HHS Public Access

Author manuscript

J Affect Disord Rep. Author manuscript; available in PMC 2023 January 12.

Published in final edited form as:

J Affect Disord Rep. 2022 December ; 10: . doi:10.1016/j.jadr.2022.100430.

Portability of natural language processing methods to detect suicidality from clinical text in US and UK electronic health records

Marika Cusick[a,c,1,*], **Sumithra Velupillai**[b,c,1], **Johnny Downs**[b,c], **Thomas R. Campion Jr.**[a,c], **Evan T. Sholle**[a,c], **Rina Dutta**[b,c], **Jyotishman Pathak**[a,c]

[a] WeiCornell Medicine, 402 E. 67th St., New York, NY 10065, USA

[b] IoPPN, King's College London, London, UK

[c] South London and Maudsley NHS Foundation Trust, London, UK

Abstract

Background: In the global effort to prevent death by suicide, many academic medical institutions are implementing natural language processing (NLP) approaches to detect suicidality from unstructured clinical text in electronic health records (EHRs), with the hope of targeting timely, preventative interventions to individuals most at risk of suicide. Despite the international need, the development of these NLP approaches in EHRs has been largely local and not shared across healthcare systems.

Methods: In this study, we developed a process to share NLP approaches that were individually developed at King's College London (KCL), UK and Weill Cornell Medicine (WCM), US - two academic medical centers based in different countries with vastly different healthcare systems. We tested and compared the algorithms' performance on manually annotated clinical notes (KCL: n = 4,911 and WCM = 837).

Results: After a successful technical porting of the NLP approaches, our quantitative evaluation determined that independently developed NLP approaches can detect suicidality at another

* Corresponding author. mac2364@med.cornell.edu (M. Cusick).

[1] This author contributed equally to this work.

Author contributions

SV and MC contributed to the main analysis of this paper. They developed the NLP approaches, conducted the portability experiment, and evaluated the results of the experiment. They contributed heavily to the methods and discussion section. JD and RD provided clinical expertise on suicidality and contributed to the introduction and discussion section. THC and ES supervised the development of the NLP approach on the WCM side and provided expertise in clinical informatics. JP supervised the entire project and contributed revisions to all sections of the paper.

Declaration of Competing Interest

RD and SV declare previous research funding received from Janssen. The remainder of the authors (MC, ES, THC, JD, JP) do not declare any competing interests.

Ethics declarations

This study was approved by the WCM Institutional Review Board (IRB). The de-identified CRIS database has received ethical approval for secondary analysis: Oxford REC C, reference 18/SC/0372. The data is used in an anonymized and data-secure format under strict governance procedures.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jadr.2022.100430.

healthcare organization with a different EHR system, clinical documentation processes, and culture, yet do not achieve the same level of success as at the institution where the NLP algorithm was developed (KCL approach: F1-score 0.85 vs. 0.68, WCM approach: F1-score 0.87 vs. 0.72).

Limitations: Independent NLP algorithm development and patient cohort selection at the two institutions comprised direct comparability.

Conclusions: Shared use of these NLP approaches is a critical step forward towards improving data-driven algorithms for early suicide risk identification and timely prevention.

Keywords

Suicide; Natural language processing; Electronic health records; Portability

1. Introduction

Suicide is a major public health concern across the world. The World Health Organization (WHO) reports that nearly 700,000 people die per year by suicide, accounting for nearly 1.3% of all deaths worldwide (Ritchie et al., 2015; World Health Organization, 2021). To prevent suicide, many healthcare institutions have attempted to predict deaths by suicide, but this has not been particularly successful (Kessler et al., 2020). This can be attributed to the fact that within most populations, suicidal phenomena, especially those resulting in death, are relatively rare events, making it difficult to identify at-risk individuals (Pokorny, 1983). To improve suicide risk detection, researchers need to collaborate with other institutions to aggregate data or, when governance limits the pooling of protected health information (PHI), perform meta-analytic evaluations, improving statistical power.

Electronic health records (EHRs) are distributed documentation systems that offer a substantial advantage compared to paper records or charts by aggregating and collecting a wide variety of patient health information (Coorevits et al., 2013). Making use of EHR data, particularly unstructured clinical notes, offers a novel avenue for suicide risk modeling. EHRs can bring together very large samples for researchers to scrutinize and provide real-world insights into a patient's mental state (McCoy et al., 2016; Walsh et al., 2017). This is particularly true for suicidality, a common precursor to death by suicide (Brown et al., 2000), as it has been established that many providers document suicidality in notes rather than as structured data elements in EHR systems (Anderson et al., 2015; Haerian et al., 2012). Based on this evidence, several investigators and health systems are now working to develop natural language processing (NLP) approaches to detect suicidality from unstructured clinical notes in the EHRs (Fernandes et al., 2018). Currently, such NLP algorithms are considered useful for retrospective suicide-related research and can aid domain experts on patterns of suicide precursors and identification of suicide prevention interventions (Levis et al., 2022).

