DOI:
10.2196/15852

Document Version
Peer reviewed version

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

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

Daniel Leightley[1*], David Pernet[1], Sumithra Velupillai[2,3], Robert J. Stewart[2,3], Katharine M. Mark[1], Elena Opie[1], Dominic Murphy[1,4], Nicola T. Fear[1,5ǂ] and Sharon A. M. Stevelink[1,6ǂ]

[1]King's Centre for Military Health Research, King's College London, Weston Education Centre, London, SE5 9RJ, UK;
[2]Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, SE5 9RJ, UK;
[3]South London and Maudsley NHS Foundation Trust, London, UK;
[4]Research Department, Combat Stress, Leatherhead, KT22 0BX, UK;
[5]Academic Department of Military Mental Health, King's College London, Weston Education Centre, London, SE5 9RJ, UK.
[6]Department of Psychological Medicine, King's College London, Institute of Psychiatry, Psychology and Neuroscience, London, SE5 8AF, UK.

*Corresponding author
ǂJoint last author

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

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

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

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

Abstract

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

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

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

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

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

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

Introduction

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

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

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

NLP approaches cover wide-ranging solutions to the analysis of text such as retrieval, analysis, transformation and classification of text, such as those found in EHR and free-text clinical documents [13,14]. NLP sub-themes, such as text mining, are represented as a set of programmatic rules or machine learning algorithms (e.g. automated learning from labelled data) to extract meaning from 'naturally-occurring' text (e.g. human generated text) [11,14]. The result is often an output that can be interpreted by humans and that can be processed computationally more efficiently [15]. It may be possible to apply NLP for the identification of veterans, if not already defined from structured fields, for which, in the UK, are sparely coded (Mark *et* al; Submitted). The ability to identify veterans at scale could significantly improve our understanding of their health and well-being, navigation of care pathways and allow for the exploration of the longer-term impacts of service.

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

Methods

Data Source – Clinical Record Interactive Search system

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

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

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

Study Design

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

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

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

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

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

Generating the gold standard dataset and inter-rater agreement

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

Developing a rule-based approach for veteran identification

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

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

Developing the Military Service Identification Tool

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

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

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

Availability of materials and data

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

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

---

[1] Using regular expressions.
[2] Words/phrases were required to exactly match those contained in the gold standard annotated dataset.

Statistical analyses

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

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

Results

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

Document characteristics

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

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

Performance: Positive predictive value, Sensitivity and Accuracy

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

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

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

Applying MSIT to the gold standard test dataset correctly identified 290 veteran documents, incorrectly identified 30 civilian documents as veteran documents, and incorrectly identified 27 civilian documents as being a veteran document. Misclassification was observed, with manual inspection of the documents revelling that use of military-related terms were used to describe events, occupations or items for civilians such as “Legion” or “Mess Hall”. This created confusion with the classifier. This may be due to the clinician potentially being former military thus using military vernacular, or the patient being aware of military terminology. This resulted in an overall accuracy of 0.97, a positive predictive value of 0.90, negative predictive value of 0.95, a sensitivity of 0.91 and F1 score of 0.91. Additional analyses were conducted using leave-one-out methodology, please see Supplementary Material.

Discussion

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

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

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

Strengths and limitations

An important strength of this work was the exploitation of NLP, which is advantageous for automating the process of identification and reducing the possibility of human error and bias. Considering the current research focus, this is the first time that NLP has successfully been used to identify veterans from free-text clinical documents using detailed occupational history that clinicals routinely record. The MSIT described in this work does not rely on any codes (clinical or otherwise) or structured fields, which broadens its application to others, such as diagnosis and occupation detection. Further, veterans may not always be willing, or think it is necessary to state their veteran status, particularly in the UK, which has no department for veterans' affairs. As such, NLP is advantageous as it may pick up veterans based on small details that are discussed and recorded during clinical interactions rather than having to reply on disclose of veteran status by an individual upon registration with clinical services.

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

Conclusions

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

References

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

16. Reeves RM, Ong FR, Matheny ME, Denny JC, Aronsky D, Gobbel GT, et al. Detecting temporal expressions in medical narratives. Int J Med Inform. 2013;82(2):118–127.

17. Gundlapalli A V, Carter ME, Palmer M, Ginter T, Redd A, Pickard S, et al. Using natural language processing on the free text of clinical documents to screen for evidence of homelessness among US veterans. AMIA Annu Symp Proc. 2013;2013:537–46. PMID: 24551356

18. Mowery DL, Chapman BE, Conway M, South BR, Madden E, Keyhani S, et al. Extracting a stroke phenotype risk factor from Veteran Health Administration clinical reports: an information content analysis. J Biomed Semantics. 2016;7(1):26.

