Measuring Data Collection Diligence for Community Healthcare

Ramesha Karunasena
rameshak@smu.edu.sg
Singapore Management University
Singapore

Mohammad Sarparajul Ambiya
sarfraz@khushibaby.org
Khushi Baby
India

Arunesh Sinha
aruneshs@smu.edu.sg
Singapore Management University
Singapore

Ruchit Nagar
ruchit@khushibaby.org
Khushi Baby
India

Saachi Dalal
saachi@khushibaby.org
Khushi Baby
India

Divy Thakkar
dthakkar@google.com
Google Research
India

Dhyanesh Narayanan
dhyaneshn@google.com
Google Research
India

Milind Tambe
milindtambe@google.com
Google Research
India

ABSTRACT

Data analytics has tremendous potential to provide targeted benefit in low-resource communities, however the availability of high-quality public health data is a significant challenge in developing countries primarily due to non-diligent data collection by community health workers (CHWs). Our use of the word non-diligence here is to emphasize that poor data collection is often not a deliberate action by CHW but arises due to a myriad of factors, sometime beyond the control of the CHW. In this work, we define and test a data collection diligence score. This challenging unlabeled data problem is handled by building upon domain expert's guidance to design a useful data representation of the raw data, using which we design a simple and natural score. An important aspect of the score is relative scoring of the CHWs, which implicitly takes into account the context of the local area. The data is also clustered and interpreting these clusters provides a natural explanation of the past behavior of each data collector. We further predict the diligence score for future time steps. Our framework has been validated on the ground using observations by the field monitors of our partner NGO in India. Beyond the successful field test, our work is in the final stages of deployment in the state of Rajasthan, India. This system will be helpful in providing **non-punitive** intervention and necessary guidance to encourage CHWs.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Artificial intelligence**.

KEYWORDS

Community Healthcare, Data Quality, Data Collection Diligence, Clustering, Social Impact

ACM Reference Format:
Ramesha Karunasena, Mohammad Sarparajul Ambiya, Arunesh Sinha, Ruchit Nagar, Saachi Dalal, Divy Thakkar, Dhyanesh Narayanan, and Milind Tambe. 2021. Measuring Data Collection Diligence for Community Healthcare. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21), October 5–9, 2021, –, NY, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3465416.3483292

1 INTRODUCTION

Community health workers (CHWs) play a vital role in health care systems, especially in developing countries where they provide health care to the rural communities. Among the many contributions made by frontline CHWs, an important one is *collection of population health data*. Such data is crucial for many tasks, such as health resource allocation, health predictions etc. However, even with extensive human infrastructure to enable data collection in *developing countries*, some CHWs are not able to deliver high quality data [9] due to various social factors. Data quality challenges have prevented the use of such data for downstream applications and when used, occasionally led to sub-optimal outcomes. Poor data collection can be attributed [14, 24] to various factors such as overburdened CHWs, lack of functional equipment, etc. While we recognise several factors are embedded in complex social problems, in this paper we specifically focus on factors associated with data diligence that can be improved through interventions such as mentoring and motivation. We recognize the importance and challenges of CHWs [14, 15, 23, 36] in providing medical care to rural and marginalized communities. Field experts (including our NGO partner Khushi Baby) struggle with fundamental design techniques [3] of how to effectively measure the diligence of CHWs in terms of accurate data collection, how to predict diligence, and how to design targeted interventions for improving data collection.

We find that designing the data collection diligence score problem has 2 main challenges: (1) absence of labeled data for indicators of diligence and (2) the diligence of a CHW is not expected to be the same over time, which brings in a temporal aspect. While the

EAAMO '21, October 5–9, 2021, –, NY, USA

ACM ISBN 978-1-4503-8553-4/21/10…$15.00
https://doi.org/10.1145/3465416.3483292

problem may seem solvable by an unsupervised anomaly detection approach [7, 10], we found (and know from the NGO) that a large percentage of the reported data is suspected to be of poor quality, which rules out unsupervised anomaly detection. Another approach, which is commonly used, is to utilize rules designed by domain experts for non-diligence detection. However, we found that such rules with fixed hard thresholds are too stringent as all CHWs violates some of these rules and the fixed threshold fails to take the local context into account.

Our first technical contribution is to provide a framework for converting rules of non-diligence detection, designed by domain experts with decades of experience, into probabilities of non-diligence for each CHW. All these rules measure some percentages in a time window, for example, one rule tracks the percentage of number of blood pressure readings in a camp over a month, that are all same. We posit that a CHWs diligence must be measured *relatively* with respect to the performance of other CHWs in the local area; such relative measurement implicitly accounts for the local context, in which a health measurement task might be difficult. Further, we are interested in ranking the CHWs for the purpose of later intervention to improve diligence. Thus, for each rule, we consider the relative position of a given percentage reading of a CHW in a month within the distribution of percentage readings of all CHWs in all months; this relative position is converted to a probability of non-diligence for the CHW for that rule in a time window. Then, the multiple rules provide a multi-dimensional vector of non-diligence probabilities for every month. We call these health-worker specific probability vectors as their *behavior*. However, a multi-dimensional vector does not allow ranking CHWs.

