



## King's Research Portal

DOI:  
[10.2196/15852](https://doi.org/10.2196/15852)

*Document Version*  
Peer reviewed version

[Link to publication record in King's Research Portal](#)

*Citation for published version (APA):*

Leightley, D. J., Pernet, D., Velupillai, S. U., Stewart, R. J., Mark, K. M., Opie-Bassano, E. M. T., Murphy, D., Fear, N. T., & Stevelink, S. (2020). The Development of the Military Service Identification Tool: Identifying Military Veterans in a Clinical Research Database using Natural Language Processing and Machine Learning. *JMIR Medical Informatics*, 8(5), [e15852]. <https://doi.org/10.2196/15852>

### **Citing this paper**

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

### **General rights**

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

### **Take down policy**

If you believe that this document breaches copyright please contact [librarypure@kcl.ac.uk](mailto:librarypure@kcl.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.

1           The Development of the Military Service Identification Tool: Identifying  
2           Military Veterans in a Clinical Research Database using Natural Language  
3           Processing and Machine Learning

4           Daniel Leightley<sup>1\*</sup>, David Pernet<sup>1</sup>, Sumithra Velupillai<sup>2,3</sup>, Robert J. Stewart<sup>2,3</sup>, Katharine M.  
5           Mark<sup>1</sup>, Elena Opie<sup>1</sup>, Dominic Murphy<sup>1,4</sup>, Nicola T. Fear<sup>1,5†</sup> and Sharon A. M. Stevelink<sup>1,6†</sup>

6           <sup>1</sup>King's Centre for Military Health Research, King's College London, Weston Education Centre, London,  
7           SE5 9RJ, UK;

8           <sup>2</sup>Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, SE5 9RJ, UK;

9           <sup>3</sup>South London and Maudsley NHS Foundation Trust, London, UK;

10          <sup>4</sup>Research Department, Combat Stress, Leatherhead, KT22 0BX, UK;

11          <sup>5</sup>Academic Department of Military Mental Health, King's College London, Weston Education Centre,  
12          London, SE5 9RJ, UK.

13          <sup>6</sup>Department of Psychological Medicine, King's College London, Institute of Psychiatry, Psychology and  
14          Neuroscience, London, SE5 8AF, UK.

15  
16          \*Corresponding author

17          †Joint last author

18  
19          Funding: Forces in Mind Trust (Project: FiMT18/0525KCL).

20  
21          Competing Interests: N.T.F, D.P and S.A.M.S are part funded by the United Kingdom's Ministry of  
22          Defence. N.T.F sits on the Independent Group Advising on the Release of Data at NHS Digital. N.T.F is  
23          also a trustee of two military related charities. D.M is employed by Combat Stress, a national charity  
24          in the UK that provides clinical mental health services to veterans. The views expressed are those of  
25          the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health and Social  
26          Care or the UK Ministry of Defence.

27  
28          Author Contributions: S.A.M, D.M and N.T.F conceived the concept of the study and obtained funding.  
29          DL and DP led on the natural language processing procedure. D.L, K.M.M and E.O performed data  
30          annotation. SV and R.S provided substantial improvements to the manuscript after drafting. All author  
31          reviewed the final manuscript.

32  
33          Acknowledgments: S.V, R.S SAMS' salary is partly paid by the National Institute for Health Research  
34          Biomedical Research Centre at the South London and Maudsley National Health Service Foundation  
35          Trust and King's College London. In addition to the listed authors, the study involved support from the  
36          NIHR Biomedical Research Centre. This department is a partnership between the South London and  
37          Maudsley National Health Service Foundation Trust and the Institute of Psychiatry, Psychology and  
38          Neuroscience at King's College London. We would particularly like to thank Megan Pritchard (lead in  
39          Clinical Record Interactive Search training and development), Debbie Cummings (administrator),  
40          Karen Birnie (researcher) and Larisa Maria (researcher) for their help and support in undertaking this  
41          study.

42  
43

44

45

46 **Abstract**

47 **Background:** Electronic healthcare records (EHRs) are a rich source of health-related information, with  
48 potential for secondary research use. In the United Kingdom (UK), there is no national marker for  
49 identifying those who have previously served in the Armed Forces, making analysis of the health and  
50 well-being of veterans using EHRs difficult.

51  
52 **Objective:** The aim of this study was to develop a tool to identify veterans from free-text clinical  
53 documents recorded in a psychiatric EHR database.

54  
55 **Methods:** Veterans were manually identified using the South London and Maudsley Biomedical  
56 Research Centre Clinical Record Interactive Search – a database holding secondary mental health care  
57 electronic records for the South London and Maudsley National Health Service Foundation Trust. An  
58 iterative approach was taken, first a Structured Query Language (SQL) method was developed which  
59 was then refined using Natural Language Processing and machine learning to create the Military  
60 Service Identification Tool (MSIT) to identify if a patient was a civilian or veteran. Performance, defined  
61 as correct classification of veterans compared to incorrect classification, was measured using positive  
62 predictive value, negative predictive value, sensitivity, F1 score and accuracy (otherwise termed  
63 Youden Index).

64  
65 **Results:** A gold standard dataset of 6672 free-text clinical documents were manually annotated by  
66 human coders, 66% of were then used to train the SQL and MSIT approaches, and 34% used for testing  
67 the approaches. To develop the MSIT, an iterative two-stage approach was undertaken. In the first  
68 stage, a SQL method was developed to identify veterans using a keyword rule-based approach. This  
69 approach obtained an accuracy of 0.93 in correctly predicting civilians and veterans, a positive  
70 predictive value of 0.81, a sensitivity of 0.75 and negative predictive value of 0.95. This method  
71 informed the second stage, which was the development of the MSIT using machine learning, which,  
72 when tested, obtained an accuracy of 0.97, a positive predictive value of 0.90, a sensitivity 0.91 and a  
73 negative predictive value of 0.98.