19. Gundlapalli A V., Jones AL, Redd A, Divita G, Brignone E, Pettey WBP, et al. Combining Natural Language Processing of Electronic Medical Notes With Administrative Data to Determine Racial/Ethnic Differences in the Disclosure and Documentation of Military Sexual Trauma in Veterans. Med Care. 2019;57:S149–S156.

20. Perera G, Broadbent M, Callard F, Chang C-K, Downs J, Dutta R, et al. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. BMJ Open. 2016;6(3):e008721.

21. Downs JM, Ford T, Stewart R, Epstein S, Shetty H, Little R, et al. An approach to linking education, social care and electronic health records for children and young people in South London: a linkage study of child and adolescent mental health service data. BMJ Open. 2019;9(1):e024355.

22. Velupillai S, Hadlaczky G, Baca-Garcia E, Gorrell GM, Werbeloff N, Nguyen D, et al. Risk Assessment Tools and Data-Driven Approaches for Predicting and Preventing Suicidal Behavior. Front Psychiatry. 2019;10:36.

23. Jackson RG, Patel R, Jayatilleke N, Kolliakou A, Ball M, Gorrell G, et al. Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. BMJ Open. 2017;7(1):e012012.

24. Kovalchuk Y, Stewart R, Broadbent M, Hubbard TJP, Dobson RJB. Analysis of diagnoses extracted from electronic health records in a large mental health case register. Abe T, editor. PLoS One. 2017;12(2):e0171526.

25. Mueller C, Perera G, Hayes RD, Shetty H, Stewart R. Associations of acetylcholinesterase inhibitor treatment with reduced mortality in Alzheimer’s disease: a retrospective survival analysis. Age Ageing. 2018;47(1):88–94.

26. NIHR Biomedical Research Centre (BRC) - King’s College London [Internet]. 2019.

27. Juckett D. A method for determining the number of documents needed for a gold standard corpus. J Biomed Inform. 2012;45(3):460–70. PMID: 22245601

28. Leng CJ, South B, Shen S. Extensible Human Oracle Suite of Tools. University of Utah and SLC VA; 2011.

29. Loper E, Bird S. NLTK. Proc ACL-02 Work Eff tools Methodol Teach Nat Lang Process Comput Linguist -. Morristown, NJ, USA: Association for Computational Linguistics; 2002. p. 63–70.

30. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. JMLR.org; 2011;12:2825–2830.

31. Leightley D, Darby J, Baihua Li, McPhee JS, Moi Hoon Yap. Human Activity Recognition for Physical Rehabilitation. 2013 IEEE Int Conf Syst Man, Cybern. IEEE; 2013. p. 261–266.

32. Leightley D, McPhee JS, Yap MH. Automated Analysis and Quantification of Human Mobility Using a Depth Sensor. IEEE J Biomed Heal Informatics. 2017;21(4):939–948.

33. Ahad MAR, Tan JK, Kim HS, Ishikawa S. Human activity recognition: Various paradigms. 2008 Int Conf Control Autom Syst. COEX, Seoul, Korea: IEEE; 2008. p. 1896–1901.

34. Cunningham R, Sánchez M, May G, Loram I. Estimating Full Regional Skeletal Muscle Fibre Orientation from B-Mode Ultrasound Images Using Convolutional, Residual, and Deconvolutional Neural Networks. J Imaging. 2018;4(2):29.

35. López Pineda A, Ye Y, Visweswaran S, Cooper GF, Wagner MM, Tsui F (Rich). Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. J Biomed Inform. 2015;58:60–69.

36. Cronin RM, Fabbri D, Denny JC, Rosenbloom ST, Jackson GP. A comparison of rule-based and machine learning approaches for classifying patient portal messages. Int J Med Inform. 2017;105:110–120.

37. Stewart R. The big case register. Acta Psychiatr Scand. 2014;

38. Stewart R, Soremekun M, Perera G, Broadbent M, Callard F, Denis M, et al. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. BMC Psychiatry. 2009;9(1):51.

39. Wickersham A, Petrides PM, Williamson V, Leightley D. Efficacy of mobile application interventions for the treatment of post-traumatic stress disorder: A systematic review. Digit Heal. 2019;5:205520761984298.

