Identifying Military Veterans in a Clinical Research Database using Natural Language Processing

Dr Daniel Leightley

King’s Centre for military Health Research

@_Dr_Daniel | @kcmhr

forces in mind trust

SUCCESSFUL SUSTAINABLE TRANSITION

COMMUNITY FUND

KING'S CENTRE FOR MILITARY HEALTH RESEARCH

ACADEMIC DEPARTMENT OF MILITARY MENTAL HEALTH

Introduction

Estimates of the UK's veteran population range from 2.5 – 5million

Between 7-22% of veterans experience psychiatric conditions

86% of serving and ex-serving personnel seek some form of support (e.g. from friends and family)

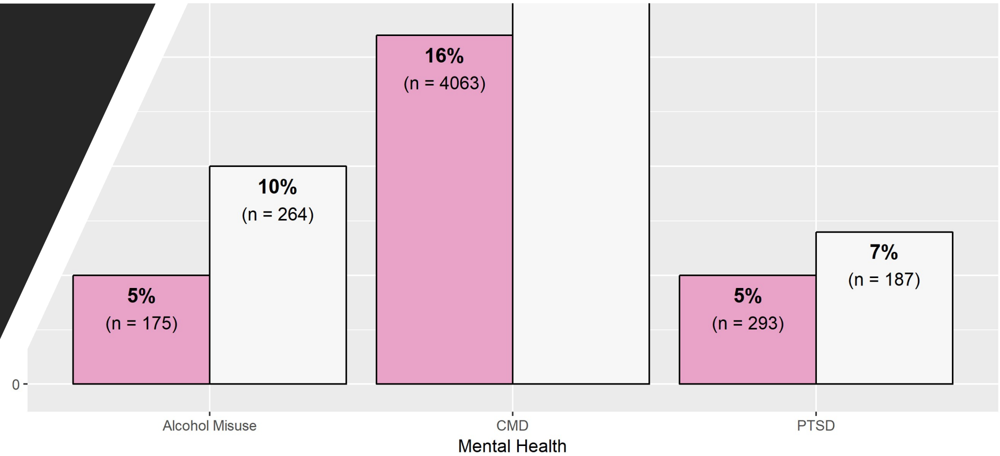
BUT only 55% of serving and ex-serving personnel seek formal medical support

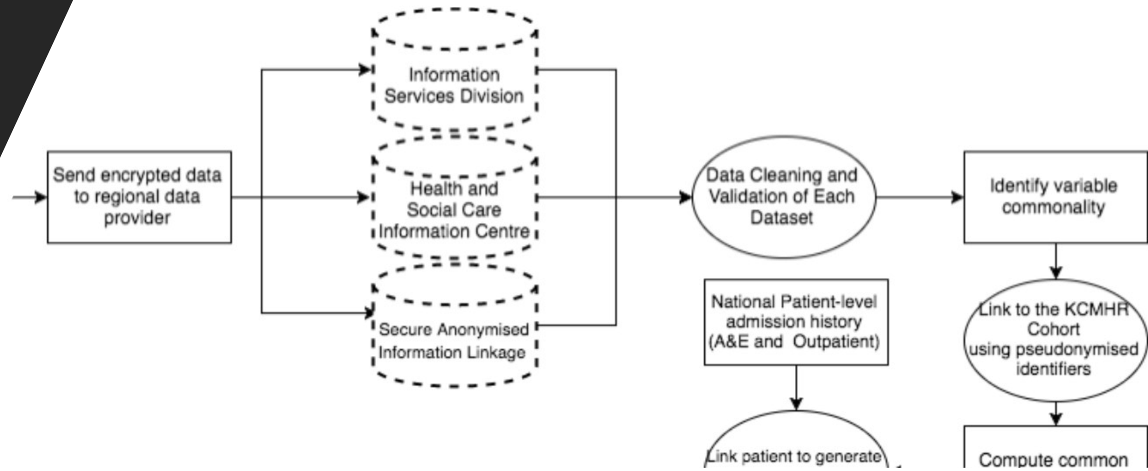
There is <u>no national marker in England and Wales</u> for identifying veterans in these records.

Previous KCMHR research

KCMHR Cohort: Data linkage between EHRs of England, Scotland and Wales and the KCMHR cohort study. Outpatients, Admitted Patient Care and Accident & Emergency. **NHS number required for linkage**.

APMS comparison: Veteran data extracted from the KCMHR cohort was compared to the Adult Psychiatric Morbidity Survey and the UK Household Longitudinal Study. **Limitation: Self-report and anonymised.**





What about CRIS?