74  
75 **Conclusion:** The MSIT has the potential to be used in identifying veterans in the UK from free-text  
76 clinical documents, providing new and unique insights into the health and well-being of this  
77 population and their use of mental healthcare services.

78  
79 **Key Words:** Natural Language Processing; Machine Learning; Armed Forces; Electronic Healthcare  
80 Records; Mental Health; Veteran.

81

## 82 Introduction

83 Estimates of the United Kingdom's (UK) military veteran population, defined by the British  
84 Government as those who have served in the military for at least one day [1], is approximately 2.5  
85 million, equivalent to around 5% of household residents aged 16 years or over in the UK [2]. UK military  
86 veterans receive healthcare provision from the National Health Service (NHS) alongside civilians, with  
87 care recorded in local, regional and national Electronic Healthcare Records (EHRs) [3]. EHRs –  
88 structured and unstructured (i.e. free text) – can be used to evaluate disease prevalence, surveillance,  
89 to perform epidemiological analyses and investigate quality of care and to improve clinical decision-  
90 making [4,5].

91 Veterans of the UK experience a range of mental health problems (estimates range from 7% to 22%  
92 across psychiatric conditions), some resulting from their experiences in the line of duty [6]. A large UK  
93 cohort study set up to investigate the health of serving personnel and veterans has also shown that  
94 veterans report higher levels of probable Post-Traumatic Stress Disorder and alcohol misuse than  
95 serving personnel [7]. Recent research suggests that 93% of veterans who report having a mental  
96 health difficulty seek some form of help for their problems, including informal support through family  
97 and friends [8]. However, there is no national marker in UK EHRs to identify veterans, nor is there a  
98 requirement for healthcare professionals to record it, making it difficult to evaluate the unique  
99 healthcare needs of those who have served in the UK Armed Forces [9]. Furthermore, the ability to  
100 identify veterans would allow for comparisons between civilian and military cohorts and to allow for  
101 direct comparison of their physical and mental health.

102 In England and Wales, only two studies exist which analyse secondary care delivered through the NHS  
103 for Armed Forces personnel. In the first, Leightley *et al.* (2018) [3] developed a method to link the  
104 EHRs of military personnel in England, Scotland and Wales (three Nations of the UK). This study used  
105 a longitudinal cohort consisting of serving personnel and veterans to establish a link to national EHRs  
106 (England, Scotland and Wales). Then, statistical analyses were performed to identify the most  
107 common reasons to admission into hospital, diagnoses and treatment pathways. The second, by Mark  
108 *et al.* (2019; [10]), on which this study is based, systematically searched for veterans using a military-  
109 related search term strategy on free-text clinical documents using a manual approach. While this  
110 approach could identify veterans, it was time consuming as searches were performed manually. Each  
111 of these studies highlighted a need for novel methodological development for the identification of  
112 veterans, with natural language processing (NLP) and machine learning showing great promise [11–  
113 13]. This would enable for the automatic identification of veterans without the need for manual  
114 annotation and validation.

115 NLP approaches cover wide-ranging solutions to the analysis of text such as retrieval, analysis,  
116 transformation and classification of text, such as those found in EHR and free-text clinical documents  
117 [13,14]. NLP sub-themes, such as text mining, are represented as a set of programmatic rules or  
118 machine learning algorithms (e.g. automated learning from labelled data) to extract meaning from  
119 'naturally-occurring' text (e.g. human generated text) [11,14]. The result is often an output that can  
120 be interpreted by humans and that can be processed computationally more efficiently [15]. It may be  
121 possible to apply NLP for the identification of veterans, if not already defined from structured fields,  
122 for which, in the UK, are sparsely coded (Mark *et al*; Submitted). The ability to identify veterans at scale

123 could significantly improve our understanding of their health and well-being, navigation of care  
124 pathways and allow for the exploration of the longer-term impacts of service.

125 NLP tools have been used extensively in military health research, predominantly in the United States  
126 of America, for the detection of veteran homelessness and clinical diagnosis [16–19]. However, to the  
127 best of our knowledge, none exist to identify veteran status using either a rule-based or machine  
128 learning approaches. The aim of this work is to describe the development of the Military Service  
129 Identification Tool (MSIT) for the identification of veterans using free-text clinical documents and to  
130 evaluate the tool’s performance against a manually annotated dataset (gold standard). This work is  
131 inspired by Fernandes *et al.* (2018, [14]) but we propose a different approach to the way in which  
132 features are generated and used for training machine learning classifiers, the annotation of the  
133 training and testing data, the way in which we evaluate the performance of MSIT across different  
134 classifiers and we make publicly available our source code.

135

136 **Methods**

137 **Data Source – Clinical Record Interactive Search system**

138 The Clinical Record Interactive Search (CRIS) system provides de-identified EHRs from the South  
139 London and Maudsley (SLaM) NHS Foundation Trust, a secondary and tertiary mental healthcare  
140 provider serving a geographical catchment of roughly 1.3 million residents of four south London  
141 boroughs (Lambeth, Southwark, Lewisham, and Croydon) [20]. The CRIS system has supported a range  
142 of research projects [20–23]. Many of these have aimed to answer specific clinical or epidemiological  
143 research questions and have drawn on particular sub-populations being identified in the database –  
144 such as ethnic minorities and those with Alzheimer’s disease [24,25].

145 Ethical approval for the use of CRIS as an anonymised database for secondary analysis was granted by  
146 the Oxford Research Ethics Committee (reference: 08/H0606/71+5). The current study described here  
147 has been approved by the CRIS Patient Data Oversight Committee of the National Institute of Health  
148 Research Biomedical Research Centre (reference: 16-056).

149 The documents used in this study are ‘Correspondence’, which are created by clinical staff to provide  
150 a summary of admission/care received and are sent to a patients General Practitioner, and, in some  
151 cases, to the patient themselves. Correspondence were used as they routinely provided a detailed  
152 history of a patient’s life events including employment history.