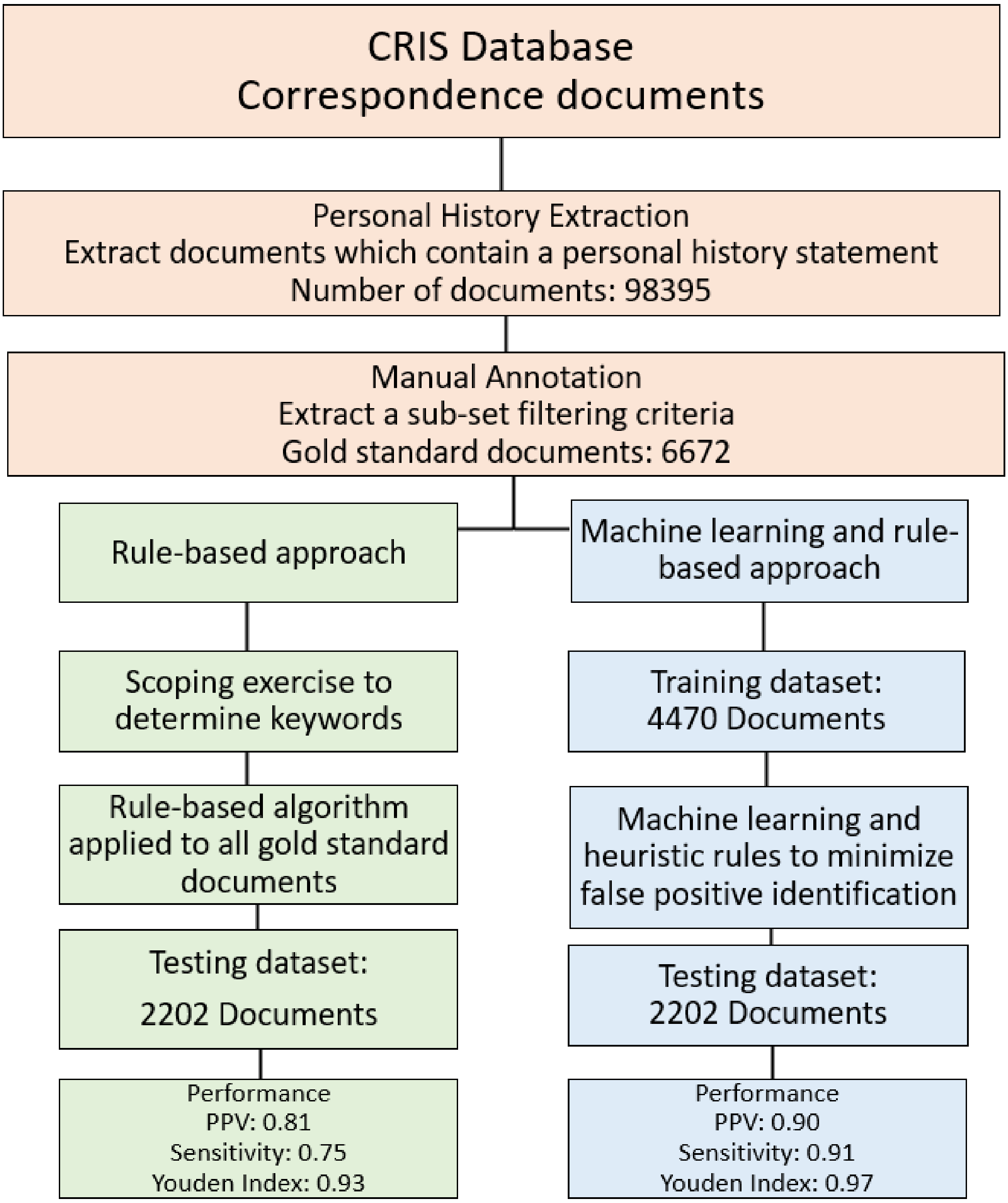

Figure 1: Flow diagram of the Military Service Identification Tool. Correspondences are used to define any communications between a patient and clinical staff or between clinical staff members.

Table 1: Machine learning classifier *n*-fold cross validation accuracy, Standard Deviation (SD) and 95% Confidence Interval (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	**0.95 (0.01, 0.94-0.95)**
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)

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

Characteristic	Civilian (*n*=5630)	Veteran (*n*=1042)
	average n (SD)	*average* n (SD)
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)

Table 3: Top 5 occurring military word and phrases identified during manual annotation of the gold standard training 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)

Table 4: SQL-based approach and Military Service Identification Tool performance result comparison for detecting veterans 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
Performance				
Positive predictive value	0.81		0.90	

Negative predictive value	0.95	0.98
Sensitivity	0.75	0.91
F1 score	0.78	0.91
Youden Index	0.93	0.97