NLP</b>	100%	99.58%	85.71%	99.59%	0.9231
<b>ML Models with BOW</b>					
NB + BOW	66.67%	91.12%	16.00%	90.52%	0.2581
SVM <sup>#</sup> + BOW	58.33%	97.46%	36.84%	96.49%	0.4516
LR + BOW	66.67%	95.77%	28.57%	95.05%	0.4000
<b>ML Models with TF-IDF</b>					
NB + TF-IDF	58.33%	94.50%	21.21%	93.61%	0.3111
SVM <sup>#</sup> + TF-IDF	58.33%	91.54%	14.89%	90.72%	0.2373
LR + TF-IDF	58.33%	92.39%	16.28%	91.55%	0.2545
<b>DL Models with word2vec</b>					
Word2Vec + CNN	41.67%	84.14%	6.25%	83.09%	0.1087
Word2Vec + HAN	75.00%	91.97%	19.15%	91.55%	0.3051
<b>DL Models with contextual representation</b>					
BERT	33.33%	99.79%	80.00%	98.14%	0.4700
ClinicalBERT	33.33%	99.58%	66.67%	97.94%	0.4444
BlueBERT	58.33%	97.67%	38.89%	96.70%	0.4667

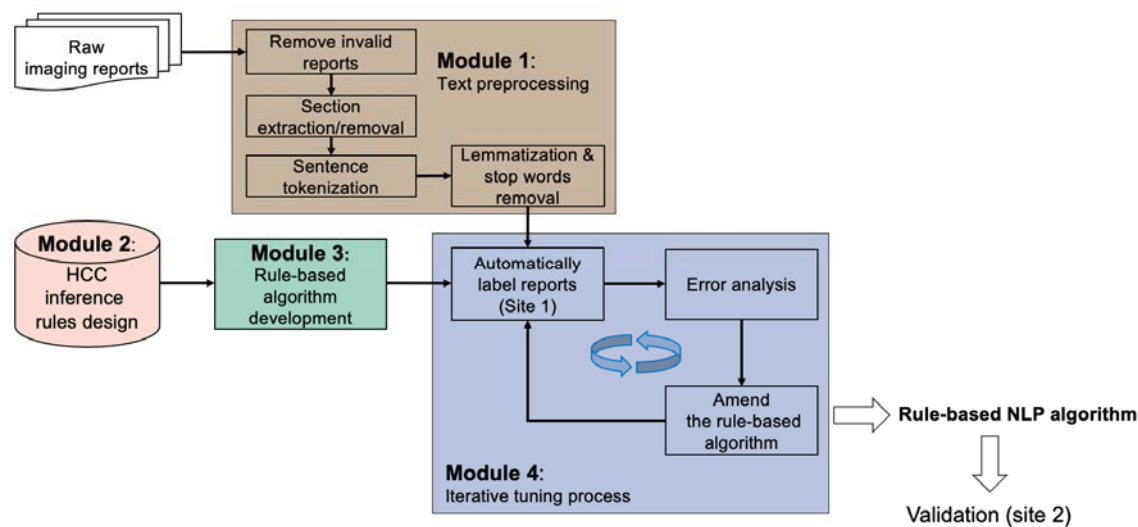
<sup>#</sup>SVM classifier with linear kernel were used.

*DL, deep learning; CNN, convolutional neural network; HAN, hierarchical attention networks; BERT, bidirectional encoder representations from transformers; BOW, bag of words; ML, machine learning;*