153 **Study Design**

154 There are approximately 300,000 correspondence documents available in CRIS. Due to the large  
155 volumes of data a sub-set was extracted for the development of the MSIT. This subset (hereafter  
156 termed personal history dataset) was extracted using the Personal History Detection tool which has  
157 been developed by the CRIS team [26]. This tool identifies documents which have a sub-heading or  
158 section entitled ‘personal history’ (or similar) before extracting the proceeding text (see Extract 1 for  
159 an example). Each personal history record contains an outline of each patient’s life events since birth;  
160 these include educational attainment, childhood adversity, employment and relationship information.  
161 Each record is written by a clinician. The personal history dataset contains 98395 documents sampled  
162 from records recorded in CRIS since 2006, which was the first year the CRIS database was operational.

163 *“Mrs X was born in X. Her father was a Normandy D-Day veteran who had sustained a bullet wound*  
164 *to his left arm during the war. He subsequently worked as a bus driver in and around X. Mrs X*  
165 *describes her upbringing as old-fashioned, traditional and one of poverty. She describes her school*  
166 *years as happy and fun and says she got on well with her parents. She acknowledged that during her*  
167 *teenage years that she was difficult to manage. She met her husband X while on holiday in X; X was*  
168 *stationed there in a military unit conducting NATO exercises. After they began a relationship, in 1983,*  
169 *they moved to X. Mrs worked in various jobs including in a supermarket and as a hotel receptionist,*  
170 *before taking an administrative job in academia.”*

171 Extract 1. Synthetic generated personal history statement by the research team for a female patient  
172 who father and husband served in the military. X denotes personal identifier being removed. Due to  
173 patient confidentiality we are not able to share real examples from the personal history dataset.

174 After an informal scoping exercise, discussions with NLP experts with experience of using CRIS and  
175 timing constraints of the study, the decision was made to retain only 6672 documents (hereafter  
176 termed gold standard dataset), which represented 4200 patients (civilian: 3331, veteran: 869). A  
177 patient could have multiple documents which represent different timepoints of care. The decision to  
178 retain 4200 patients (which in total had 6672 documents) was made considering resources limitations  
179 of the study which included staff time to annotation and balancing patient privacy as to only process  
180 a minimum number of records to allow us to archive the study aim. A sample size calculation was not  
181 performed due to these considerations.

182 For evaluating the performance of MSIT, a decision was made to retain 66% (4470 documents) of the  
183 dataset for training, and the remainder 34% (2202 documents) was used for testing and evaluation.  
184 Patients and their documents were sampled either to the training or testing; a patient's documents  
185 would not appear in both samples. There is no defined approach for determining the size of the  
186 training and testing set needed, with most research using ad hoc reasoning depending on data,  
187 financial, time or personal constraints [27]. This study followed an iterative approach to the  
188 development of the MSIT, first by developing a Structured Query Language (SQL) rule-based method,  
189 with lessoned learned informing the development of MSIT, a Natural Language Processing and  
190 machine learning method.

#### 191 ***Generating the gold standard dataset and inter-rater agreement***

192 A set of classification rules for the annotation of each document were developed and agreed upon by  
193 DL, EO, DP and SAMS. The Extensible Human Oracle Suite of Tools (*eHost*) software package was used  
194 to perform annotations [28]. The following words and phrases were annotated: 1) those that  
195 described a patient's military service (i.e. 'he served in the Army'); 2) those that described an individual  
196 other than the patient's military service (i.e. 'dad served in the Forces'); and 3) those that may cause  
197 confusion (i.e. 'Navy Blue'). This led to the creation of a gold standard dataset which contained  
198 veterans and civilians annotated free-text clinical documents. Veterans were labelled as such based  
199 on a clear statement that the patient themselves had served in the military. The protocol, including  
200 classification rules, is available upon request from the corresponding author.

#### 201 ***Developing a rule-based approach for veteran identification***

202 Civilians and veterans were classified using SQL rule-based method based on a corpus of known words  
203 and phrases related to military service (See Supplementary Material). The corpus was composed of;  
204 1) primary search terms: common words or phrases used to describe military service; 2) secondary  
205 search terms: used to validate that the document describes a patient who has served in the military;  
206 3) exclusion terms: used to exclude documents that may describe an others persons military service  
207 and not the patient.

208 The SQL rule-based method was developed using a combination of the research team's expert  
209 knowledge of the military, relevant research literature and analysis of personal history statements.  
210 The gold standard training dataset was used to refine the SQL rule-based approach. The code was  
211 iteratively tested on the training set, reviewed and refined to ensure full coverage of known military  
212 words and phrases. The SQL rule-based method operated by searching for the occurrence of a primary  
213 search term in a document. If the term was found, text surrounding the term would be extracted (up  
214 to 50 characters, where available). The extracted text was then evaluated against a list of secondary

215 terms to classify the document as a civilian or veteran. The SQL rule-based approach informed the  
216 development of the MSIT.

### 217 ***Developing the Military Service Identification Tool***

218 A machine learning classification framework was used to create MSIT. It was developed in Python  
219 using the Natural Language Processing Toolkit (3.2.5) [29] and *Scikit-learn* (0.20.3) [30]. The gold  
220 standard dataset was pre-processed to remove: 1) punctuations<sup>1</sup>; 2) words/phrases<sup>2</sup> related to  
221 another individuals military service; 3) stop words and frequently occurring (except military terms);  
222 and 4) word/phrases that may cause confusion with correctly identifying a veteran. The remaining  
223 features were then converted into term frequency–inverse document frequency (tf-idf) features.