As health systems begin to achieve siloed success in detecting clinical conditions from EHR data and clinical notes, a critical next step forward in suicide risk detection is sharing and implementing these approaches broadly across other organizations, as most do not have the infrastructure and resources to develop and build these approaches *de novo*. Wide

adoption of existing approaches may minimize duplicative effort in the development of NLP algorithms (Chapman et al., 2011). While the literature details multiple efforts to enhance the portability of phenotype algorithms, such as through the eMERGE network (McCarty et al., 2011) or the OHDSI consortium (Hripcsak et al., 2016), and apply them to detecting physical conditions, such as rheumatoid arthritis (Carroll et al., 2012) and heart disease (Kashyap et al., 2020), to our knowledge, there has been little work in the sharing of NLP algorithms to detect more complex clinical phenomena specific to psychiatry (Edgcomb and Zima, 2019), such as suicidality and traumatic life events such as violence or abuse. The process is further complicated when sharing across institutions located internationally where the practice of clinical psychiatry, coding and documentation varies greatly.

In this study, we evaluate cross-institutional NLP portability of such complex clinical phenomena, with suicidality detection across the US and UK as our use case. As described by Silverman and De Leo (2016), the lack of an internationally agreed-upon set of terms, definitions, and classifications that indicate suicidality make it difficult to conduct and compare suicide-related research and further, make generalizable conclusions on findings. Psychiatrists in the US and UK document suicide-related issues according to the phenomenology they were taught in medical school and during clinical training, they and are inevitably “likely to emulate their supervisors’ EHR use” (Gagliardi and Turner, 2016). Each clinician seeks to follow best practice national guidelines, such as the American Psychiatric Association (APA) guidelines in USA (American Psychiatric Association, 2006) and the Department of Health Best Practice in Managing Risk guidance in the UK (National Risk Management Programme, 2007). Currently, there is no prior work that can aid international collaborations for sharing and evaluating NLP algorithms to detect suicidality from clinical notes in EHR systems, nor are there empirical findings on performance when NLP algorithms developed in one institution are implemented in another organization. Use of another institution’s algorithm may give unique insights on suicide precursors not previously studied and recognized, aiding the development of suicide prevention measures implemented at the institution.

To address this, we set out to evaluate how independently developed NLP approaches that detect suicidality translate across differing EHR platforms and classification objectives. In this study, we conducted a portability experiment using NLP approaches and datasets developed independently at two separate academic medical centers in two different countries (UK and the USA) known to have very different rates of national suicide rates (UK:US odds ratios for suicide 1:1.79) reflecting societal cultural differences (gun-related suicide deaths in the US) (Pritchard et al., 2021) and healthcare systems. Results from our experiment can inform other institutions on how to share NLP algorithms that detect clinically complex psychiatric phenomena, such as suicidality, a phenotype with important implications in improving international collaboration for suicide prevention efforts.

2. Materials and methods

2.1. Data source

We used NLP algorithms and EHR data from two large, academic healthcare institutions: South London and Maudsley Foundation National Health Service (NHS) Trust, based in South London, UK and Weill Cornell Medicine based in New York City, USA.

The KCL team used data extracted from electronic clinical records from the South London and Maudsley (SLaM) Foundation. NHS Trust is one of the UK's largest and oldest mental health trusts, providing a wide range of inpatient, outpatient and community-based mental health services. The main catchment area is four boroughs in south London (Croydon, Lambeth, Lewisham and Southwark), serving a local population of around 1.3 million people with more than 4500 employees. Jointly with the Institute of Psychiatry, Psychology & Neuroscience, KCL, SLaM hosts the National Institute for Healthcare Research (NIHR) Maudsley Biomedical Research center (BRC). The Clinical Record Interactive Search (CRIS) database is a resource developed at the Maudsley BRC, making de-identified EHR records available for secondary research use under an extensive governance model (Perera et al., 2016). The de-identified CRIS database has received ethical approval for secondary analysis: Oxford REC C, reference 18/SC/0372. The data is used in an anonymized and data-secure format under strict governance procedures. All experiments were performed in accordance with guidelines and regulations. The data were used in an entirely anonymized and data-secure format, and patients have the choice to opt-out of their anonymized data being used, and therefore, under UK law, does not require informed consent from patients whose data are represented here.

Weill Cornell Medicine (WCM) is an academic medical center in New York City with 1600 physicians, over 50 locations throughout the New York City metropolitan region, and 3 million annual patient encounters. WCM has an affiliation with NewYork-Presbyterian Hospital (NYPH), which serves as the primary emergency and inpatient setting for WCM patients. While clinical care is documented in different EHR systems at WCM and NYPH, the Architecture for Research Computing in Health (ARCH) database facilitates the secondary use of EHR data for research by capturing novel research measures and integrating data from multiple EHR systems (Sholle et al., 2017). This study was approved by the WCM Institutional Review Board (IRB). All experiments were performed in accordance with guidelines and regulations. This study was approved for a full waiver of informed consent, as it involves no more than minimal risk to the subjects.