- Veteran status not routinely collected (optional)
- Potential to be recorded in **10+** documents types
- Best source for veteran status: **free text clinical notes**

Important to develop a scalable and automatic approach

Psychiatric History

- Personal details
- Source, mode and reason for referral
- Presenting complaint
- History of presenting complaint
- Past psychiatric history
- Past medical history
- Family history
- Personal history
- Social history
- Drug & Alcohol history
- Forensic history
- Premorbid personality

Example

"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."

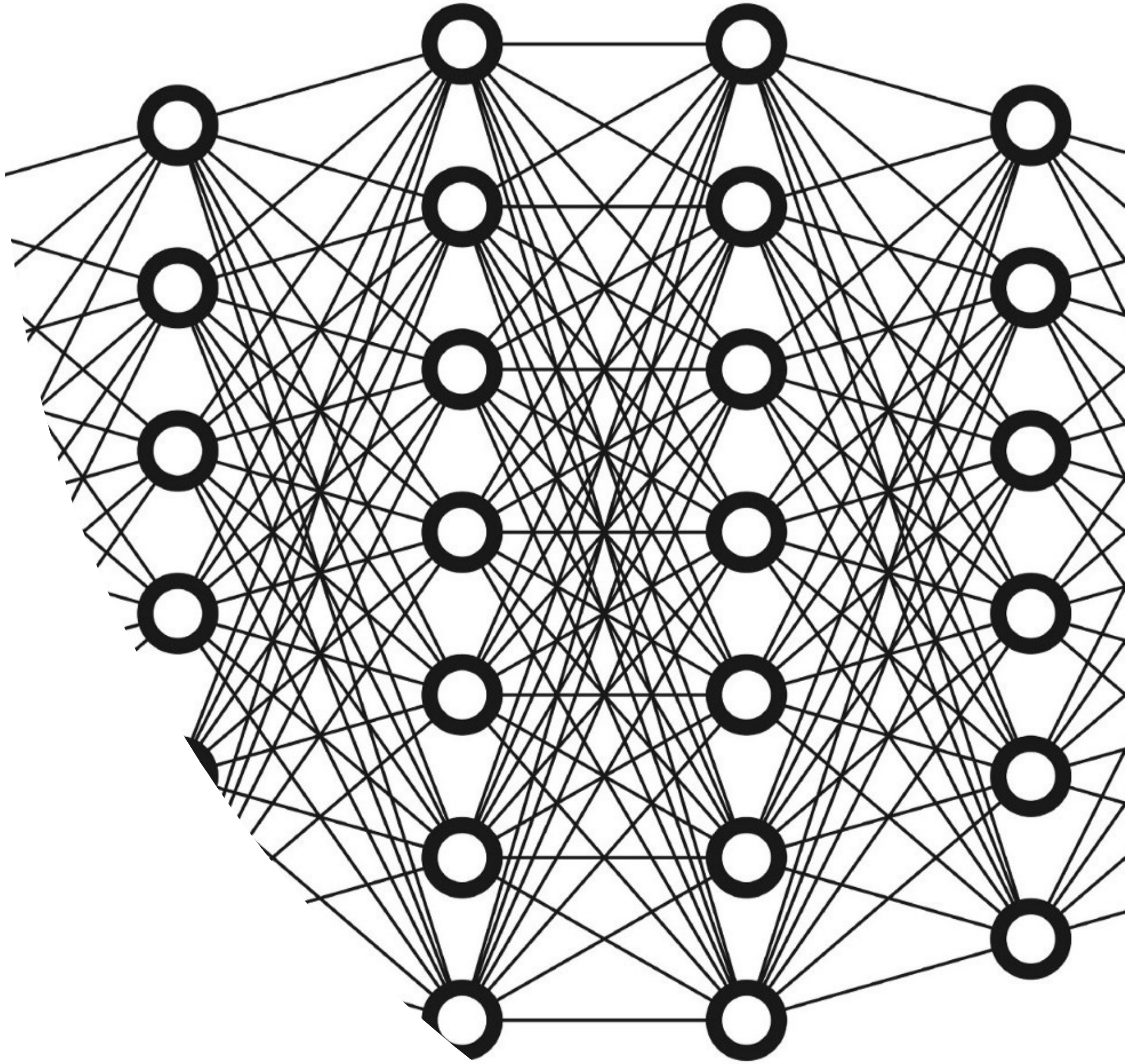
Why Natural Language Processing?