224 The classification framework was trained to identify veterans based on the use of military terms and  
225 phrases with the outcome being binary (1: veteran, 0: not a veteran). A training set of 4470 annotated  
226 documents was used to select a machine learning classifier. There is sparse literature on which  
227 machine learning algorithms are best suited for specific tasks, not only in the field of NLP but also  
228 in areas such as healthcare, agricultural and security [31–34]. To ensure the appropriate selection of  
229 classifier used for the MSIT, a comparison was made based on ten-fold cross validation accuracy using  
230 tf-idf features as an input of the following machine learning classifiers (which are part of the *Scikit-*  
231 *learn* package): Random Forest, Decision Tree, Linear Support Vector Classifier, Support Vector  
232 Classifier, Multinomial Naïve Bayes, k-Nearest Neighbour, Logistic Regression and Multi-layered  
233 Perception. Each machine learning classifier used default parameters. Linear Support Vector Classifier  
234 obtained the highest accuracy (see *Table 1*, 0.95, Standard Deviation: 0.01, 95% Confidence Interval:  
235 0.94-0.95) and was used as the machine learning classifier for MSIT.

236 To improve the *true positive rate* of the MSIT, and to reduce the potential for *false positives*, a post-  
237 processing of the Linear Support Vector Classifier outcome was applied based on the SQL rule-based  
238 approach described earlier, as has been used in similar works [14]. For each document that was  
239 predicted as being that of a veteran, a SQL operation was performed to ensure the document used a  
240 military term of phrase (e.g. ‘joined the army’, ‘left the army’, ‘demobbed from the army’).

### 241 ***Availability of materials and data***

242 The datasets used in this study are based on patient data which is not publicly available. While the  
243 data is pseudonymised, that is, patient personal details are removed, the data still contains  
244 information which could be used to identify a patient. Access to this data requires a formal application  
245 to the CRIS Patient Data Oversight Committee of the National Institute of Health Research Biomedical  
246 Research Centre. On request, and after suitable arrangements are put in place, the data and modelling  
247 employed in this study can be viewed within the secure system firewall. The corresponding author can  
248 provide more information about the process.

249 A Jupyter Notebook demonstrating the tool with artificial data can be found here ([link provided upon  
250 acceptance]).

---

<sup>1</sup> Using regular expressions.

<sup>2</sup> Words/phrases were required to exactly match those contained in the gold standard annotated dataset.



251 **Statistical analyses**

252 All analyses were performed using Python 3.5 with standard mathematical packages and *Scikit-learn*  
253 (0.20.3) [30]. Cohen's kappa values are presented for civilian and veteran annotations separately, with  
254 a two-tailed statistical test applied to determine significance of the finding. Machine learning classifier  
255 10-fold cross validation was reported as the highest accuracy obtained, with Standard Deviation and  
256 95% Confidence Interval (CI) reported to represent the *n*-fold result. Document characteristics was  
257 reported as the average frequency in which words, sentences, whitespaces, stop-words and non-  
258 alphanumeric across documents stratified by civilian and veteran. The most frequent military terms  
259 and phrases annotated during the study were restricted to the top 5 and reported as a count with  
260 percentage out of the denominator. For evaluating SQL rule-based approach, the algorithm was tested  
261 by measuring the output results against the results from manual annotations (the gold standard  
262 testing dataset) allowing for computation of positive predictive value, negative predictive value  
263 sensitivity, F1 score and accuracy at a document level. For evaluating MSIT, each classifier model was  
264 tested by measuring its results against the results from manual annotations (the gold standard testing  
265 dataset) allowing for computation of positive predictive value, negative predictive value sensitivity, F1  
266 score and accuracy at a document level.

267 In this study, positive predictive value was defined as the proportion of correctly identified *true*  
268 veterans over the total number of *true* veterans identified by the classifier. Negative predictive value  
269 was defined as the proportion of correctly identified *true* civilians over the total number of *true*  
270 civilians identified by the classifier. Sensitivity was defined as the proportion of *true* veterans identified  
271 by the classifier over the total number of actual veterans (identified by manual annotation). F1 score  
272 considers both positive predictive value and sensitivity and produces a harmonic mean, where the  
273 best value lies at 1, and the worst at 0. Accuracy was measured using Youden Index which considers  
274 sensitivity and specificity (summation minus one), which results in a value that lies between 0 (absence  
275 of accuracy) and 1 (perfect accuracy).

276

277 **Results**

278 An iterative approach to developing MSIT was employed. See Figure 1 for a flow diagram of the MSIT  
279 and evaluation process. The datasets used in this study was independently annotated by DL, EO and a  
280 researcher (see acknowledgements) with acceptable inter-rater agreement as indicated by a Cohen's  
281 kappa of 0.83 for veterans and 0.89 for civilians ( $p = 0.147$ ).

282 ***Document characteristics***

283 Of the 6672 documents annotated to generate the gold standard dataset, there were 5630 civilian  
284 and 1042 veteran documents (civilian: 3331, veteran: 869). Descriptive characteristics (see Table 2)  
285 indicate that often civilian documents had more words, sentences, stop-words and non-alphanumeric  
286 characters.

287 A total of 2611 words and 2016 phrases that describe a patient's military service were annotated (see  
288 Table 3). Most of the words and phrases annotated described the service branch (e.g. 'served in the  
289 army', 'national service in the RAF', 'demobbed from the army', 'was a pilot in the RAF'), with only a  
290 small number including the length of service (e.g. 'served for two years in the army', 'served two years  
291 for national service', 'demobbed from the army after two years').

292 ***Performance: Positive predictive value, Sensitivity and Accuracy***

293 The performance of each approach was evaluated against the manually annotated gold standard test  
294 dataset producing positive predictive value, negative predictive value, sensitivity, F1 score and  
295 accuracy statistics. The gold standard test dataset contained 2202 documents which included 1882  
296 civilian and 320 veteran documents (see Table 4).

297 The SQL rule-based approach correctly identified 262 veteran documents, incorrectly identified 87  
298 civilian documents as veteran documents, and incorrectly identified 58 civilian documents as veteran.  
299 Misclassification was due to the rigidity of the keywords used to search the records, with confusion  
300 observed between the individual's serving status and a family members status. For example, phrases  
301 such as "had served" were used to describe another person's military service, such as father or  
302 brother. This resulted in an overall accuracy of 0.93, a positive predictive value of 0.81, negative  
303 predictive value score of 0.95, a sensitivity of 0.75 and F1 score of 0.78.