2.2. Study population

The KCL test data consists of 4911 documents (progress notes, assessments and correspondence notes) from a random sub-cohort of 500 adolescents (13–18 years old) diagnosed with autism spectrum disorders (International Classification of Diseases (ICD-10): F84.0, F84.1, F84.5, F84.9) derived from a previously studied clinical sample (Downs et al., 2018; World Health Organization, 2004). Cohort demographic characteristics are provided in the supplementary section. The clinical documents were annotated for mentions of suicide-related information by trainee clinical psychologists, under senior

clinician supervision. As described in Downs et al. (2018), suicidality was defined as "as either the reporting of the intention to engage in a potentially lethal act towards oneself, or undertaking such acts themselves." Each note contained at least one instance of a suicide-related term (e.g. 'suicid*', 'kill him/her/themself', 'want to die'), that were then labeled as positive, negated, or uncertain. From the individual annotations, each *document* has then been further labeled as either affirmed/relevant for suicidality (True) or negated (False). There are in total 3069 documents labeled as True (62.5%) and 1842 as False (37.5%).

The WCM test data set consists of 837 suicide-related notes for 30 patients selected from a pre-established depression cohort, defined as any patient diagnosed with depression or prescribed an antidepressant. Of the 30 patients, 10 patients had an encounter diagnosis of suicidal ideation (V62.84 (ICD9), R45.85 (ICD10)) in their medical history. The remaining 20 patients were considered to be potentially suicidal, as they had at least ten notes with a key suicidal phrase ("suicidal", "suicide", "suicidal ideation (SI)", or "suicidality"). Cohort demographic statistics are provided in the supplementary section. A large majority of these notes (83%) were documented in the outpatient office setting at psychiatry and internal medicine departments between January of 2006 and December of 2019, further described in Cusick et al. (2021). The dataset was annotated for current suicidality, defined as patients discussing, thinking about or planning for suicide during the documented encounter, by two investigators at WCM with established annotation guidelines. Each note contained at least one instance of a suicidal mention ("suicidal", "suicide", "SI", or "suicidality"), that were then labeled as positive or negative for current suicidality. Based on these annotations, 134 (16.0%) of the documents were classified as either affirmed/relevant for suicidality (True) and 703 (84.0%) documents were classified as negated (False).

2.3. NLP approaches

Two symbolic rule-based NLP approaches were applied from each of the institutions. We henceforth refer to the KCL approach as KCL-neg (Velupillai et al., 2019) and the WCM approach as WCM-si (Cusick et al., 2021). They were both developed on the basis of the NegEx algorithm (Chapman et al., 2001), an approach to identify negated findings in unstructured clinical text. This algorithm relies on two lexicons: one defining target concepts (e.g. *suicidal)* and the other defining modifiers (e.g. *not).*

The KCL-neg approach was designed to detect any mention of suicidality, regardless of temporality (i.e. current or historical). The thirteen target lexicons of the KCL-neg approach included both direct and indirect mentions of suicidality. Direct mentions are any word with the regular expression basis of "suicid", which includes "suicidal," "suicidality," and "suicide." Indirect mentions include expressions such as "take (his|her|their) life", "wish to die", and "life not worth living." The WCM-si approach was designed for detecting *current* suicidality, a predictor for lethal suicide attempts (Shelef et al., 2021). The target lexicons for the WCM-si approach were the four key suicidal ideation terms—"suicidal", "suicide", "suicidality", "si"—used to select the EHR note cohort. The two approaches had overlap in target lexicons that were direct mentions of suicidality.

The KCL team implemented different sets of modifiers to study the impact on algorithm performance when using previously published lists of modifier terms compared to adapted

lists for new use cases. We used two of the KCL modifier sets, anySI-1 (44 modifiers) and anySI-2 (248 modifiers). The WCM modifier lexicon set, henceforth called currentSI, included 108 modifiers to negate current suicidality. The WCM modifier set shared 10 and 25 modifiers with anySI-1 and anySI-2, respectively. We categorized both the KCL and WCM modifiers into four different categories: negated, historical, conditional, and unrelated. Examples of each of these modifiers are provided in Table 1. The entire set of modifiers are available on our respective GitHub[2],[3] websites, with additional details on the NLP approaches in the supplementary section.

2.4. NLP algorithm portability process

To initiate the portability experiment, investigators from WCM and KCL held several meetings to present and discuss each of the NLP algorithms from both a clinical and technical standpoint. First, it was critical to share the algorithm's main clinical objectives, key details about the study population and site, and guidelines behind the manual annotation of the test set. Next, we shared the technical details of the algorithms and developed a plan of action to transfer compatible code and execution guidelines. Upon realizing that versioning issues might arise, we created a virtual environment with a consistent Python installation and package versions. Finally, we confirmed the format of each institution's data set, including the text format. To maintain security and privacy of each institution's dataset, we ran both algorithms within our respective firewalls, with no data sharing. Code for the algorithms is available on our respective GitHub1,[2] websites.

2.5. Evaluation

After executing the two NLP algorithms on the manually annotated datasets, we evaluated the results using both quantitative and qualitative methods. First, to assess portability quantitatively, we compared the algorithms' results using traditional intrinsic evaluation metrics, such as accuracy, precision, recall, and F1-scores (Resnik and Lin, 2010). Second, to assess portability with an eye towards the underlying details of each NLP approach, we conducted a thorough qualitative manual error analysis to characterize the most common misclassification scenarios. Based on the specification of each of the approaches, we can identify some classification errors to be expected and re-analyze relevant quantitative metrics after removing notes with such errors. No changes were made to either of the algorithms during the portability experiment, leaving improvements for generalizability as future work.

