Change on the fly: using machine learning in the monitoring and evaluation of digital health programs to improve evidence generation and implementation in real time

Diwakar Mohan MD, DrPH*
dmohan3@jhu.edu
Department of International Health, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St, Baltimore, Maryland, USA

Jean Juste Harrisson Bashingwa PhD
jeanjuste@aims.ac.za
Computational Biology, Department of Integrative Biomedical Sciences, & Member of the Institute of Infectious Disease and Molecular Medicine (IDM), Faculty of Health Sciences, University of Cape Town

Pierre Dane MS
pierre@jembi.org
Division of Epidemiology and Biostatistics, School of Public Health and Family Medicine, University of Cape Town, Cape Town, South Africa

Sara Chamberlain
sara.chamberlain@in.bbcmediaaction.org
BBC Media Action, E-21, Market Lane, Hauz Khas, New Delhi, Delhi 110016, India

Nicola Mulder PhD
nicola.mulder@uct.ac.za
Computational Biology, Department of Integrative Biomedical Sciences, & Member of the Institute of Infectious Disease and Molecular Medicine (IDM), Faculty of Health Sciences, University of Cape Town

Nicki Tiffin PhD, MPH
Nicki.Tiffin@uct.ac.za
Centre for Infectious Disease Research in Africa, Institute of Infectious Disease and Molecular Medicine, University of Cape Town.
Computational Biology, Department of Integrative Biomedical Sciences, Faculty of Health Sciences, University of Cape Town

Amnesty E. LeFevre PhD MHS
alefevre@gmail.com
Division of Epidemiology and Biostatistics, School of Public Health and Family Medicine, University of Cape Town, Cape Town, South Africa
Department of International Health, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St, Baltimore, Maryland, USA

Corresponding author*
Abstract

Background: Digital health programs, which encompass the sub-sectors of health information technology, mobile health, electronic health, telehealth, and telemedicine, have the potential to generate ‘big data’. Appropriate implementation of machine learning (ML) methodologies can facilitate dynamic, real-time interventions to improve program implementation, evaluation, and ultimately decision-making on the optimal use of technology in the health sector. As part of efforts to evaluate two digital health programs in India – the maternal mobile messaging service (Kilkari) and the mobile training resource for frontline health workers (Mobile Academy) – we illustrate possible applications of ML for public health practitioners which can be applied to generate evidence on effectiveness and improve program implementation.

Methods: Study participants include pregnant and postpartum women (Kilkari) as well as frontline health workers (Mobile Academy) across 13 states in India. Data elements are drawn from system generated databases used in the routine implementation of programs to provide users with health information. We explain the structure and elements of the extracted data and the proposed process for their linkage. We then outline the various steps to be undertaken to evaluate and select final algorithms for identifying gaps in data quality, poor user performance, predictors for call receipt, user listening levels, and linkages between early listening and continued engagement.

Conclusions: Rigorous evaluations of digital health programs are limited, and few have included applications of ML. By describing the steps to be undertaken in the application of ML approaches to the analysis of routine system generated data, we aim to demystify the use of ML not only in evaluating digital health education programs but in improving their performance. Where analysis articles offer an explanation of the final model selected, here we aim to emphasize the process; thereby illustrating to program implementors and evaluators with limited exposure to ML its relevance and potential use within the context of broader program implementation and evaluation.
1. Introduction

Machine learning (ML) is an application of artificial intelligence which aims to allow computers to learn automatically from the analysis of large, highly granular datasets, with minimal human intervention {Mooney, 2018 #1084}. In ML, models are created using training data to make predictions about future events. Applications of ML in global public health are emerging, particularly in the context of digital health solutions which have the potential to generate large amounts of ‘big data’. Digital health encompasses the sub-sectors of health information technology, mobile health (mHealth), electronic health (eHealth), telehealth, and telemedicine.

ML approaches in digital health have been mainly in the area of analyzing data generated by wearable sensors and accelerometers including efforts to predict and personalize monitoring systems for mobile patients {Clifton, 2014 #8}; predict physical activity type and energy expenditure {Ellis, 2014 #9; Kerr, 2016 #13; Ellis, 2014 #10; Pärkkä, 2006 #1088}; and falls {Fahmi, 2012 #1086}. Beyond analyses of accelerometer data, studies have explored user engagement with different apps on mobile phones and tablets {Rahman, 2017 #29}, and responses to patient feedback on health services {Engelhard, 2018 #1070}. Similar applications to social media data have sought to improve the detection of online illegal drug sales {Mackey, 2017 #17}; depression {De Choudhury, 2013 #1087; Yang, 2015 #1090}; as well as explore vaccination sentiment trends and improve disease identification {Du, 2017 #14}. Analyses of Geographic Information System data have been used to map risk of exposure to disease {Wang, 2006 #1085}. Collectively these varied applications of ML have been classified by Mooney et al into three broad categories of (1) surveillance, including systems to monitor trends in disease incidence, health behaviors, and environmental conditions; (2) hypothesis-generating research, and (3) causal inference {Mooney, 2018 #1084}.