304 During initial development of the MSIT, model sensitivity was skewed towards commonly occurring  
305 words. To overcome this bias, a 4-step pre-processing step was introduced to identify and remove  
306 these frequent words and phrases, punctuation and stop words which improved positive predictive  
307 value and sensitivity of the tool (training dataset: positive predictive value: 0.78; sensitivity: 0.88). To  
308 further improve the prediction of the tool and reduce the potential for *false positives*, a post-  
309 processing step was introduced to ensure a military word or phrase was present in the documents  
310 predicted as describing a veteran. The addition of this step improved positive predictive value and  
311 sensitivity of the MSIT (training dataset: positive predictive value: 0.82; sensitivity: 0.91).

312 Applying MSIT to the gold standard test dataset correctly identified 290 veteran documents,  
313 incorrectly identified 30 civilian documents as veteran documents, and incorrectly identified 27 civilian  
314 documents as being a veteran document. Misclassification was observed, with manual inspection of  
315 the documents revealing that use of military-related terms were used to describe events, occupations

316 or items for civilians such as “Legion” or “Mess Hall”. This created confusion with the classifier. This  
317 may be due to the clinician potentially being former military thus using military vernacular, or the  
318 patient being aware of military terminology. This resulted in an overall accuracy of 0.97, a positive  
319 predictive value of 0.90, negative predictive value of 0.95, a sensitivity of 0.91 and F1 score of 0.91.  
320 Additional analyses were conducted using leave-one-out methodology, please see Supplementary  
321 Material.

322

323 **Discussion**

324 This research has demonstrated that it is possible to identify veterans from free-text clinical  
325 documents using NLP. A tool to identify veterans and civilians is described, which performed well, as  
326 indicated by high positive predictive value, sensitivity and accuracy results. To the authors' knowledge,  
327 this is the only study to have developed, applied and tested NLP for the identification of veterans in  
328 the UK using a large psychiatric database. The MSIT presented superior results to the SQL rule-based  
329 approach developed, due to the former's ability to adapt to different military terms. The SQL rule-  
330 based approach was, on the other hand, fixed on set keywords.

331 This study is the first that seeks to identify military veterans from a case register in the UK using NLP  
332 and machine learning. Although military literature is sparse, NLP techniques have been used in the  
333 detection of sexual trauma, temporal expressions in medical narratives and for screening  
334 homelessness [16,17,19]. While it is difficult to compare our study to the aforementioned studies  
335 similar methodologies are employed. This includes each developing a gold standard (annotated  
336 dataset) manually annotated dataset, developing a set of rules to support identification and finally  
337 generated features from free-text. While this study used Linear Support Vector Classification, as it was  
338 determined to be the most optimal, Reeves *et al.* (2013; [16]) used a maximum entropy classifier to  
339 detect temporal expressions. Outside of the military literature, Fernandes *et al.* (2018) sought to  
340 identify suicidal attempts using a psychiatric database with Support Vector Machines, they were able  
341 to detect suicidal attempt with a sensitivity of 0.98, which is higher than what was achieved in this  
342 study (MSIT: 0.91). Other studies have compared different classification algorithms for clinical NLP  
343 tasks with varying conclusions – achieving optimal performance is highly task- and use-case dependent  
344 [35,36].

345 The ability to identify veterans could provide insights into the physical and mental health of military  
346 personnel and their navigation through, and use of, healthcare services including primary and  
347 secondary services. This would overcome the current need to either manually identify veterans, or to  
348 perform large-scale cohort and data linkage studies, such as that by Leightley *et al.* (2018; [3]). EHR-  
349 based case registers, such as CRIS, function as single, complete and integrated electronic versions of  
350 traditional paper health records [3]. These registers have been positioned as a 'new generation' for  
351 health research and are now mandatory in the UK [3]. The methodological advantages of case registers  
352 – including their longitudinal nature, largely structured fields and detailed coverage of defined  
353 populations – make them an ideal research and surveillance tool [37]. EHRs in mental health care  
354 provide extremely rich material and analysis of their data can reveal patterns in healthcare provisions,  
355 patient profiles and mental and physical health problems [3,38]. This is hugely advantageous for  
356 investigating vulnerable sub-groups within the wider population [20–22], potential for developing  
357 digital interventions [39] and to support data-driven decision making [11].

358 ***Strengths and limitations***

359 An important strength of this work was the exploitation of NLP, which is advantageous for automating  
360 the process of identification and reducing the possibility of human error and bias. Considering the  
361 current research focus, this is the first time that NLP has successfully been used to identify veterans  
362 from free-text clinical documents using detailed occupational history that clinicals routinely record.  
363 The MSIT described in this work does not rely on any codes (clinical or otherwise) or structured fields,  
364 which broadens its application to others, such as diagnosis and occupation detection. Further,

365 veterans may not always be willing, or think it is necessary to state their veteran status, particularly in  
366 the UK, which has no department for veterans' affairs. As such, NLP is advantageous as it may pick up  
367 veterans based on small details that are discussed and recorded during clinical interactions rather than  
368 having to reply on disclose of veteran status by an individual upon registration with clinical services.