2.6. Results

During a five-month span, the two teams met at least ten times (twice a month) to outline the technical requirements necessary to port our algorithms to unseen datasets at the other institution. Once all necessary information and details were made available on GitHub[1], each team successfully executed the ported algorithm on their own test data set with little difficulty. In the event of questions, we communicated via email, striving to respond within two days to all critical communications. Fig. 1 illustrates the complete portability

[2] https://github.com/wcmc-research-informatics/SI_Ideation
[3] https://github.com/KCL-Health-NLP/camhs_pycontext_adaptation

experiment workflow. Pre-experiment, the WCM and KCL teams independently developed and validated their respective algorithms. Within the portability experiment, the teams engaged in preliminary discussions and planning, executed code sharing through GitHub, and conducted quantitative evaluation followed by qualitative error analysis. Finally, future work will involve algorithm improvements on generalizability and further cross-institutional collaborations.

2.7. NLP approach results

As demonstrated in Table 2, the ported algorithms did detect suicidality, yet they did not replicate the same level of success in detecting suicidality as at the institution where the algorithm was originally developed. Using the two modifier sets, the KCL-neg approach achieved a maximum macro-average f1-score of 0.85 on its own KCL test dataset, but only resulted in a maximum score of 0.68 on the WCM dataset. Similarly, the WCM-si approach resulted in a macro-average f1-score of 0.87 on their own test dataset, but only achieved a maximum score of 0.72 on the KCL dataset.

We observed the same phenomenon on each of the performance metrics (e.g. precision on affirmed instances), as neither of the approaches outperformed the success of the "home algorithms". While the WCM-si approach was able to achieve higher precision (0.87) than the KCL-neg approach using AnySI-1 modifiers (0.74) for positive instances of suicidality on the KCL data, the KCL-neg approach using AnySI-2 modifiers yielded a similar precision (0.87).

Although the KCL-neg approach using the AnySI-2 modifier set had a better overall performance (macro-average f1-score: 0.85 vs. 0.68) in comparison to the AnySI-1 modifier set on their own KCL data set, the opposite was observed on the WCM data set (macro-average f1-score: 0.53 with AnySI-1 vs. 0.68 with AnySI-2).

2.8. Error analysis

2.8.1. KCL-neg with AnySI-1 modifiers on WCM data—Of the two KCL-neg modifier sets, the AnySI-I modifier lexicon set proved the most successful (macro-average f1-score of 0.68 vs. 0.53). Thus, using this set, we conducted our qualitative error analysis to determine the reasons for misclassification.

Out of the 206 total errors, there were 177 and 29 false positive and false negative errors, respectively. Of the 29 false negative errors, the majority (69%) can be attributed to the KCL-neg algorithm's target lexicons not including the term "si." For this reason, the KCL algorithm was not programmed to detect this type of mention from the clinical notes and automatically classified the note as negative. After removing these expected instances from the note set, recall for positive mentions of suicidal ideation increased from 0.80 to 0.93.

The 177 false positive errors (Table 3) can be grouped into the following scenarios: missing a negation modifier, non-patient person reference, structured references, conditional mentions, and historical mentions. The table below displays the number of cases in each scenario and several examples. Because the KCL algorithm was not configured to negate historical suicidality, the 45 false positive errors classified to this scenario were to be

expected. By removing these expected errors from the note set, precision for positive mentions increased from 0.39 to 0.46.

2.8.2. WCM-si with currentSI modifiers on KCL data—A similar error analysis was performed on the KCL data with the WCM-si approach. Out of the 1398 total errors, there were 279 and 1119 false positive and false negative errors, respectively. Of the 1119 false negative errors, 529 (47%) were attributed to the WCM-si algorithm not including target terms such as “kill him/herself” or “end his/her life”. An analysis of 250 of the remaining 590 errors revealed that 58 (23%) related to references to the past, which the WCM-si approach was designed to exclude given that its primary focus is on “current” suicidality; 69 (28%) related to complex, long documents where there were several references to suicidal behavior but the algorithm only picked up those related to “suicid*”; 40 (16%) related to missing triggers or erroneous trigger scopes; and 83 (33%) related to other issues, including errors in the gold standard annotations. As the WCM-si approach was not configured to detect indirect mentions and purposefully negated historical mentions, we considered 779 of the false negative errors to be expected, changing the recall for positive mentions of suicidal ideation from 0.64 to 0.83.

For the 279 false positive errors, similar scenarios as for the KCL-neg approach on the WCM dataset were observed (Table 4), with some notable differences: historical mentions were not considered false positives in this case; but missing negation modifiers were observed (e.g. “nil”), as well as structured mentions in forms. Additionally, some false positives were related to documents where there were both negative and positive mentions. These were classified as negated in the KCL gold standard, but the WCM-si approach classified them as positive. Other examples include mentions of routine checks by the clinician, hypotheticals, and mentions related to someone other than the patient. We did not consider any of these errors to be expected.