Evidence gathering on the effectiveness of digital health solutions is a growing field {Free, 2013 #413}. However, very few evaluations have sought to incorporate ML algorithms {Engelhard, 2018 #1070} and broader guidelines on the monitoring and evaluation of digital health solutions have stopped short of outlining options for predictive modeling {LeFevre, 2016 #658}. Appropriate implementation of ML methodologies can facilitate dynamic, real-time interventions to improve data collection to assess program effectiveness; and can also inform how data are used prospectively to improve program implementation.

In this paper, we outline methods proposed for the application of ML to the evaluation of two large scale mHealth initiatives in India, which have scaled to over 13 states in India since their initiation in 2012-2013, with data currently held in different databases located in Gurugram1. Kilkari is an outbound service that delivers weekly, gestational age appropriate audio messages about pregnancy, childbirth, and childcare directly to families on their mobile phones, starting from the second trimester of pregnancy until the child is one-year-old. Accredited Social Health Activists (ASHAs) mobilize women in the community to attend outreach and primary health center activities where Auxiliary Nurse Midwives (ANMs) collect and register details of mothers and their pregnancies and, after delivery, children born in their catchment areas. Mobile Academy (MA) is an Interactive Voice Response (IVR) audio training course for ASHAs in India. The training material delivered over the phone is designed to refresh their knowledge of life-saving preventative health behaviors and improve their interpersonal communications skills.

Through the use of these two digital health examples – the maternal mobile messaging service (Kilkari) and the mobile training resource for frontline health workers (MA) – we illustrate possible applications of ML which can be applied to generate evidence on effectiveness, as well as to more broadly improve

1 With the exception of the Mobile Network Operator (MNO) call data records which are head in the MNO’s datacenter in Delhi.
program implementation. We intend this paper to be an easy reference for public health practitioners considering the applicability of ML for digital health solutions, rather than a comprehensive review of the field of ML. In the course of the paper, we seek to provide an explanation in layman terms of the methods under consideration, for an audience unfamiliar with ML algorithms or advanced statistical methodologies. We also provide references throughout the text for further in-depth reading.

2. Methods
Study aims and objectives
Study aim 1: Identify factors influencing dropouts at different points along the continuum of RCH and program databases
Objective 1a. Determine differences in database records that remain subscribed to Kilkari throughout the duration of service versus those that dropped at different points along the continuum of RCH and program databases;
Objective 1b. Develop a classifier for identifying different categories of dropouts.

Study aim 2: Facilitate the program’s ability to identify and target ASHAs likely to perform poorly on digital health training programs and knowledge assessments
Objective 2a. Determine predictors of training course completion (overall and time to completion) by ASHAs based on performance on early modules of the course and other characteristics including, reported motivation, knowledge, individual characteristics, and mobile literacy.
Objective 2b. Develop a classifier for the routine identification of ASHAs likely to perform poorly.

Study aim 3: Understand the factors underpinning successful receipt of calls (Are calls received?)
Objective 3a. Determine what proportion of calls successfully reach the end-user’s device for Kilkari;
Objective 3b. Identify the proportion of content specific to infant feeding and family planning received by end-users for Kilkari.

Study aim 4. Determine whether users are listening to calls received
Objective 4a. Determine predictors for exposure to Kilkari content based on user characteristics
Objective 4b. Measure exposure to Kilkari content based on technological and behavioral (end-user engagement) performance
Objective 4c. Develop a classifier for measuring listening levels

Study Aim 5: Understand optimal message delivery options for maximal impact.
Objective 5a. Determine the time of day with the most success in engaging client and assess patterns of listening including duration and frequency of messages;
Objective 5b. Determine the effect of early listening patterns (time of day of listening, duration and frequency of listening, content listening patterns) on postpartum engagement and overall exposure.

Data sources and flow
Data collection
ANMs collect and register details of pregnant women and, after delivery, of postpartum women and children born in their catchment areas. These data are captured in print registers and uploaded at the block level by a data entry operators, forming the data in the MCTS and/or RCH databases. The data collected includes personal identifiers including geographic location, names of women and a mandatory mobile phone number and, where available, details of the pregnancy and childbirth. Data uploads happen in two

2 Many states in India are currently transitioning from MCTS to RCH. Ultimately, it is envisioned that the RCH will replace the MCTS and serve as the core Reproductive and Child health tracking system for the public health sector in India.
key time periods – (1) the earliest is the registration of the woman at the time of the identification of pregnancy; (2) following childbirth, when the birth details for the pregnancy and delivery care are available. In actual practice, these events may happen many days or months after the event (pregnancy registration or birth of child) has happened.