369 It must be noted that there are several limitations to the tool described in this work. First, the study  
370 relied on patients' self-reporting that they have served in the military, which could be influenced by  
371 the patient's mental health or failing memory. Second, the need for a clinician to ask a patient's  
372 military status. Third, the accuracy of recording by the clinician could have had a negative impact on  
373 MSIT's performance, or results in misidentification of veterans. Fourth, the MSIT relied upon personal  
374 history section being present in a correspondence which may limit scalability. Fifth, while different  
375 approaches to stating veteran service were annotated, spelling and additional permutations were not  
376 considered. This could limit generalisability of the algorithms on other datasets. Sixth, identified  
377 veterans were not validated against Ministry of Defence databases or contacted directly to validate  
378 veteran status. Seventh, a sample size calculation was not computed for this study. This was due to  
379 resource limitations, as a result this could limit the generalisability of the algorithms on other datasets.  
380 Finally, documents were misclassified, often due to military vernacular being used by civilians and/or  
381 the clinician, or that a family member had served and not the patient. Further work should be  
382 undertaken to improve reliability and reducing the rate of misclassification.

### 383 ***Conclusions***

384 We have shown that it is possible to identify veterans using either a SQL-based or NLP and machine  
385 learning based approach. Both approaches are robust in correctly identifying civilians and veterans,  
386 with high accuracy, sensitivity and negative predictive values observed. The MSIT has the potential to  
387 be used in identifying veterans in the UK from free-text clinical documents, providing new and unique  
388 insights into the health and well-being of this population and their use of mental healthcare services.  
389 Despite our success in the current work, the tools are tailored to the CRIS dataset and future work is  
390 needed to develop a more agnostic framework.