3. Discussion

Our study showed that NLP approaches developed to detect complex clinical constructs, such as suicidality, can be successfully ported and shared across institutions, with proper emphasis on clear and effective communication. In this experiment, we saw success in technical seamlessness, the algorithms’ ability to gain signal on an unseen new dataset, and the valuable insight for more nuanced NLP evaluation techniques and opportunities for future work. First, our informative discussions on the technical compatibility of our NLP algorithms (now hosted on GitHub[1]) made porting the algorithms a seamless experience. Second, while the traditional quantitative measures of portability, such as accuracy and F1-scores, indicated that ported NLP approaches were not able to achieve the same level of success as at their home institution, much of this was attributed to the approaches’ slight differences in clinical objectives. Our qualitative error analysis, which took this into account, indicated that many of the errors were to be expected based on the institution’s guidelines for defining and annotating suicidality. In fact, if expected errors were taken out of consideration, we found the ported algorithm’s results to improve significantly. For the KCL-neg on the WCM data, 65 (32%) of the 206 errors were to be expected, changing the overall F1 score of the algorithm from 0.75 to 0.83. Similarly, of the 1398 errors that the

WCM-si approach made on the KCL data, 587 (42%) of the errors were to be expected based on the algorithm's configuration and objective, changing the overall F1 score from 0.72 to 0.83. Adjustment for expected errors led to reductions in the instances of false negative errors, one of the most important metrics for NLP algorithms in being useful for detecting suicidality and common precursors to suicide. Finally, through the collaborative process, we were able to gain a better understanding of the other institution's clinical objectives and patient cohorts of interest, develop a framework for a more informative qualitative evaluation, and identify potential areas of improvement for future work.

In addition to a more meaningful evaluation, understanding the underlying details of the NLP approaches may inform how to develop a more generalizable approach. Similar to prior work, we confirmed in this study that suicidality is interpreted and defined differently across institutions and healthcare systems. Among our two institutions, the biggest differences in the definition of suicidality include decisions on temporality, specifically WCM-si's focus on current suicidality, and inclusion of indirect suicidality-related terms, specifically KCL-neg's use of phrases such as "wish to die", and "life not worth living." While these differences exist, our approaches did also have a number of similarities, including terms in the target lexicons and modifier terms within negation, conditional, and non-experiencer (unrelated categories). This commonality suggests that with further experimentation and collaborations, we can continue to improve on the detection of this complex clinical condition, by developing a portable and generalizable approach.

The NLP methods used in this study used relatively simple rule-based approaches to tackle the clinically complex phenomena of suicidality. Both of our institutions have also implemented more novel, state-of-the-art methods for the detection of suicidality, such as a text classification convolutional neural network (CNN) (Cusick et al., 2021; Song et al., 2020), and support vector machines (SVM) (Velupillai et al., 2019). However, we decided to experiment with the more basic rule-based NLP algorithms for two reasons: ease of portability and human interpretability. With the eventual goal of building generalizable and portable NLP approaches to detect suicidality, we determined that rule-based approaches could be more widely implemented across institutions as they require significantly less technical expertise, computational power, and other resources. In addition, many studies have concluded that simple rule-based approaches achieve similar levels of success to these novel implementations (Chiticariu et al., 2013; Cusick et al., 2021; Tan et al., 2018; Topaz et al., 2019). A second advantage of rule-based NLP algorithms is their human interpretability. While effective, state of the art machine learning methods are challenging to interpret (Martens et al., 2011; Vellido et al., 2012) and even more difficult to adjust. In the case of rule-based NLP algorithms, an external institution would have the ability to determine the most common sources of error and make changes to the approach, as it sees fit to the organization's use case.

3.1. Limitations

There are several limitations to this study. First, because the suicidality NLP approaches and test data sets were developed completely independently of each other, we recognize that they may not be directly comparable. However, these differences show the robustness of our

approaches – as they were able to detect suicidality in disparate populations. We view this to be a real-world experiment and thus replicable by other institutions. Second, our test data sets were relatively small, and similar results may not hold in larger patient populations. Third, our methods of data extraction, specifically data filtering, create patient cohorts with known suicidal issues, which may be problematic when applying NLP at scale. While the WCM team only extracted notes with explicit mentions of suicidality, such as “suicidal”, the KCL team extracted notes with wider search criterion, which included both explicit and “implicit” terms, such as “wish to die.” Fourth, because we were not able to interact with the clinicians who wrote the clinical notes, we were unable to comment and interpret how specific institutional differences in front-end EHR systems, clinical documentation process, and culture impacted the results of our portability experiment. Fifth, incorporation of diagnostic codes, such as ICD-9/10 may improve our algorithms’ results. However, the aim of this study was to evaluate and assess portability of our suicidality NLP approaches, rather than prediction results. Results from this study will make NLP algorithms more widely accessible and bolster results of existing suicide prediction algorithms currently reliant on structured EHR data such as diagnosis codes. Finally, because we did not port and evaluate our more advanced machine-learning based algorithms, we cannot say for certain whether these models would have achieved higher success than our lexicon-based NLP approaches. However, based on our goal of developing NLP approaches that can be widely used, we believe this is out of scope for this study and a future area of research.