**Databases**

Figure 1 summarizes the databases and flow of data for both MA and Kilkari. The following is list of existing databases:

1) State-based databases that pre-date existing Mother-Child Tracking System (MCTS) and are integrated with MCTS;
2) Mother-Child Tracking System (MCTS) database at the national level;
3) Reproductive Child Health (RCH) database at the national level;
4) Call Data Records captured by Mobile Network Operator (MNO) stored separately in the MNO’s databases and mainly used for billing purposes;
5) Call Data Records of Kilkari subscribers and ASHAs usage of MA captured by the IVR system and stored in the IVR database in a data center contracted by the government;
6) The MOTECH database, which is integrated with the RCH and MCTS databases and integrates information from call data records with a small set of MCTS and RCH data for each user;
7) The MIS database, which extracts data from MOTECH and generates reports. The sampling frame for both MA and Kilkari are derived from the MCTS and RCH data. Data from the RCH and MCTS databases are pulled into the MOTECH database on a pre-determined schedule every day.

**Data Processing**

For the Kilkari program, the pregnant woman or postpartum woman’s data are captured in the RCH and MCTS systems, or in state based systems that then pass data to RCH or MCTS, and from there are sent to MOTECH system. Before the data is accepted by MOTECH, the system automatically runs validations to check that the mobile numbers are in the correct format, locations match location masters in the MOTECH database and LMP and DoB are within the Kilkari timeframe. The MOTECH system uses the last menstrual period or the delivery date to determine the schedule of messages to be delivered. The MOTECH engine provides the list of phone numbers (clients) to be called each day to the IVR system, which then calls the numbers and plays the appropriate pre-recorded message, which is stored in the IVR system’s content management system. If the call is not answered, then the IVR system attempts to call again at least 3 times every day for 4 days until the call is answered.

For the MA program, details on ASHAs including their names, phone numbers, geographic location, and age are contained in either the RCH or MCT databases, or in state-based databases integrated with MCTS and used to register them to MA. The MOTECH engine captures these data on ASHAs from the RCH or MCTS databases and following registration to MA, ASHAs are eligible to call in to the IVR system using the same phone number provided in the RCH database. The IVR system validates the phone number against the MOTECH system and then retrieves the ‘bookmark’ information which details the status of the ASHA and her progress on the list of content expected to be covered. Based on this ‘bookmark’ information, the appropriate content is delivered to the ASHA via the IVR system and the updated ‘bookmark’ data return to the MOTECH database.
Figure 1. Summary of data flow for Kilkari and Mobile Academy
<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
<th>Sample of data elements anticipated for use in analyses</th>
<th>Variables</th>
<th>Reference to figure 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCTS and RCH databases</td>
<td>National databases on reproductive, maternal newborn and child health care seeking among mothers and children under 5 yrs Physically located in the NIC data center in Delhi</td>
<td>District, Mother &amp; Child unique IDs Pregnancy IDs ASHA IDs Last Menstrual Period, Date of birth of child</td>
<td>Geographic identifiers Unique Episode / Beneficiary identifier</td>
<td>DB2, DB3</td>
</tr>
<tr>
<td>Mobile Technology for Community Health (MOTECH) database</td>
<td>Program database containing data on women and children as well as registered ASHAs. Database works in conjunction with two algorithms – (1) Kilkari and (2) Mobile Academy – which function as ‘engines’ for running the program. Physically located in Railtel data center, Gurugram, Haryana</td>
<td>Kilkari Activation date, deactivation date, Calls answered Duration of content heard per message</td>
<td>Duration of enrollment in the program, Status of enrollee</td>
<td>DB4</td>
</tr>
<tr>
<td>Interactive Voice Response database</td>
<td>Records when calls are triggered, what happens after the call is triggered (i.e. does it get answered or does it fail, and if it fails why has it failed – i.e. network errors, device errors) and whether the call needs to be retried, if yes how many times. Records similar data for incoming calls. Physically located in Railtel data center, Gurugram, Haryana</td>
<td>Kilkari Content File Name, Message Duration*</td>
<td>Message identifier, duration of listening to message</td>
<td>DB6</td>
</tr>
<tr>
<td>Call records database</td>
<td>Operator database on call handling; Physically located in the Reliance data center, New Delhi</td>
<td>Call status, call failure reasons, Call start time, Call stop time (and additional elements)</td>
<td>Duration of listening to content</td>
<td>DB7</td>
</tr>
</tbody>
</table>

Table 1. Sample of data elements by source for Kilkari & Mobile Academy
*Also captured in MOTECH
Analysis

**Extraction and linking of program data from databases**
The data from the databases (Figure 1) will be extracted onto secure password protected hard drives from each server storage. Merging data files will be complex given the nature of identifiers across databases. An MCTS record does not have a beneficiary ID, rather it has a ‘mother’ (pregnancy) ID or a ‘child’ IDs. In other words, MCTS tracks pregnancies and births, rather than women. When Kilkari first went live in October 2015, it mirrored the MCTS approach, and generated subscription IDs for each pregnancy and then birth. However, the new RCH database does have a unique beneficiary ID, which enables the government to track an individual woman through her multiple pregnancies and the births of children. The architecture of the MOTECH database and Kilkari was changed in December 2016 to introduce a unique beneficiary ID. MOTECH was then integrated with RCH in mid-2017. There is an additional complexity – namely that MOTECH used to allow multiple Kilkari subscriptions on one mobile number, assuming a single phone could be shared by a number of women in a joint family. However, a decision was made to remove this feature in 2017 (28th July for RCH and 6th October for MCTS) due to the complexity it created in analyzing system generated data. Hence the analytic time horizon assumed in the analysis may span from 2017-2018 after the MCTS-RCH integration occurred and the unique beneficiary ID was introduced into MOTECH, and multiple subscriptions on one mobile number removed. The merging of datasets will occur in India and only de-identified data will be stored on the hard drives and used in this analysis. As part of Study Aim 1, we will examine the quality of the data for completeness, including patterns and any geographic clustering in missingness.