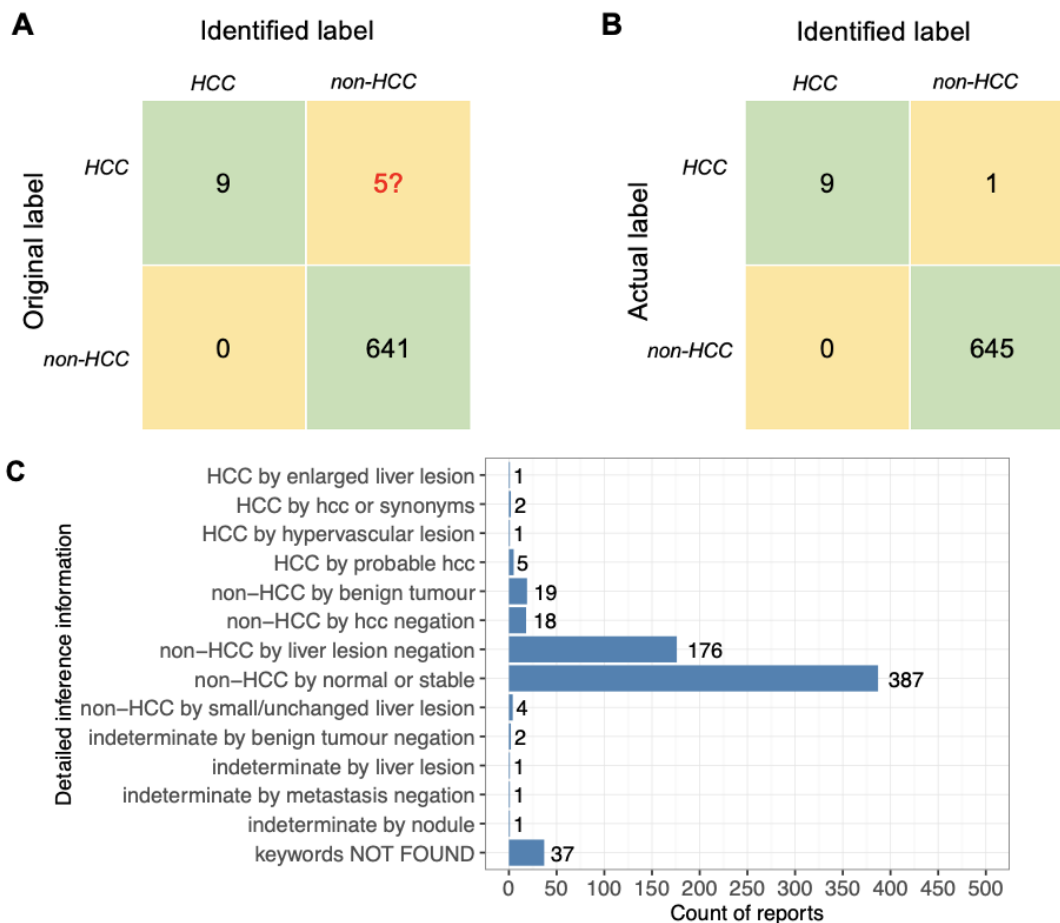


*TF-IDF*, term frequency-inverse document frequency. *TPR*, true positive rate; *TNR*, true negative rate.  
*NB*, Naive Bayes, *LR*, Logistic Regression; *SVM*, Support Vector Machine.

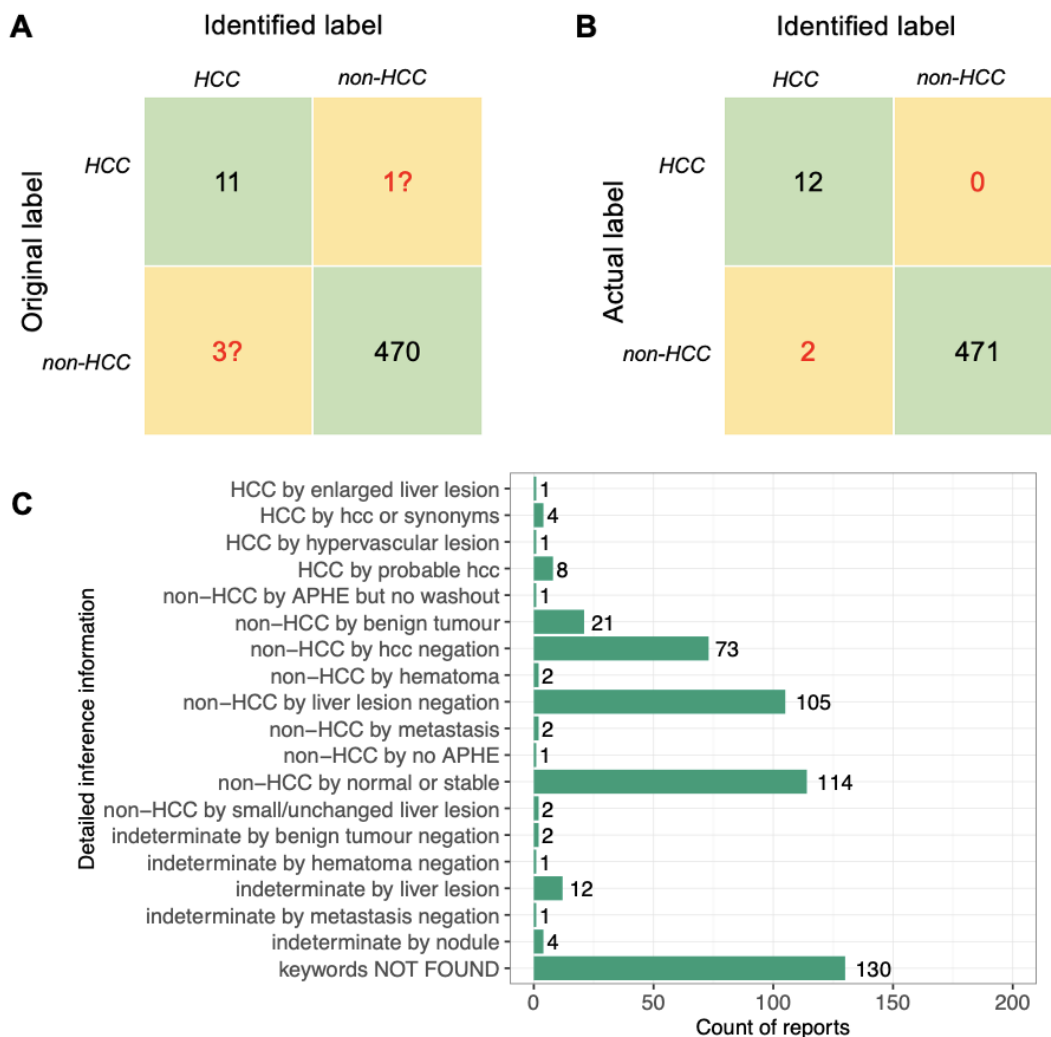
## Figures



**Figure 1.** Framework of the proposed NLP method for identifying HCC from imaging reports



**Figure 2.** Results in development set (site 1): **(A)** confusion matrix compared to raw labels; **(B)** confusion matrix compared to actual (ground-truth) labels; **(C)** Distribution of each criteria used by the algorithm in development set. *Noting that the categories of 'keywords not found' or 'indeterminate' were classified into 'non-HCC' by the algorithm.*



**Figure 3.** Results in test set (site 2): **(A)** Normalised confusion matrix compared to raw labels, **(B)** Normalised confusion matrix compared to ground-truth labels, **(C)** Distribution of each criteria used by the algorithm in test set. *Noting that the categories of 'keywords not found' or 'indeterminate' were classified into 'non-HCC' by the algorithm.*