4. Conclusion

In an effort to understand patients who are at risk for death by suicide, routinely collected data from healthcare institutions, such as EHRs, can be a valuable resource at scale. Information about suicidal risk behavior is predominantly documented in free text notes in EHRs, leading to an increase in the development of NLP approaches to detect suicidality in unstructured clinical text. However, due to the clinical complexity of suicidality, the lack of consensus on how to define this condition, differences in how clinical assessments are documented in EHRs, and the various ways the task can be modeled for information extraction, the development of relevant NLP approaches have largely been local to each institution’s definitions and interpretations. Thus, these NLP approaches are much less generalizable and portable in comparison to phenotype algorithms for well-defined clinical conditions, such as rheumatoid arthritis (Carroll et al., 2012). However, with a well-defined process to understand the underlying details of an approach, institutions can be well-equipped to make use of an external approach, allowing a larger number of institutions to participate in suicide-related research. This is a critical step forward in learning how to develop more robust, portable, and generalizable NLP methods that can be applied to any clinical text, regardless of the origin EHR system.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

This research has been funded in part by the US National Institutes of Health (NIH) Grants R01MH119177, R01MH121907, R01GM105688, R01MH121922, and P50MH113838.

Role of the funding source

Research by MC, ES, TC, and JP was funded in part by NIH grants R01MH105384, R01MH119177 and R01MH121922. JD is supported by NIHR Clinician Science Fellowship award (CS-2018-18-ST2-014) and has received support from a Medical Research Council (MRC) Clinical Research Training Fellowship (MR/L017105/1) and Psychiatry Research Trust Peggy Pollak Research Fellowship in Developmental Psychiatry. This paper by SV, JD, and RD represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research center at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. SV and JD are part-funded by the Medical Research Council (MRC) Mental Health Data Pathfinder Award to King's College London. RD is funded by a Clinician Scientist Fellowship (RE11682: research project e-HOST-IT) from the Health Foundation in partnership with the Academy of Medical Sciences.

Data availability

The datasets generated during and/or analyzed during the current study are not publicly available because they include personally identifiable data, which each investigator on this study obtained access through our institutions' review boards. CRIS data is made available to researchers with appropriate credentials (provided by the South London and Maudsley NHS Trust) working on approved projects. Projects are approved by a CRIS Oversight Committee, a body set up by and reporting to the South London and Maudsley Caldicott Guardian. On request, and after appropriate credentials have been obtained as well as arrangements with the lead of the respective CRIS project, data presented in this study can be viewed within the secure system firewall. In a similar fashion, WCM data is made available to investigators who are a part of WCM IRB approved research projects. Researchers are given credentials and access instructions to a subset of EHR data as described in the project's study protocol.

As mentioned in our manuscript, code for our NLP algorithms is available on our GitHub repositories at the following links: https://github.com/KCL-Health-NLP/camhs_pycontext_adaptation, https://github.com/wcmc-research-informatics/SI_Ideation.

References

American Psychiatric Association, 2006. Practice Guideline for the Assessment and Treatment of Patients With Suicidal Behaviors. APA Practice Guidelines for the Treatment of Psychiatric Disorders: Comprehensive Guidelines and Guideline Watches. American Psychiatric Association, Arlington, VA.

Anderson HD, Pace WD, Brandt E, Nielsen RD, Allen RR, Libby AM, West DR, Valuck RJ, 2015. Monitoring suicidal patients in primary care using electronic health records. J. Am. Board Fam. J. Am. Board Fam. Med. JABFM 28, 65–71.

Brown GK, Beck AT, Steer RA, Grisham JR, 2000. Risk factors for suicide in psychiatric outpatients: a 20-year prospective study. J. Consult. Clin. Psychol. 68, 371–377. [PubMed: 10883553]

Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, Pacheco JA, Boomershine CS, Lasko TA, Xu H, Karlson EW, Perez RG, Gainer VS, Murphy SN, Ruderman EM, Pope RM, Plenge RM, Kho AN, Liao KP, Denny JC, 2012. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J. Am. Med. Inform. Assoc. JAMIA 19, e162–e169. [PubMed: 22374935]

Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG, 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. J. Biomed. Inform. 34, 301–310. [PubMed: 12123149]

Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O, 2011. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J. Am. Med. Inform. Assoc. JAMIA 18, 540–543. [PubMed: 21846785]

Chiticariu L, Li Y, Reiss FR, 2013. Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems!. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Presented at the EMNLP 2013, Association for Computational Linguistics, Seattle, Washington, USA, pp. 827–832.

Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, Dugas M, Dupont D, Schmidt A, Singleton P, De Moor G, Kalra D, 2013. Electronic health records: new opportunities for clinical research. J. Intern. Med. 274, 547–560. [PubMed: 23952476]

Cusick M, Adekkanattu P, Campion TR, Sholle ET, Myers A, Banerjee S, Alexopoulos G, Wang Y, Pathak J, 2021. Using weak supervision and deep learning to classify clinical notes for identification of current suicidal ideation. J. Psychiatr. Res. 136, 95–102. [PubMed: 33581461]

Downs J, Velupillai S, George G, Holden R, Kikoler M, Dean H, Fernandes A, Dutta R, 2018. Detection of suicidality in adolescents with autism spectrum disorders: developing a natural language processing approach for ese in electronic health records. AMIA Annu. Symp. Proc. 2017, 641–649. [PubMed: 29854129]

Edgcomb JB, Zima B, 2019. Machine learning, natural language processing, and the electronic health record: innovations in mental health services research. Psychiatr. Serv. Wash. DC 70, 346–349.