**Additional data from baseline surveys of evaluation of Kilkari and Mobile Academy**
Analyses mentioned here are being carried out as part of a large external evaluation of Kilkari. Elsewhere we describe concurrent efforts to undertake a randomized controlled trial (RCT) in the state of MP for Kilkari, inclusive of baseline surveys with pregnant (n=4,500) and postpartum (n=880) women, and ASHA workers (n=1,200). The sample of pregnant women from the RCT is, as yet, unregistered on the MOTECH platform. Once identified as part of baseline survey activities and randomized to receive Kilkari content (or no content at all), phone numbers will be fed directly into the MOTECH database for provision of program services. For pregnant women, additional data collected as part of baseline household surveys include demographic (age, education, parity, literacy), socioeconomic (household assets, conditions), health care seeking and practices, as well as data on mobile literacy and phone access. These data can be linked to MOTECH, IVR and call center records to provide additional data elements. Overall, these data as well as data on technology performance (receipt of messages) and user engagement (behavioral performance) with content will help to yield the estimate of exposure to Kilkari used in the assessment of causality as part of the RCT. For ASHAs, baseline survey data will include similar data elements on demographic, socioeconomic, and mobile literacy and phone access as well as knowledge and work-related variables linked to reported motivation and satisfaction. Overall these added data elements can be linked to IVR and call record data for this sub-population of MA and Kilkari users in 4 districts of MP where the RCT is underway.

**Data processing and analysis**
Descriptive statistics, including univariate plots like histograms, will be used to understand the distribution of each variable, including skewness and outliers. Multivariate plots like scatterplots and lowess lines will be used to understand the relationships between different variables. Efforts to prepare the data are divided into two parts: 1. Splitting data into training and testing groups; and 2. Data processing.

1. **Splitting into training & testing**
To avoid overfitting models that work well for the data in hand but fail to predict well with other datasets, the data will be split into three components, which is possible due to the large size of the dataset. The training set will comprise 60%, the test dataset 20% and the validation dataset 20% of the data. The test set will be used to test and fine-tune the accuracy of predictive models; and the final selected model will be applied to the validation dataset. We anticipate having data from 2017, 2018 and 2019 and will ensure equal representation by random sampling. To ensure that the data is controlled for time as a confounder, subsets will be equally represented across different time periods. Descriptive statistics will be generated to ensure that key characteristics of the three subsets are similar.

2. Processing
Data processing is the act of preparing the data from its raw format into one usable by the ML models. Indications for data processing will include the following: (1) Making the data easier to use - new indicators will need to be created to facilitate their use as predictors; (2) Reducing computational cost of many algorithms by decreasing the number of variables, especially correlated and collinear variables; (3) Removing noise due to outliers; (4) Making the results easier to understand.

The most common methods by which algorithms learn about data to make predictions are supervised, unsupervised, and semi-supervised learning {Mooney, 2018 #1084}. Supervised learning trains algorithms using example input and output data, previously labeled by humans. Data may be labeled – a term used to denote that the outcome (or class) is known (e.g. ASHA has completed the training module or not completed the training module) – or unlabeled. In contrast, unsupervised learning is concerned with uncovering structure and patterns within complex datasets based on information that is neither classified nor labeled. In unsupervised ML, the algorithms learn to inherent structure based on unlabeled input data using clustering techniques. Semi-supervised learning is a hybrid analytic technique; applied in contexts where the majority of data points are missing outcome information and yet prediction remains the goal {Mooney, 2018 #1084}.

In this program context, supervised ML algorithms are expected to be the primary analytic method employed because analyses are focused on classification using predictors and available data are expected to be predominately labeled. The transformation of variables may be achieved by a variety of techniques including the creation of composite indicators and box-cox transformations. Unsupervised ML techniques, including dimensionality reduction techniques like principal components analysis (PCA) or K-means clustering will be carried out, as appropriate. PCA uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The first principal component has the largest possible variance, and accounts for highest proportion of the variance in the data, with each succeeding component accounting for the highest variance possible after accounting for the previous components. K-means clustering is a way to use data to uncover natural groupings within a heterogeneous population (Table 1). To uncover patterns, the algorithm starts by first assigning data points into random groups. The group centers are then calculated, and the group memberships are re-assigned based on the distances between each data point and the group centers. This process is repeated until there are no changes in the group memberships from the previous iteration {Armstrong, 2012 #41}. In its application to MA, K-means clustering will be used to detect patterns in ASHA engagement with training content, including training initiation and completion. Amongst Kilkari users, K-means clustering will be used to assess patterns in exposure to content by user characteristics based on data elements available in the RCH, including parity, age, and geographic area.