- Manual identification is time consuming and resource intensive
  - **6 – 16 minutes**
- Human bias and error
- Volumes of data, document types and linguistic variations
  - Finding the right document(s)
- Knowledge and tracking of military terms and phrases

NLP offers flexibility, scalability and repeatability

NLP and machine learning. Why?

- Examples are easier to create than rules
  - usually from an annotated gold standard corpus
- Rules may miss low frequency cases (edge cases)
  - Single term use
  - Obscure word usage
- Many factors involved in language interpretation
  - Able to model the linguistic relationship between terms and phrases
- Scalable and adaptable to different settings



The Military Service Identification Tool (MSIT)

Python

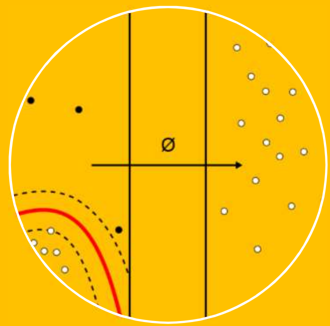
Natural Language Processing Toolkit

Scikit-learn

Development of the MSIT

4,200 patients extracted from the Personal History Extraction Dataset

Manually annotated each statement (*n*=6672)



Train a machine learning classifier

Rule-check to ensure prediction is a veteran

Prediction –

Civilian and Veteran

Total Class Numbers: Civilian 5630 Veteran 1042

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)
Royal Navy	124 (4.76)	Two years in the Army	64 (3.19)

Manual Annotation

Inter-rater agreement as indicted by a Cohen's kappa of 0.83 for veterans and 0.89 for civilians

Machine learning

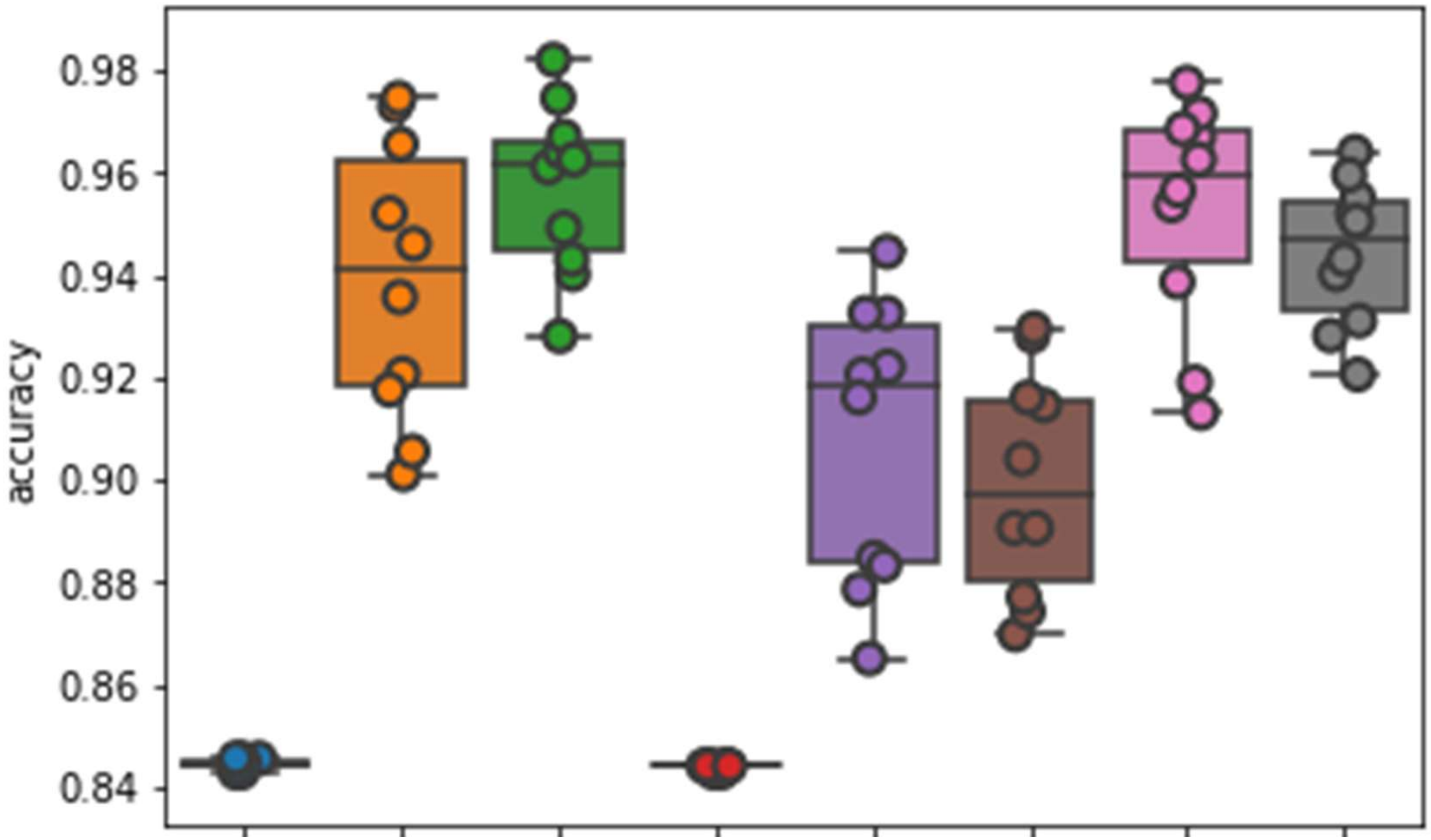
- Pre-processing
  - Common word removal EXCEPT military specific terms and phrases
  - Stemming (removal of affixes)
  - Noise removal
  - Feature representation
- Machine learning evaluation
  - Sub-set defined (training dataset, *n*=4470), with an equal proportion of civilian/veteran records
  - 8 machine learning algorithms evaluated against training dataset
  - Highest performance selected for further refinement