Fernandes AC, Dutta R, Velupillai S, Sanyal J, Stewart R, Chandran D, 2018. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. Sci. Rep. 8, 7426. [PubMed: 29743531]

Gagliardi JP, Turner DA, 2016. The electronic health record and education: rethinking optimization. J. Grad. Med. Educ. 8, 325–327. [PubMed: 27413432]

Haerian K, Salmasian H, Friedman C, 2012. Methods for identifying suicide or suicidal ideation in EHRs. AMIA Annu. Symp. Proc. 2012, 1244–1253. [PubMed: 23304402]

Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, Suchard MA, Schuemie MJ, DeFalco FJ, Perotte A, Banda JM, Reich CG, Schilling LM, Matheny ME, Meeker D, Pratt N, Madigan D, 2016. Development and validation of phenotype classifiers across multiple sites in the observational health data sciences and informatics network. Proc. Natl. Acad. Sci. 113, 7329–7336. [PubMed: 27274072]

Kashyap M, Seneviratne M, Banda JM, Falconer T, Ryu B, Yoo S, Hripcsak G, Shah NH, 2020. Development and validation of phenotype classifiers across multiple sites in the observational health data sciences and informatics network. J. Am. Med. Inform. Assoc. JAMIA 27, 877–883. [PubMed: 32374408]

Kessler RC, Bossarte RM, Luedtke A, Zaslavsky AM, Zubizarreta JR, 2020. Suicide prediction models: a critical review of recent research with recommendations for the way forward. Mol. Psychiatry 25, 168–179. [PubMed: 31570777]

Levis M, Levy J, Dufort V, Gobbel GT, Watts BV, Shiner B, 2022. Leveraging unstructured electronic medical record notes to derive population-specific suicide risk models. Psychiatry Res 315, 114703. [PubMed: 35841702]

Martens D, Vanthienen J, Verbeke W, Baesens B, 2011. Recent advances in data, text, and media mining & information issues in supply chain and in service system design. Decis. Support Syst. 51, 782–793.

McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM, Struewing JP, Wolf WA, Team, em, 2011. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med. Genom. 4, 13.

McCoy TH, Castro VM, Roberson AM, Snapper LA, Perlis RH, 2016. Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. JAMA Psychiatry 73, 1064–1071. [PubMed: 27626235]

National Risk Management Programme, 2007. Best practice in managing risk - principles and evidence for best practice in the assessment and management of risk to self and others in mental health services. Department of Health.

Perera G, Broadbent M, Callard F, Chang CK, Downs J, Dutta R, Fernandes A, Hayes RD, Henderson M, Jackson R, Jewell A, Kadra G, Little R, Pritchard M, Shetty H, Tulloch A, Stewart R, 2016. 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 6, e008721.

Pokorny AD, 1983. Prediction of suicide in psychiatric patients. Report of a prospective study. Arch. Gen. Psychiatry 40, 249–257. [PubMed: 6830404]

Pritchard C, Porters S, Rosenorn-Lanng E, Williams R, 2021. Mortality in the USA, the UK and other Western countries, 1989–2015: what is wrong with the US? Int. J. Health Serv. 51, 59–66. [PubMed: 33059529]

Resnik P, Lin J, 2010. Evaluation of NLP Systems. The Handbook of Computational Linguistics and Natural Language Processing. John Wiley & Sons, Ltd, pp. 271–295.

Ritchie H, Roser M, Ortiz-Ospina E, 2015. Suicide. Our World Data.

Shelef L, Rabbany JM, Gutierrez PM, Kedem R, Ben Yehuda A, Mann JJ, Yacobi A, 2021. The role of past suicidal behavior on current suicidality: a retrospective study in the Israeli military. Int. J. Environ. Res. Public Health 18, E649.

Sholle ET, Kabariti J, Johnson SB, Leonard JP, Pathak J, Varughese VI, Cole CL, Campion TR, 2017. Secondary use of patients' electronic records (SUPER): an approach for meeting specific data needs of clinical and translational researchers. AMIA Annu. Symp. Proc. AMIA Symp. 2017, 1581–1588.

Silverman MM, De Leo D, 2016. Why there Is a need for an international nomenclature and classification system for suicide. Crisis 37, 83–87. [PubMed: 27232426]

Song X, Downs J, Velupillai S, Holden R, Kikoler M, Bontcheva K, Dutta R, Roberts A, 2020. Using deep neural networks with intra- and inter-sentence context to classify suicidal behaviour. In: Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France. European Language Resources Association, pp. 1303–1310. Presented at the LREC 2020.