Training of algorithms
Once data have been processed, algorithm testing will be carried out. Table 1 summarizes the algorithms proposed for training along with their intended applications to MA and Kilkari. To determine the model with the best fit, we will explore several machine learning approaches in turn. Models will be fit on the
training set, and the fitted model used to predict the responses for the observations in the validation set. The preferred analytic approach(es) will be selected based on their ability to minimize the total error of the classification; where the latter is defined as the probability that a solution will classify an object under the wrong category. We describe each approach considered below in lay terminology, along with indications for use, and its proposed application in the evaluations of MA and Kilkari.

Table 2. Summary of algorithms proposed for testing and their intended application to MA and Kilkari

<table>
<thead>
<tr>
<th>Name of algorithm</th>
<th>Description</th>
<th>Usage examples from the literature</th>
<th>Intended Evaluation Application</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Supervised</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic regression (LR)</td>
<td>Classification (non-linear model)</td>
<td>Determinants of early registration into a maternal messaging program in South Africa [LeFevre, 2018 #1042]</td>
<td>Classification of ASHA workers by user characteristics and patterns of training initiation by ASHAs</td>
</tr>
<tr>
<td>Linear discriminant analysis (LDA)</td>
<td>Classification (linear model). It is a linearization of Gaussian naïve Bayes.</td>
<td>Digital health related to Telemedicine [Chakraborty, 2016 #35]</td>
<td>Classification of ASHA workers by user characteristics and patterns of training initiation, completion and performance</td>
</tr>
<tr>
<td>Support Vector Machines (SVMs)</td>
<td>SVMs are techniques based on the calculation of the maximum margin hyperplane for the classification problems</td>
<td>Classify data from Digital Database for Screening Mammography [Soares Servulo de Oliveira, 2015 #36]</td>
<td>Classification of ASHA workers by user characteristics and patterns of training initiation and completion</td>
</tr>
<tr>
<td>Classification and Regression Trees (CART)</td>
<td>Predictive model that consists of leaves that represent the target and branches that represent conjunctions of inputs features. Considered a subset of decision trees. Random forests operate by constructing multiple decision trees during training and aggregating their results to avoid overfitting by single trees.</td>
<td>To evaluate Wrist accelerometers [Ellis, 2014 #9]</td>
<td>Classification of ASHA workers by user characteristics and patterns of training initiation and completion</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>Classification model based on probabilities.</td>
<td>Optimise mHealth</td>
<td>Classification of ASHA workers by user characteristics</td>
</tr>
</tbody>
</table>

User characteristics associated with exposure to messaging content and duration in MP where face to face survey data have been collected.
<table>
<thead>
<tr>
<th>Neural Networks (NN)</th>
<th>NN are powerful models for ML. They are a generalization of linear and non-linear models</th>
<th>To assess the feasibility of cardiac auscultation using smartphones with no add-on devices {Kang, 2018 #37}</th>
<th>Classification of ASHA workers by user characteristics and patterns of training initiation and completion</th>
<th>User characteristics associated with exposure to messaging content and duration</th>
</tr>
</thead>
</table>

**Unsupervised**

| K-means     | K-means clustering is a way to use data to uncover natural groupings within a heterogeneous population | Improving the screening of user reports on disrespect and abuse among incoming SMS messages to a national helpdesk in South Africa {Engelhard, 2018 #1070} | Grouping of ASHA workers by user characteristics and patterns of training initiation and completion | User characteristics associated with exposure to messaging content and duration |

*Assumes criteria for differentiating high versus low are clearly defined*

**Logistic regression (LR)**

LR models, while commonly associated with biostatistics, are used in supervised machine learning to predict correctly the target variable given the input set of predictors. They are suitable when the outcome variable of interests is a dichotomous variable. Evaluations of digital health programs have frequently used LR to assess a range of outcomes, including the effect of sociodemographic predictors of eHealth {Kontos, 2014 #32} and mobile app {Shelus, 2017 #31} as well as user characteristics associated with early registration into programs {LeFevre, 2018 #1042}. In the context of the evaluation for MA, logistic regression will be used to predict training completion of ASHA workers. For Kilkari, LR will similarly be used to predict factors underpinning exposure to health information content including user characteristics, IVR and call center records. Limited data on user characteristics will be drawn from the basic data elements available in the MCTS/ RCH as well as the more comprehensive data on characteristics collected as part of face to face surveys in four districts of MP.

**Linear Discriminant Analysis (LDA)**

LDA is based upon the theory of obtaining a linear combination of variables (predictors) that best separates two classes (targets). LDA, closely related to principal component analysis (PCA), attempts to look for linear combinations of variables which best explain the data and explicitly attempts to model the difference between classes. However, LDA uses labeled data (supervised ML) whereas PCA relies on data which are not labeled (Unsupervised ML). LDA works best when the predictor variables are continuous and used when groups are known a priori. We describe a simple example of how LDA could work in our study of the Mobile Academy program using the dichotomous outcome variable of completion of the
training course within one month of enrollment (Early and late finishers) and continuous predictor variables - (i) Motivation of ASHA (ii) Score obtained by ASHA in the first module of the training course.