Machine learning – Selection Results

Training Dataset:
4470 Documents

Classifier
Random Forest
Decision Tree
Linear Support Vector Classifier
Support Vector Classifier
Multinomial Naïve Bayes
k-Nearest Neighbour
Logistic Regression
Multi-layered Perception

Testing Dataset:
2202 Documents



Post-processing – Veteran Prediction

- For those predicted as being a veteran, a check is performed to ensure the presence of a military specific word.

Army	Navy
RAF	Royal Air Force
National Service	Demobbed
Soldier	Conscripted
Corporal	Enlisted
Serviceman	Servicewoman
Falklands	Iraq
Afghanistan	Bosnia

Final Prediction

Performance

	SQL rule-based approach		MSIT	
	Veteran	Civilian	Veteran	Civilian
Veteran	262	58	290	30
Civilian	87	1795	27	1855
		Performance		
Precision	0.81		0.90	
Recall	0.75		0.91	
Accuracy	0.93		0.97	

Thank you

Any questions?

daniel.leightley@kcl.ac.uk

Research team:

PI: Dr Sharon Stevelink | PI: Professor Nicola Fear | PI: Dr Dominic Murphy | Researcher: Dr Daniel Leightley | Researcher: David Pernet

KCMHR

KING'S CENTRE FOR MILITARY HEALTH RESEARCH

ADMMH

ACADEMIC DEPARTMENT OF MILITARY MENTAL HEALTH

Currently submitted to: Journal of Medical Internet Research
Date Submitted: Aug 13, 2019
Open Peer Review Period: Aug 13, 2019 - Oct 3, 2019
(closed for review but you can still tweet)

Tweet

NOTE: This is an **unreviewed** Preprint

Preprint

Identifying Military Veterans in a Clinical Research Database using Natural Language Processing and Machine Learning

Daniel Leightley; David Pernet; Sumithra Velupillai; Katharine M. Mark; Elena Opie; Dominic Murphy; Nicola T. Fear; Sharon A.M. Stevelink;

ABSTRACT

Background:

Electronic healthcare records (EHRs) are a rich source of health-related information, with huge 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 notes 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 Trust. An iteratively developed Natural Language Processing and machine learning approach called the Veteran

Citation

Please cite as:
Leightley D, Pernet D, Velupillai S, Mark KM, Opie E, Murphy D, Fear NT, Stevelink SA
Identifying Military Veterans in a Clinical Research Database using Natural Language Processing and Machine Learning
JMIR Preprints. 13/08/2019:15852
DOI: 10.2196/preprints.15852
URL: https://preprints.jmir.org/preprint/15852

Current Preprint Settings

(as selected by the authors)

1. When the manuscript is submitted, allow peer review from:

✓ (a) Anybody (open community peer review)
(b) Editor-selected reviewers (closed peer review)

2. When the manuscript is submitted, display the preprint PDF to:

(a) Anybody, anytime
✓ (b) Logged-in users only
(c) Anybody, anytime (title and abstract only)
(d) Nobody

3. When the manuscript is accepted, display the accepted manuscript PDF to:

✓ (a) Anybody, anytime
(b) Logged-in users only
(c) Anybody, anytime (title and abstract