**A. SIGNIFICANCE**

**A1. Depression is common in peripartum women and has negative consequences for maternal and child health**

Kenya mental health strategy 1 WHO 2017 depression report 2

**A2. Mobile health (mHealth) approaches may be resource-appropriate strategies to support mental health**

**A3. The potential for interactive SMS interventions to be efficient and adaptive has not been fully realized.**

Many SMS interventions involve automated transmission of pre-scheduled communication. However, interaction with participants is managed by live agents. Automation may increase efficiency. In order to implement automation, participant status must be inferred from digital communications, which are structured as natural language. Data are needed on what can be automated while preserving efficacy

**A4. Natural language processing (NLP) enables unstructured human communication to be analyzed computationally.** NLP refers to a set of methods that transform and represent natural language in a form that can be used for computation.3,4 It is the method that underlies now-ubiquitous tools such as internet searching and email spam filtering, and is receiving growing attention in biomedical fields as a way to harness rich data captured as large quantities of free text, which are prohibitively resource-intensive to manually summarize.5–7 Essentially, NLP breaks a piece of text (a *corpus*) into its constituent *tokens* (typically words) and represents each token numerically. Many approaches have been developed to accomplish numerical representation. They range from simply representing each word type encountered as an arbitrary integer, and summarizing the corpus by word frequencies, to representing each word as a multidimensional vector of *features*.8 Features can be assigned based on expert linguistic knowledge, for example a feature could be a categorical variable indicating which part of speech it is (e.g. verb, noun) or to which root word it is related.3,4 Recently, machine learning algorithms have been developed that perform *contextual word embedding* – non-parametric feature definition based on a word’s distribution in other corpora, based on the idea that words found in similar contexts (e.g. close to the same words) have similar meanings.9–11 These algorithms have been shown to have excellent performance as predictors in a variety of classification tasks,8 but the features they identify are generally not interpretable by humans.

NLP figure TBD

**A5. Machine learning methods can be used to predict clinical outcomes based on natural language communication.** Machine learning (ML) refers to computer algorithms that derive parameters of mathematical models by iteratively learning from observed data.12,13 ML can be supervised, meaning *labeled* data with linked predictor and outcome variable values are available to train the model; or unsupervised, meaning data are unlabeled and the algorithm clusters similar data together without considering outcome. ML methods can be used in analysis of unstructured natural language communications such as SMS messages in several ways. As described above, unsupervised ML models known as *neural networks* can be used to generate feature vectors from text.(ref) Additionally, feature vectors generated by (ML or non-ML) NLP can in turn be used as predictors in supervised ML tasks such as classifying whether text indicates a particular health outcome.

A number of studies have employed NLP and ML methods to extract mental health-related information from unstructured text. Physician notes from medical records have been most widely used as data sources, but social media posts have recently been used as abundant textual windows into patients’ subjective psychological experiences.7,14–16 Social media posts and SMS messages pose unique analytical challenges: they are short, contain abbreviations and misspellings, and often discuss a range of topics. Nonetheless, **several models have been developed that successfully predict depression,**17–20 **suicidality,**21,22 **and mental health crisis**23,24 **from social media and SMS messages.** Sensitivity and specificity of reported models were both 60-90%. Most models employed non-parametric ML approaches to define features, making it difficult to identify individual human-interpretable features associated with depression; those studies that used rule-based feature definition reported that users experiencing depression are more likely to use certain key words, first-person pronouns, and message less frequently.15,17

ML/NLP schematic TBD

Prior studies validate that NLP and ML approaches can be used to monitor mental health based on digital communications, but current literature has the following limitations. Few studies have applied the **most recent generation of contextual word embedding algorithms, which may yield higher sensitivity and specificity**.10,25 The vast majority of studies have focused on English language corpora; **validation of these methods in non-English languages is needed**.15 Finally, **few studies to date have expanded beyond *identification* of mental illness and applied these approaches to inform mHealth *treatment***.24