Figure 2 describes a simplified application of LDA for the classification of ASHAs into two classes (1) Early finishers, (blue dots) and (2) Late finishers (red dots). The two predictor characteristics - ASHA motivation (y-axis) and ASHA score on knowledge assessments (x-axis) are plotted on a scatter plot. The solid line is LDA decision boundary. For a new observation, if it falls above the solid line, the LDA will assign it to the first class (Early finisher class), otherwise if it falls below the solid line, it will be assigned to the second class. LDA have been used to assess digital health programs in telemedicine framework {Chakraborty, 2016 #35}.

Support Vector Machine
A Support Vector Machine (SVM) performs classification by finding the “hyperplane” that maximally separates two classes (or outcomes). The vectors (cases) that define the hyperplane are the support vectors. In the simplest case with two predictor variables, a straight line would separate two classes which is the “hyperplane” for this case. This method is best used with continuous predictor variables and categorical outcomes.

Figure 3 describes a simplified application of SVM for classification. Similar to Figure 2, in this example, two classes of ASHAs are shown: (1) Early finishers, (blue dots) and (2) Late finishers (red dots). The maximal margin hyperplane is shown as a solid line. The margin is the distance from the solid line to either of the dashed lines. Although SVM and LDA look pictorially similar, they use completely different algorithms.
Classification and Regression Trees
Decision tree builds classification or regression models in the form of a tree structure by breaking down a dataset into smaller and smaller sub-datasets building a decision tree in the process. The final result is a tree with decision nodes and leaf nodes with the decision node (predictor variable) having two or more branches (e.g., categories of predictor variable). Leaf node (Outcome) represents a classification or decision, a dichotomous (usually) outcome. The topmost decision node in a tree corresponds to the best predictor and is called root node. They usually handle categorical variables with those using continuous ones called as regression trees. Regression trees work similar to classification trees, with the exception that the criterion for choosing a split is the mean squared error in each leaf node.

Classification trees are predictive models that consists of leaves representing training completion or training performance status, and branches that represent conjunctions of inputs features. The training phase of the algorithm consists of constructing the decision tree, i.e. learning the branches that lead to a tree that correctly classifies as many examples in the training data set as possible. We use recursive binary splitting to grow a classification tree and predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs.

Random Forest (RF)
Random forests are a collection of randomized decision trees {Ellis, 2014 #9}. To classify a test example, the outputs from each decision tree are averaged to determine the overall output. A probability score is assigned according to the ratio of training examples of each class that belong to the leaf node. These probability scores are averaged over each tree in the forest to obtain an overall probability score for the example. Finally, the class type with highest probability is predicted for that example.

Naive Bayes
The Naive Bayesian classifier is based on Bayes’ theorem of independence between predictors. A Naive Bayesian model (used for discrete variables) is easy to build, since it does not have iterative parameter estimation making it useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier is widely used since it often outperforms more sophisticated classification methods.

Neural Networks (NN)
NN (named after neurons in an animal brain) are ML algorithms based on a collection of connected units or nodes called artificial neurons. They consist of input layer (features), hidden layers and output layer (target). To create nonlinearity, an activation function (differentiable function) is applied for each hidden layer. In NN if the activation function is the identity, we have a linear model. If the activation function is a sigmoid function with one hidden layer, we have a logistic regression model.
Our choice of methods includes a mix of algorithms based on their strengths and weaknesses with relation to the data. A comprehensive comparison of supervised learning methods is provided in literature {Kotsiantis, 2007 #39;Kotsiantis, 2006 #38}. SVM and NNs perform better with continuous data while the Naïve Bayes method and decision trees perform better with discrete/categorical variables. Naïve Bayes and decision trees have good tolerance to missing values, while NNs and SVM do not. NNs and Naïve Bayes have difficulty handling irrelevant and redundant attributes (extra variables with no useful information or variables with too many categories and too few numbers), while SVM and decision trees are insensitive towards them. Variables with high correlation negatively affect the performance of both Naïve Bayes and NNs, whereas SVM are relatively robust to correlated variables. While Naïve Bayes is robust to noise, NNs are sensitive to noise and susceptible to overfitting. NNs and SVM perform well with multi-dimensional data and when there is a non-linear relationship between predictor and outcome. Naïve Bayes require less memory for both training and validation phase, whereas NN requires large memory allocation across all phases. SVM and NNs usually outperform other methods while Naïve Bayes may yield less accurate results. Table 3 compares the strengths and weakness of different supervised machine learning methods.