391

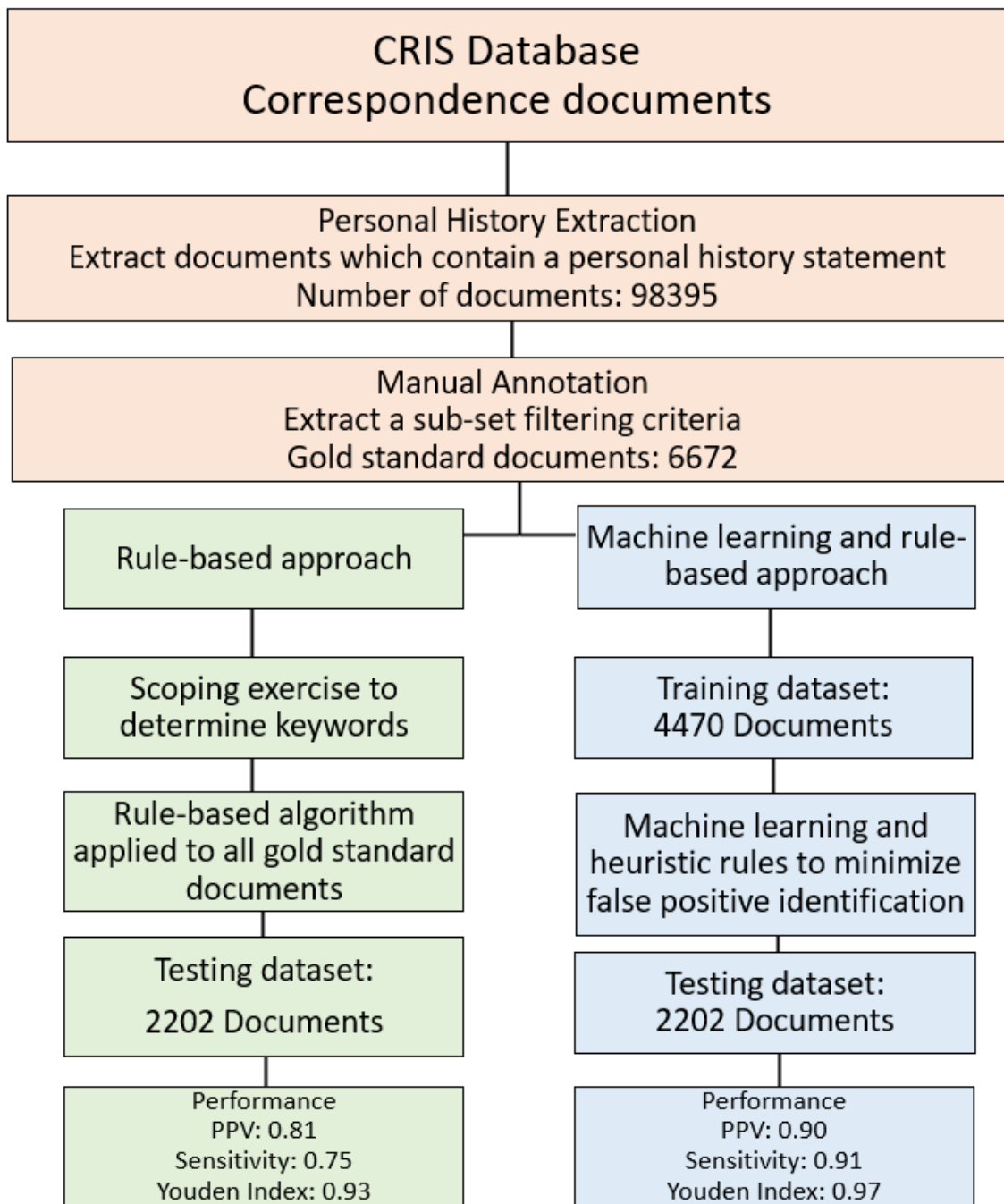
392 **References**

- 393 1. Veterans: Key facts [Internet]. Ministry of Defence; 2016.
- 394 2. Population Projections: UK Armed Forces Veterans residing in Great Britain, 2016 to 2028  
395 [Internet]. London, UK; 2019.
- 396 3. Leightley D, Chui Z, Jones M, Landau S, McCrone P, Hayes RD, et al. Integrating electronic  
397 healthcare records of armed forces personnel: Developing a framework for evaluating health  
398 outcomes in England, Scotland and Wales. *Int J Med Inform.* 2018;113:17–25.
- 399 4. Payne RA, Abel GA, Guthrie B, Mercer SW. The effect of physical multimorbidity, mental  
400 health conditions and socioeconomic deprivation on unplanned admissions to hospital: a  
401 retrospective cohort study. *Can Med Assoc J.* 2013;185(5):E221–E228.
- 402 5. Simmonds SJ, Syddall HE, Walsh B, Evandrou M, Dennison EM, Cooper C, et al. Understanding  
403 NHS hospital admissions in England: linkage of Hospital Episode Statistics to the Hertfordshire  
404 Cohort Study. *Age Ageing.* 2014;43(5):653–660.
- 405 6. Stevelink SAM, Jones M, Hull L, Pernet D, MacCrimmon S, Goodwin L, et al. Mental health  
406 outcomes at the end of the British involvement in the Iraq and Afghanistan conflicts: a cohort  
407 study. *Br J Psychiatry.* 2018;213(6):1–8.
- 408 7. Fear NT, Jones M, Murphy D, Hull L, Iversen AC, Coker B, et al. What are the consequences of  
409 deployment to Iraq and Afghanistan on the mental health of the UK armed forces? A cohort  
410 study. *Lancet.* 2010;375(9728):1783–1797.
- 411 8. Stevelink SAM, Jones N, Jones M, Dyball D, Khera CK, Pernet D, et al. Do serving and ex-  
412 serving personnel of the UK armed forces seek help for perceived stress, emotional or mental  
413 health problems? *Eur J Psychotraumatol.* 2019;10(1):1556552. PMID: 30693074
- 414 9. Morgan VA, Jablensky A V. From inventory to benchmark: quality of psychiatric case registers  
415 in research. *Br J Psychiatry.* 2010;197(01):8–10.
- 416 10. Mark KM, Leightley D, Pernet D, Murphy D, Stevelink SAM, Fear NT. Identifying Veterans  
417 Using Electronic Health Records in the United Kingdom: A Feasibility Study. *Healthcare.*  
418 2019;8(1):1.
- 419 11. Leightley D, Williamson V, Darby J, Fear NT. Identifying probable post-traumatic stress  
420 disorder: applying supervised machine learning to data from a UK military cohort. *J Ment  
421 Heal.* 2019;28(1):34–41.
- 422 12. Karstoft K-I, Statnikov A, Andersen SB, Madsen T, Galatzer-Levy IR. Early identification of  
423 posttraumatic stress following military deployment: Application of machine learning methods  
424 to a prospective study of Danish soldiers. *J Affect Disord.* 2015;184:170–175.
- 425 13. Cambria E, White B. Jumping NLP Curves: A Review of Natural Language Processing Research.  
426 *IEEE Comput Intell Mag.* 2014;9(2):48–57.
- 427 14. Fernandes AC, Dutta R, Velupillai S, Sanyal J, Stewart R, Chandran D. Identifying Suicide  
428 Ideation and Suicidal Attempts in a Psychiatric Clinical Research Database using Natural  
429 Language Processing. *Sci Rep.* 2018;8(1):7426.
- 430 15. Dalianis H. *Clinical Text Mining* [Internet]. Cham: Springer International Publishing; 2018.

- 431 16. Reeves RM, Ong FR, Matheny ME, Denny JC, Aronsky D, Gobbel GT, et al. Detecting temporal  
432 expressions in medical narratives. *Int J Med Inform.* 2013;82(2):118–127.
- 433 17. Gundlapalli A V, Carter ME, Palmer M, Ginter T, Redd A, Pickard S, et al. Using natural  
434 language processing on the free text of clinical documents to screen for evidence of  
435 homelessness among US veterans. *AMIA Annu Symp Proc.* 2013;2013:537–46. PMID:  
436 24551356
- 437 18. Mowery DL, Chapman BE, Conway M, South BR, Madden E, Keyhani S, et al. Extracting a  
438 stroke phenotype risk factor from Veteran Health Administration clinical reports: an  
439 information content analysis. *J Biomed Semantics.* 2016;7(1):26.
- 440 19. Gundlapalli A V., Jones AL, Redd A, Divita G, Brignone E, Pettey WBP, et al. Combining Natural  
441 Language Processing of Electronic Medical Notes With Administrative Data to Determine  
442 Racial/Ethnic Differences in the Disclosure and Documentation of Military Sexual Trauma in  
443 Veterans. *Med Care.* 2019;57:S149–S156.
- 444 20. Perera G, Broadbent M, Callard F, Chang C-K, Downs J, Dutta R, et al. Cohort profile of the  
445 South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC)  
446 Case Register: current status and recent enhancement of an Electronic Mental Health Record-  
447 derived data resource. *BMJ Open.* 2016;6(3):e008721.
- 448 21. Downs JM, Ford T, Stewart R, Epstein S, Shetty H, Little R, et al. An approach to linking  
449 education, social care and electronic health records for children and young people in South  
450 London: a linkage study of child and adolescent mental health service data. *BMJ Open.*  
451 2019;9(1):e024355.
- 452 22. Velupillai S, Hadlaczy G, Baca-Garcia E, Gorrell GM, Werbeloff N, Nguyen D, et al. Risk  
453 Assessment Tools and Data-Driven Approaches for Predicting and Preventing Suicidal  
454 Behavior. *Front Psychiatry.* 2019;10:36.
- 455 23. Jackson RG, Patel R, Jayatilleke N, Kolliakou A, Ball M, Gorrell G, et al. Natural language  
456 processing to extract symptoms of severe mental illness from clinical text: the Clinical Record  
457 Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ Open.*  
458 2017;7(1):e012012.
- 459 24. Kovalchuk Y, Stewart R, Broadbent M, Hubbard TJP, Dobson RJB. Analysis of diagnoses  
460 extracted from electronic health records in a large mental health case register. Abe T, editor.  
461 *PLoS One.* 2017;12(2):e0171526.
- 462 25. Mueller C, Perera G, Hayes RD, Shetty H, Stewart R. Associations of acetylcholinesterase  
463 inhibitor treatment with reduced mortality in Alzheimer’s disease: a retrospective survival  
464 analysis. *Age Ageing.* 2018;47(1):88–94.
- 465 26. NIHR Biomedical Research Centre (BRC) - King’s College London [Internet]. 2019.
- 466 27. Juckett D. A method for determining the number of documents needed for a gold standard  
467 corpus. *J Biomed Inform.* 2012;45(3):460–70. PMID: 22245601
- 468 28. Leng CJ, South B, Shen S. Extensible Human Oracle Suite of Tools. University of Utah and SLC  
469 VA; 2011.
- 470 29. Loper E, Bird S. NLTK. *Proc ACL-02 Work Eff tools Methodol Teach Nat Lang Process Comput*  
471 *Linguist - . Morristown, NJ, USA: Association for Computational Linguistics; 2002. p. 63–70.*