**B. INNOVATION**

**B1. Leveraging existing data to improve the efficiency of Mobile WACh**

Availability of labeled longitudinal data

Availability of both patient and provider communication

**B2. Non-English language**

**B3. Inclusion of latest embedding methods?**

**B4. Probing effective messages for mental health treatment**

**C. APPROACH**

**C1. Overall strategy**

Summary figure

Links to training plan

Relationship with mentor research

**C2. Preliminary studies**

**Mobile WACh platform overview**

* Design, history, efficacy
* Need for improved efficiency
* Number of participant, nurse and system SMS messages in each study

**Mental health data in Mobile WACh studies**

**Impact of Mobile WACh intervention on mental health**

* XY
* Qualitative accounts
* Plans in Neo

**Analyses of Mobile WACh SMS content**

**C4. Aim 1. To develop a computational model that uses *Mobile WACh* participants’ SMS messages to predict their depression symptoms.**

**Rationale:** Published studies and our preliminary data suggest that NLP and ML methods can be used to predict maternal mental health status from the SMS messages they send to the system. The goal of this aim is to train and test the performance of a predictive model to identify Mobile WACh participants with elevated depression symptoms. This will result in a computational model that can be tested for its ability to identify messages indicative of depression in practice (Aim 3).

**Figure 1. Aim 1 overview**

**Figure 1** provides an overview of the model development process for this aim. Details of each step are described below.

**Data source:** A corpus will be generated by pooling SMS messages sent by participants in the Mobile WAChX, Mobile WACh XY, Mobile WACh Neo pilot and Mobile WACh Neo RCT studies.26,27 **Table 1** summarizes data available from each study. We currently have a total of ~34,000 SMS received from 1205 participants. An additional 2500 participants will be enrolled in Mobile WACh Neo by the K18’s period of performance, increasing the expected number of participant messages in our corpus to 100,000 assuming similar levels of participant messaging.

**Data labels:** We will take a supervised modeling approach, meaning that all SMS messages will be labeled with known values of the outcome of interest (depression). All participants included in this analysis have multiple longitudinal depression score measures every few months, with either the Patient Health Questionnaire (PHQ9) or the Edinburgh Postpartum Depression Scale (EPDS) (**Table 1**). Each SMS in the corpus will be assigned the depression score closest in time to it. Depression scores will be dichotomized into depressed vs. not depressed at a cutoff score of 10, based on instrument guidelines.(refs)

**Feature definition:** As described in section A4, multiple approaches can be taken to define features of natural language. In order to identify the optimal predictive model, we will explore multiple methods to define features and compare the performance of different models, an approach often taken in NLP.3,4 An initial set of feature definitions is summarized in **Table 2**; approaches will be refined based on completion of training activities and discussion with mentors. Features to be explored include pre-defined (rule-based) features hypothesized to indicate depression based on published literature, such as the length of a message, frequency of messaging, or

specific parts of speech,(refs) as well as machine learning approaches such as sentiment analysis and contextual word embedding.(refs)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Study** | **Status** | | **Primary outcome** | **N participants** | **N SMS** | | | **Follow-up period** | **N depression measures (instrument)** |
| **Automated system** | **Personalized nurse** | **Participant** |
| **Mobile WACh XY**27 | Complete | | Family Planning | 130 | 6,852  (144 unique) | 3,356 | 3,577 | Preg - 6mo | 4 (EPDS) |
| **Mobile WACh Neo pilot** | Complete | | Neonatal Health | 799 | 26,021  (62 unique) | 13,535 | 19,952 | Preg - 3mo | 2 (EPDS) |
| **Mobile WACh X**26 | Ongoing | | HIV PMTCT | 276 | 39,993  (1149 unique) | 6,361 | 10,122 | Preg - 2yr | 6 (PHQ-9) |
| **Mobile WACh Neo RCT** | Planned by mid-2020 | | Neonatal Health, Maternal Depression | 2,500 | *NA* | *NA* | *NA* | Preg - 2mo | 2 (EPDS) |
|  | **Total to date** | | | 1,205 | 72,866  (1355 unique) | 23,252 | 33,651 |  | |
| **Table 1. SMS data sets** | | **Anticipated total** | | **3,705** | **224,000** | **71,500** | **100,000** |

**Model training and testing**: help – how to select which model to use? Do people try multiple approaches? Logistic regression? Support vector machine? Non-linear approaches? I don’t think for this model probability or interpretability is important, but I’m still not sure how to choose.