Tan WK, Hassanpour S, Heagerty PJ, Rundell SD, Suri P, Huhdanpaa HT, James K, Carrell DS, Langlotz CP, Organ NL, Meier EN, Sherman KJ, Kallmes DF, Luetmer PH, Griffith B, Nerenz DR, Jarvik JG, 2018. Comparison of natural language processing rules-based and machine-learning systems to identify lumbar spine imaging findings related to low back pain. Acad. Radiol. 25, 1422–1432. [PubMed: 29605561]

Topaz M, Murga L, Gaddis KM, McDonald MV, Bar-Bachar O, Goldberg Y, Bowles KH, 2019. Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches. J. Biomed. Inform. 90, 103103. [PubMed: 30639392]

Vellido A, Martín-Guerrero JD, Lisboa PJG, 2012. Making machine learning models interpretable. In: Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning.

Velupillai S, Epstein S, Bittar A, Stephenson T, Dutta R, Downs J, 2019. Identifying suicidal adolescents from mental health records using natural language processing. Stud. Health Technol. Inform. 264, 413–417. [PubMed: 31437956]

Walsh CG, Ribeiro JD, Franklin JC, 2017. Predicting Risk of Suicide Attempts Over Time Through Machine Learning. Clin. Psychol. Sci. 5, 457–469.

World Health Organization, 2004. ICD-10 : International Statistical Classification of Diseases and Related Health Problems : Tenth Revision. World Health Organization.

World Health Organization, 2021. Suicide [Fact sheet]. Retrieved from https://www.who.int/news-room/fact-sheets/detail/suicide.

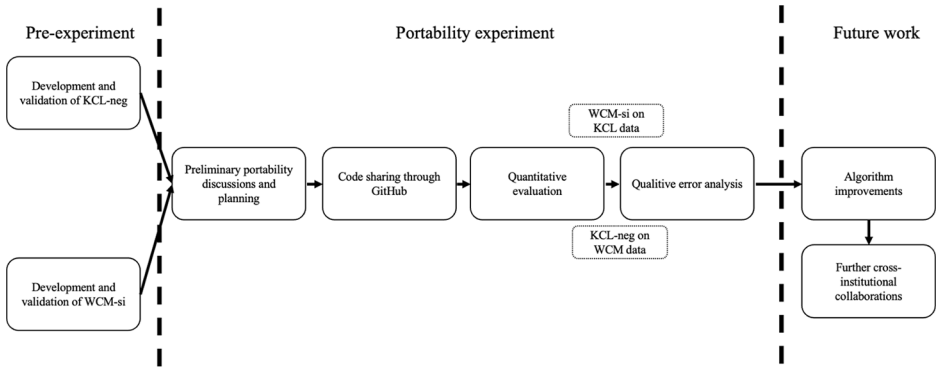

Fig. 1.
Diagram of portability experiment workflow.

Table 1

Modifier categories.

Modifier category	AnySI-1	AnySI-2	CurrentSI	Examples
Negated	Yes	Yes	Yes	Denied, negative for, never, not, no
Historical	No	No	Yes	History, previous, attempted, YEAR, DATE
Conditional	No	Yes	Yes	Lifeline, emergency number, even if, return if
Unrelated	No	Yes	Yes	Brother, sister, family history, husband, wife

Table 2

Results: Precision (P), Recall (R) and F1-score (F1) for affirmed (True) and negated (False) suicide-related instances, applied on two datasets (WCM and KCL) and using two NLP approaches (KCL-neg and WCM-si) with different modifier lexicons.

Data	NLP Approach	Modifier Set	P (T)	R (T)	F1 (T)	P (F)	R (F)	F1 (F)	F1 (avg)
WCM	KCL-neg	AnySI-1	0.39	0.80	0.53	0.95	0.74	0.83	0.68
WCM	KCL-neg	AnySI-2	0.26	0.77	0.38	0.92	0.54	0.68	0.53
WCM	WCM-si	CurrentSI	0.72	0.87	0.79	0.97	0.93	0.95	0.87
KCL	KCL-neg	AnySI-1	0.74	0.86	0.79	0.67	0.5	0.57	0.68
KCL	KCL-neg	AnySI-2	0.87	0.91	0.89	0.84	0.78	0.81	0.85
KCL	WCM-si	CurrentSI	0.87	0.64	0.74	0.58	0.85	0.69	0.72

*Results are reported without adjustment to expected errors.

Table 3

False positive error scenarios KCL-neg on WCM.

Scenario	Cases	Example
Missing a negation modifier	16	“Negative for suicidal ideation”
Non-patient references	23	“Brother (committed suicide)”
Structured mentions	63	“Suicidal ideation: denies”
Conditional mentions	30	“Children of untreated depressed mothers are also more prone to suicidal behavior”
Historical suicidality	45	“Has another suicide attempt on MM/DD”
Total	177	

Table 4

False positive error scenarios WCM-si on KCL data.

Scenario	Cases	Example
Negation and positive mention in same note (often references to the past)	149	“X said they had suicidal thoughts ... although self harm may not be true suicidal intent”
Missing a negation modifier	29	“Nil suicidal ideation”
Structured mentions	35	“Suicide risk: low”
Routine checks, hypothetical, unrelated	34	“Asked about suicidal thoughts”
Other	32	“Refused to answer questions around suicidal ideation”
Total	279	