<table>
<thead>
<tr>
<th>Table 3. Performance comparisons of learning algorithms modified from Kotsiantis et al {Kotsiantis, 2007 #39;Kotsiantis, 2006 #38}</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy in general</strong></td>
</tr>
<tr>
<td>Speed of learning with respect to number of attributes and the number of instances</td>
</tr>
<tr>
<td>Speed of classification</td>
</tr>
<tr>
<td>Tolerance to missing values</td>
</tr>
<tr>
<td>Tolerance to irrelevant attribute</td>
</tr>
<tr>
<td>Tolerance to redundant attributes</td>
</tr>
<tr>
<td>Tolerance to highly interdependent attributes</td>
</tr>
<tr>
<td>Dealing with discrete/ binary/ continuous attributes</td>
</tr>
<tr>
<td>Tolerance to noise</td>
</tr>
<tr>
<td>Dealing with danger of overfitting</td>
</tr>
<tr>
<td>Attempts for incremental learning</td>
</tr>
<tr>
<td>Explanation ability/ transparency of knowledge/ classification</td>
</tr>
<tr>
<td>Model parameter handing</td>
</tr>
</tbody>
</table>

**** stars represent the best and * the worst performance
**Testing and validation** To facilitate decision making on the optimal analytic approach, three steps will be undertaken: 1. Develop the correct model for each algorithm using the training dataset; 2. Apply the final model for each algorithm on the test dataset; and 3. Apply the best performing algorithm on the validation dataset.

In Step 1, algorithms will be run using the training dataset comprised of 60% of the total sample from across all states for which data are available. For each algorithm, iterative testing will be run to select the best model which fits the data. The emerging results are then assessed for model fit and accuracy. Table 4 below summarizes the four proposed metrics for assessing the performance of each model.

<table>
<thead>
<tr>
<th>Table 4. Metrics for assessing the performance of each model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Term</strong></td>
</tr>
<tr>
<td>Accuracy</td>
</tr>
<tr>
<td>Precision or Positive predictive value.</td>
</tr>
<tr>
<td>Sensitivity or Recall</td>
</tr>
<tr>
<td>Area Under Receiver Operating Characteristic Curve (AUROC)</td>
</tr>
</tbody>
</table>

To illustrate the definition of performance metrics for MA, we define true positives (TP) as the number of correctly classified ASHAs who have completed the training. True negatives (TN) as the number of correctly classified ASHAs who have not completed the training. False Positives (FP) as the number of false classified ASHAs who have completed the training, while false negatives (FN) are the number of false classified ASHAs who have not completed the training.

Results from the performance metrics will help to define the final model for each algorithm. In Step 2, these final models for each algorithm will be applied to the test dataset which is comprised of ~20% of the total data. Using the same performance metrics, the models with the best fit and accuracy will be applied to the validation dataset as part of Step 3. Ultimately, predictions for MA will aim to determine the probability of the ASHA finishing the course in a predetermined time frame and the possible score / performance of the individual ASHAs. For Kilkari we will determine predictors for exposure to Kilkari content based on user characteristics, as well as explore the effect of early listening patterns on postpartum engagement and overall exposure.

3. **Discussion**

This paper presents the testing of a range of ML approaches to be incorporated as part of the evaluation of two large digital health programs in India – MA and Kilkari. By utilizing ML approaches, we aim to
improve the use of data for generating evidence on program reach and exposure as well as factors underpinning uptake for both programs. We will start by measuring dropped cases and missing data along the continuum of the databases. As part of this analyses we aim to understand differentials in the characteristics of individuals lost along continuum of databases and in turn, missed by the MA and/or Kilkari programs. We then consider the technological performance as captured by four stakeholders: 1. Government’s data system (MCTS, RCH, webservises); 2. Program’s call delivery and receipt systems; 3. Mobile network operator (network coverage and quality); and 4. User device characteristics (switched on, within range of network). Finally, we explore user engagement with the program (behavioral performance) including patterns in the data which could predict key program performance metrics, including training completion for MA and user coverage and exposure for Kilkari.

Analyses to measure dropped cases and missing data along the continuum of the databases, starting with the MCTS AND RCH databases, are anticipated to generate insights into user relay of information to the Government of India including the provision of phone numbers by health workers and women. At present, there are no reliable estimates of the proportion of pregnant women with accurate mobile numbers, last menstrual periods, date of birth (for themselves or child) covered by the databases. More broadly, these analyses will help improve understanding of the quality of the RCH and MCTS data and underlying sampling frames used for MA, Kilkari, and a range of other programs, including the characteristics of individuals included compared to those excluded. The predictors identified will help identify groups of women who are more likely to be missed by the program and ASHAs less likely to complete their course. These findings will also be used to inform the sampling of respondents for qualitative research to explore issues in depth and, possibly, identify opportunities for improving program targeting like registration of users.

Analyses to understand the technological performance of the program will build off of those conducted on the MOTECH platform in Ghana {LeFevre, 2017 #911} and of MomConnect in South Africa {LeFevre, 2018 #1042}. In the former, IVR message delivery trends suggested that 25% or less of expected mobile health information messages were received by pregnant women {LeFevre, 2017 #911}. While 20% delivery rates of successful out-bound dialing (OBD) calls is standard in the mobile industry, limitations in the timeliness of their identification represent a missed opportunity for improving program exposure using, for example, call retry logics and systems, and likely effectiveness. In South Africa, over 80% of short messaging service (SMS) messages were successfully delivered as part of MomConnect. SMS as a delivery channel has a much higher success rate than OBD calls but suffers from other weaknesses in countries where illiteracy rates are high and local language fonts are not widely available on devices. In the case of MomConnect, however, challenges in the use of Unstructured Supplementary Service Data (USSD) meant that 26% of initiated registrations do not covert into successful registrants {LeFevre, 2018 #1042}. These two examples reinforce the need to understand the user journey and follow the flow of data to understand whether the technology performs as intended.