Logistic regression will be used to predict depression classification based on SMS features. In order to reduce the risk of overfitting, 10-fold cross-validation will be used.(ref) Initially, a separate predictive model will be developed with each proposed feature as the only predictor. Each model’s performance will be assessed by visual inspection of receiver operator characteristic curve (ROC), and calculation of sensitivity (recall), specificity, F1 score (the harmonic mean of sensitivity and specificity), and area under the ROC curve (AUC, a measure that combines the true positive and false positive rates). The optimal model will be selected based on consultation with clinical mentors about ideal trade-off between sensitivity and specificity in the context of SMS management. The model with optimal training performance will be selected for testing, and performance reported on the testing data.

|  |  |
| --- | --- |
| **Table 2. Proposed Aim 1 predictive features** | |
| **Feature** | **Definition / source** |
| Time of day | Obtained from message metadata |
| Gestational age | Obtained from participant study records |
| Time since enrollment | Obtained from participant study records |
| Time since previous participant SMS | Calculated from message metadata |
| Number of characters | Calculated per message |
| Bag of words | Vector representing the frequency of every word in the corpus vocabulary in an SMS |
| Sentiment | Defined by applying VADER28 |
| Language used | Obtained from message metadata |
| Pre-defined key word  (eg. “stress”, “sad”) | One-hot (dummy) encoding |
| Pre-defined parts of speech   (eg. first-person pronouns) | Defined by applying XXX |
| Contextual word embedding | Defined by applying neural network methods such as ELMo10 or BERT25 |

**Additional considerations:** A unique aspect of our dataset is the presence of multiple intermixed languages (English, Swahili, Luo, and Swahili slang known as Sheng). All non-English messages have been manually translated into English by study nurses. Our goal is to build a model that can be used to triage incoming participant messages in their original form, so to the extent possible, we will use original untranslated language. Some feature definitions can be applied to a multilingual corpus (for example features from metadata, BERT embedding25, key words, or bag of words). Others may require modification of existing packages, separation into single-language corpora, or use of English translations.

**Statistical power:** The goal of this aim is to develop a predictive model with optimal sensitivity and specificity based on the available data and clinical application – not inference of association between message features and depressive symptoms. Given the prevalence of depression in our cohorts and in similar populations,29–31 we expect that approximately 10% (10,000) SMS in our corpus will indicate depression. Published models using similar corpus sizes have achieved F1 scores of up to 0.84 using older methods,24 so we expect similar or better results. While inclusion of interpretable features in the final model may be hypothesis-generating, no formal hypothesis testing will be performed.

**C5. Aim 2. To determine core components of SMS messages sent to participants that are associated with improvement in longitudinal depression symptoms in *Mobile WACh* studies.**

|  |
| --- |
| **Table 3. Aim 2 predictors** |
| **Study SMS features** |
| Time since participant SMS |
| Number of characters |
| Bag of words |
| Sentiment |
| Pre-defined key word  (eg. “stress”, “sad”) |
| Pre-defined parts of speech   (eg. first-person pronouns) |
| Language |
| **Confounders** |
| Age |
| HIV status |
| Partner violence |
| Gestational age |
| Last participant SMS features |

**Rationale:** While Aim 1 focused on identifying messages *from* participants that indicate depression, this aim seeks to determine what message content should be sent *to* participants in order to treat depression. Our overarching goal is to provide guidance to healthcare workers managing Mobile WACh messages; we do not propose to create a computer “chatbot” that can generate natural language responses. Analysis in this aim is therefore inferential rather than predictive, necessitating two key differences in approach compared with Aim 1: (1) In order to maximize interpretability, we will only use rule-based definitions of SMS features, rather than ML word embeddings which are difficult to interpret and use as guidance. (2) Our model will take account of repeated measures and confounding to improve effect estimates.

**Data source:** This aim will use SMS sent to participants in the Mobile WACh studies listed in **Table 1**. This includes both the automated system messages (which are each sent to multiple participants), and unique personalized messages written by a study nurse in response to a participant message.