- 472 30. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn:  
473 Machine Learning in Python. *J Mach Learn Res. JMLR.org*; 2011;12:2825–2830.
- 474 31. Leightley D, Darby J, Baihua Li, McPhee JS, Moi Hoon Yap. Human Activity Recognition for  
475 Physical Rehabilitation. 2013 IEEE Int Conf Syst Man, Cybern. IEEE; 2013. p. 261–266.
- 476 32. Leightley D, McPhee JS, Yap MH. Automated Analysis and Quantification of Human Mobility  
477 Using a Depth Sensor. *IEEE J Biomed Heal Informatics*. 2017;21(4):939–948.
- 478 33. Ahad MAR, Tan JK, Kim HS, Ishikawa S. Human activity recognition: Various paradigms. 2008  
479 Int Conf Control Autom Syst. COEX, Seoul, Korea: IEEE; 2008. p. 1896–1901.
- 480 34. Cunningham R, Sánchez M, May G, Loram I. Estimating Full Regional Skeletal Muscle Fibre  
481 Orientation from B-Mode Ultrasound Images Using Convolutional, Residual, and  
482 Deconvolutional Neural Networks. *J Imaging*. 2018;4(2):29.
- 483 35. López Pineda A, Ye Y, Visweswaran S, Cooper GF, Wagner MM, Tsui F (Rich). Comparison of  
484 machine learning classifiers for influenza detection from emergency department free-text  
485 reports. *J Biomed Inform*. 2015;58:60–69.
- 486 36. Cronin RM, Fabbri D, Denny JC, Rosenbloom ST, Jackson GP. A comparison of rule-based and  
487 machine learning approaches for classifying patient portal messages. *Int J Med Inform*.  
488 2017;105:110–120.
- 489 37. Stewart R. The big case register. *Acta Psychiatr Scand*. 2014;
- 490 38. Stewart R, Soremekun M, Perera G, Broadbent M, Callard F, Denis M, et al. The South London  
491 and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register:  
492 development and descriptive data. *BMC Psychiatry*. 2009;9(1):51.
- 493 39. Wickersham A, Petrides PM, Williamson V, Leightley D. Efficacy of mobile application  
494 interventions for the treatment of post-traumatic stress disorder: A systematic review. *Digit  
495 Heal*. 2019;5:205520761984298.
- 496
- 497





499

500  
501

Figure 1: Flow diagram of the Military Service Identification Tool. Correspondences are used to define any communications between a patient and clinical staff or between clinical staff members.

502

503 Table 1: Machine learning classifier *n*-fold cross validation accuracy, Standard Deviation (SD) and 95% Confidence Interval  
 504 (CI) based on the gold standard training dataset (n=4470).

Classifier	Accuracy (SD, 95% CI)
Random Forest	0.84 (0.01, 0.83-0.84)
Decision Tree	0.91 (0.03, 0.89-0.92)
Linear Support Vector Classifier	<b>0.95 (0.01, 0.94-0.95)</b>
Support Vector Classifier	0.84 (0.01, 0.83-0.84)
Multinomial Naïve Bayes	0.90 (0.02, 0.88-0.91)
k-Nearest Neighbour	0.89 (0.02, 0.87-0.90)
Logistic Regression	0.88 (0.04, 0.85-0.90)
Multi-layered Perception	0.94 (0.02, 0.92-0.95)

505

506 Table 2: Document characteristics including frequency (*n*) and Standard Deviation (SD) for annotated personal history  
 507 statements stratified by civilian and veteran status.

Characteristic	Civilian (n=5630)	Veteran (n=1042)
	<i>average n (SD)</i>	<i>average n (SD)</i>
Words	223.76 (152.30)	197.20 (114.63)
Sentences	13.80 (8.91)	12.40 (6.50)
Whitespaces	237.99 (162.77)	208.38 (119.65)
Stop-words	32.04 (11.45)	30.09 (9.92)
Non-alphanumeric characters	26.59 (20.14)	22.22 (14.28)

508

509 Table 3: Top 5 occurring military word and phrases identified during manual annotation of the gold standard training  
 510 dataset.

Military Words (n=2611)		Military Phrases (n=2016)	
Word	Frequency (n/%)	Phrase	Frequency (n/%)
Army	553 (21.20)	Joined the army	167 (8.33)
National Service	445 (17.08)	Left the army	122 (6.07)
RAF	225 (8.65)	Demobbed from the army	101 (5.01)
Navy	166 (6.36)	National service in the army	65 (3.24)
Veteran	104 (3.98)	Two years in the army	64 (3.19)

511

512 Table 4: SQL-based approach and Military Service Identification Tool performance result comparison for detecting veterans  
 513 using the gold standard test dataset. The Military Service Identification Tool includes pre- and post-processing.

	SQL rule-based approach		Military Service Identification Tool	
	Veteran	Civilian	Veteran	Civilian
Veteran	262	58	290	30
Civilian	87	1795	27	1855
<b>Performance</b>				
Positive predictive value	0.81		0.90	

<b>Negative predictive value</b>	0.95	0.98
<b>Sensitivity</b>	0.75	0.91
<b>F1 score</b>	0.78	0.91
<b>Youden Index</b>	0.93	0.97