Beyond understanding the technological performance of the program, and the wider telecommunications network and device landscape that it operates within, analyses will aim to measure user engagement, including predictors for exposure to content based on user characteristics. In the context of Kilkari, we will additionally plan to explore the effect of early listening patterns (time of day of listening, duration and frequency of listening, content listening patterns) as a predictor for postpartum engagement and overall exposure. For MA, we will use a mixture of unsupervised and supervised ML techniques to generate predictors of training course completion by ASHAs based on performance on early modules of the course. These will be externally validated using a range of data elements on other ASHA characteristics obtained from the broader evaluation, including reported motivation, knowledge, individual characteristics, and mobile literacy.

Overall, the proposed analyses are anticipated to complement primary data collection activities proposed
as part of the summative evaluation of Kilkari and Mobile Academy. Findings emerging from this analysis will provide program implementers with tools for improving predictions of success and performance and provide insights into strategic use and collection of data. Elsewhere, deployments of similar programs including MomConnect in South Africa, Aponjan in Bangladesh as well as other maternal messaging programs may be able to apply some of the same approaches used here.

**Limitations**

Ethical issues related to identifiers are important considerations for analyses described here. We will de-identify the databases at the point of data download and all data will be secured in a password protected hard storage device with access controlled by the study Principal Investigator. Once data are accessed, we note that findings will only be as reliable as the quality of underlying data. Completeness of the coverage of data and the data elements can be limited in low-resource settings. In recognition of this challenge, we will assess the quality of data for completeness and timeliness. The inherent design of the messaging program means that only those with mobile phones can be part of the sample, which could introduce selection bias in the association between outcome and predictors. Confounding will be an issue due to the many unmeasured variables associated with exposure like mobile phone ownership, and registration into the pregnancy tracking database, and outcomes like active listening to messages and adoption of healthy behaviors. Call answering does not automatically mean listening to the message by the intended client (e.g. ASHAs or pregnant women) and as such our analyses will be limited in its ability to identify the listener (information bias). However, the potential benefits of having other family members listen to information content could be immense. Beyond challenges with the measurement of exposure, the absence of a complete set of predictors will be present due to the limitations of such large-scale data gathering processes (incomplete model).

4. **Conclusions**

This manuscript aims to provide a survey of approaches to the applications of ML to improve the implementation and evaluation of digital health programs. The two digital health examples described represent two of the largest digital health programs, based on number of active users, currently being implemented globally. Developing classifiers based on the above ML approaches will help identify gaps in data (Study aim 1), target potential slow learners (Study aim 2), measure exposure to program (Study aim 3), characterize exposure levels in population (Study aim 4), and/or evaluate program impact (Study aim 5). The scale of implementation and associated generation of data on user engagement with program content provides opportunities for big data analytics and more specifically, the utilization of ML approaches to improve the generation of evidence on program reach and exposure as well as factors underpinning uptake for both MA and Kilkari.

**Acronyms**

- ASHAs: Accredited social health activists
- ANMs: Auxiliary Nurse Midwives
- CART: Classification and Regression Trees
- IVR: Interactive Voice Response
- LDA: Linear discriminant analysis
- LR: Logistic regression
- MA: Mobile Academy
- MCTS: Maternal and Child Tracking System
- mHealth: mobile health
ML Machine learning
MP Madhya Pradesh
NN Neural Networks
PCA Principal component analysis
RF Random Forest
RCH Reproductive Child Health
SVM Support Vector Machines

Acknowledgments
We thank Diva Dhar and Suhel Bidani of the Bill and Melinda Gates Foundation for their guidance and support.

Contributions
AL conceived the overall evaluation study with input from DM and SC. DM and AL wrote the draft of the manuscript with assistance from JB. NT, SC, PD, and NM provided extensive feedback on the drafts of the manuscript.

Financial Disclosure
This work was funded by the Bill and Melinda Gates Foundation (BMGF) to the Johns Hopkins University and University of Cape Town. The contribution by NT is supported by Wellcome (203135/Z/16/Z), and the National Human Genome Research Institute (NHGRI), Office of the Director, National Institutes of Health (OD) under the H3ABioNet award, number U24HG006941.

Ethical considerations
The registration data on pregnant women and ASHAs are collected by the Ministry of Health and Family Welfare of the Government of India and the ministries of health of the states participating in the program. The data will be analyzed under a data sharing agreement with the Bill & Melinda Gates Foundation and Johns Hopkins University, University of Cape Town and BBC Media Action. The Institutional Review Boards of Johns Hopkins School of Public Health, Sigma in Delhi India, and the University of Cape Town have provided the ethical certification for the study.

References