**Analysis approach:** The primary associations of interest are between Mobile WACh SMS features and subsequent participant depression status. Features of study SMS will be defined using rule-based approaches as described in Aim 1. Features of interest include pre-defined key words (e.g. “sorry”, “understand”), parts of speech (e.g. questions, directives), sentiment, and language (Swahili, English or Luo) (**Table 3**). As in Aim 1, untranslated language will be used for feature definition to the extent possible. To account for repeated measures of both SMS features and depression, we will use generalized estimating equation (GEE) models clustered by study participant with a logit link and exchangeable correlation structure. The associations of interest are confounded by a number of covariates, including time-varying participant characteristics known to be associated with depression,(refs) and depression-predictive participant SMS features identified in Aim 1. Confounders will therefore be included in a multivariable model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 4. Minimum detectable effect sizes** | | | | |
| **Predictor prevalence** | **Depression prevalence in unexposed** | **Depression prevalence in exposed** | | **Relative depression difference** |
| 10% | 7% | 4.8% | | 0.48 |
| 10% | 4.1% | | 0.41 |
| 15% | 3.3% | | 0.33 |
| 25% | 7% | 21.7% | | 0.87 |
| 10% | 21.1% | | 0.84 |
| 15% | 20.5% | | 0.82 |
| α=0.05, β=0.8, 2-sided tests | | |

**Statistical power:** **Table 4** summarizes detectable differences in depression prevalence by predictor prevalence, assuming a total sample size of 3000 women and depression prevalence similar to that observed among Mobile WACh participants.29 Availability of multiple measures per participant is expected to increase power beyond these estimates.

**C6. Aim 3. To develop and pilot a just-in-time adaptive variant of *Mobile WACh* that identifies participant SMS indicating depression and directs HCW response.**

Relationship of proposed pilot to the Mobile WACh NEO Trial

**C7. Pitfalls and limitations**

**Messages do not lead to improvements in depression**

**Insufficient SMS data**

**Limitations of depressive symptom scales.** The proposed analyses rely on scores from the PHQ9 and EPDS as indicators of depression. Although both instruments have been validated in the Kenyan context,(ref) diagnostic instruments and classifications are known to be limited in their representation of patient experiences (consider citing 32). Nevertheless, the tools developed through this project will be adaptable to other modes of classifying maternal mental health.

JITAI framework is that participant needs change over time and as context changes. Can our model capture if one type of message works in some contexts and a different message in another context?

1. Kenya Ministry of Health. Kenya Mental Health Policy 2015 to 2030. 2015.

2. WHO. Depression and other common mental disorders: global health estimates. WORLD HEALTH ORGANIZATION. 2017. PMID: 934574

3. Eisenstein J. Natural language processing. 2018.

4. Jurafsky D, Martin JH. Speech and Language Processing: An Introduction to Natural Language and Speech Recognition. 3rd ed. 2018.

5. Hirschberg J, Manning CD. Advances in natural language processing. Science (80- ). 2015 Jul 17;349(6245):261–266.

6. Yim WW, Yetisgen M, Harris WP, Sharon WK, Kwan SW. Natural Language Processing in Oncology: A Review. JAMA Oncol. 2016 Jun 1;2(6):797–804. PMID: 27124593

7. Gonzalez-Hernandez G, Sarker A, O’Connor K, Savova G. Capturing the Patient’s Perspective: a Review of Advances in Natural Language Processing of Health-Related Text. Yearb Med Inform. Thieme Medical Publishers; 2017 Aug;26(1):214–227. PMID: 29063568

8. Smith NA. Contextual Word Representations: A Contextual Introduction. arXiv Prepr. 2019;arXiv:1902.

9. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. 2013 Jan 16;

10. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. Proc NAACL. 2018.

11. Young T, Hazarika D, Poria S, Cambria E. Recent Trends in Deep Learning Based Natural Language Processing.

12. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. New York, NY: Springer New York; 2013.

13. Daume HI. A Course in Machine Learning. 2012.

14. Abbe A, Grouin C, Zweigenbaum P, Falissard B. Text mining applications in psychiatry: a systematic literature review. Int J Methods Psychiatr Res. 2016 Jun;25(2):86–100. PMID: 26184780

15. CALVO RA, MILNE DN, HUSSAIN MS, CHRISTENSEN H. Natural language processing in mental health applications using non-clinical texts. Nat Lang Eng. 2017;23(5):649–685.

16. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychol Med. 2019 Jul 12;49(9):1426–1448. PMID: 30744717

17. De Choudhury M, Counts S, Horvitz E. Social Media as a Measurement Tool of Depression in Populations. WebSci’13. 2013.

18. Nguyen T, Phung D, Dao B, Venkatesh S, Berk M. Affective and content analysis of online depression communities. IEEE Trans Affect Comput. 2014;5(3):217–226.

19. Nguyen T, O’Dea B, Larsen M, Phung D, Venkatesh S, Christensen H. Differentiating Sub-groups of Online Depression-Related Communities Using Textual Cues. 2015. p. 216–224.

20. B.L. C, A.M. P, P. C, B. M, S. H, E. B-G, Cook BL, Progovac AM, Chen P, Mullin B, Hou S, Baca-Garcia E. Novel Use of Natural Language Processing (NLP) to Predict Suicidal Ideation and Psychiatric Symptoms in a Text-Based Mental Health Intervention in Madrid. Comput Math Methods Med. 2016;2016:8708434. PMID: 27752278

21. Moulahi B, Azé J, Bringay S. DARE to Care: A Context-Aware Framework to Track Suicidal Ideation on Social Media.

22. Coppersmith G, Leary R, Crutchley P, Fine A. Natural Language Processing of Social Media as Screening for Suicide Risk. Biomed Inform Insights. SAGE Publications; 2018;10:1178222618792860. PMID: 30158822

23. Almeida H, Queudot M, Kosseim L, Meurs M-J. Supervised Methods to Support Online Scientific Data Triage. Springer, Cham; 2017. p. 213–221.

24. Milne DN, McCabe KL, Calvo RA. Improving Moderator Responsiveness in Online Peer Support Through Automated Triage. J Med Internet Res. 2019 Apr 26;21(4):e11410. PMID: 31025945

25. Devlin J, Chang M-W, Lee K, Google KT, Language AI. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018.

26. Drake AL, Unger JA, Ronen K, Matemo D, Perrier T, DeRenzi B, Richardson BA, Kinuthia J, John-Stewart G. Evaluation of mHealth strategies to optimize adherence and efficacy of Option B+ prevention of mother-to-child HIV transmission: Rationale, design and methods of a 3-armed randomized controlled trial. Contemp Clin Trials. Elsevier Inc.; 2017;57:44–50. PMID: 28315480

27. Harrington EK, Drake AL, Matemo D, Ronen K, Osoti AO, John-Stewart G, Kinuthia J, Unger JA. An mHealth SMS intervention on Postpartum Contraceptive Use Among Women and Couples in Kenya: A Randomized Controlled Trial. Am J Public Health. 2019 Jun;109(6):934–941.

28. Hutto CJ, Gilbert E. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth Int Conf Weblogs Soc Media. Ann Arbor, MI; 2014.

29. Osborn L, Ronen K, Matemo D, Unger JA, Drake AL, Kinuthia J, John-Stewart G. The prevalence and correlates of depression in a cohort of HIV-Infected pregnant women in Kenya. AIDS 2018. 2018.

30. Osok J, Kigamwa P, Stoep A Vander, Huang KY, Kumar M. Depression and its psychosocial risk factors in pregnant Kenyan adolescents: A cross-sectional study in a community health Centre of Nairobi. BMC Psychiatry. BMC Psychiatry; 2018;18(1):1–10. PMID: 29776353

31. Madeghe BA, Kimani VN, Vander Stoep A, Nicodimos S, Kumar M. Postpartum depression and infant feeding practices in a low income urban settlement in Nairobi-Kenya. BMC Res Notes. 2016;9(1). PMID: 27931248

32. Kendler KS. The Phenomenology of Major Depression and the Representativeness and Nature of DSM Criteria. Am J Psychiatry. 2016 Aug 1;173(8):771–780. PMID: 27138588