Our second technical contribution includes a definition of a scalar non-diligence score for each CHW for each month as the Euclidean norm of the behavior vector described above. Then, we predict this scalar score using the past six months of a CHW's scalar scores as features in a linear regression predictor. We observed that, given limited data size, linear regression outperforms other complex models such as neural networks. The ranking enabled by scalar scores allows for intervention using limited resources in future.

Our final technical contribution is to provide fine grained explanation of the past behavior of each health worker by clustering the behavior vectors and interpreting each cluster. We worked closely with the non-profit partner to learn the challenges of CHWs to ensure this technology was developed to meet the needs of the community. We also perform this interpretation of the clusters in three varying information levels. The fine-grained explanation provides insights about the broad types of non-diligence exhibited in past by a CHW for targeted intervention.

Our *main highlight* of this work is a field test by our NGO partner, where the diligence scores and behavior explanations were validated against on-field observations made by monitors who observed the CHWs in Rajasthan, India. Overall we observe that 67% of our predictions match with the observations of the monitors; moreover, in additional 17% cases our tool correctly identifies non-diligence that the monitors miss out. Thus, we achieve a 84% correct non-diligence detection in the field. In addition, we are also able to explain the reasons for not matching in the remaining 16% cases, This paved the way for monitors to closely examine the algorithm outputs and retained their agency to make final decisions. Based on this field test, we have started the deployment of our work in the government controlled server for use by our NGO partner. We firmly believe that our diligence score, prediction of the same and the behavior explanations coupled with well-designed intervention, will over time result in improved data quality.

In collaboration with our NGO partners we hope to design systems to provide better care for targeted beneficiaries, using models to *predict stillbirths, infants with low birth weight and drop-outs from camps*. The non-diligence score along with the performance clusters from our framework can provide an indication of the quality of data used in these models, which might lead to better predictions and thus benefiting millions of pregnant women and newborn children.

Ethical considerations: Our work was reviewed and covered by appropriate IRB for the work done in educational institutes and internal ethics review was conducted for the work done by the NGO team as well as appropriate permission sought and obtained from the state government. As stated above, a future work topic is design of intelligent interventions that are effective in nudging the CHWs to be more diligent in data collection. The design of interventions is a complex social and organisational challenge [29] with ethical considerations. *Our interventions planned are not punitive*; in fact, in our problem the CHWs receive a fixed salary from the state government and do not receive any monetary penalties for their data work. Our intervention will focus on encouragement via phone messaging and active one-on-one engagement with the CHW to understand specific issues in data collection. Intervention for CHWs on other aspects have shown tremendous promise in addressing related problems [20, 33]. Our work was conducted in close partnership with the NGO Khushi Baby who have many years of history of working with CHWs and regular feedback was sought from CHWs through field monitors along with direct interactions with CHWs.

2 RELATED WORK

There is a long history of dealing with data quality issues in machine learning. Many of these approaches focus on scenarios where the data is noisy or adversarial corrupted but the amount of such noise or corruption is small. This low perturbation assumption naturally leads to unsupervised anomaly detection based techniques [7, 10, 21, 22, 34, 35, 37]. However, as stated earlier, poor quality data is not an anomaly in our problem; in fact, majority of the data is poor quality due to non-diligence in collecting data. Also, with no label, supervised anomaly detection [12] cannot be used. Further, distinct from these prior work, our goal is not to measure the data quality in aggregate but to measure the diligence of the data collector, improving which would naturally lead to better quality data.

More closely related to our work is a work that directly focuses on measuring anomalous data collection by CHWs [21]. However, this work also treats poor data collected by a CHW as an outlier, which is not true in our problem. In experiments, we perform a baseline comparison with anomaly detection for sake of scientific completeness. In not so closely related work, detecting human behavior in context of fabrication [5] has been studied using behavior data from a controlled user study; we do not possess behavior data of CHWs and it is quite infeasible to collect such data.

Clustering has been used in past work to detect fraud in social networks [6], in internet advertising [32], in healthcare billing [17, 19]. Also, there have been research on cluster explanations using decision trees [8, 11]. However, as far as we know, neither clustering nor cluster explanations have been used in a problem context as ours. Moreover, our problem is not one of fraud, which is easier to define, but, one where we rely on human input to define non-diligence. Further, we explain the clustering result to yield better outcomes overall.

There are other kinds of data quality issues considered in machine learning. These include data missing over a large part of the feature space [16], adversarially corrupted data [4, 25], and imputation for missing data in healthcare [13]. Our concern in this work is orthogonal to all these other issues. Also, there is work on incentivizing workers in crowd-sourced data labelling to obtain correct labels [18, 26, 28, 31]; our context is very different: the data collection task is extremely arduous and the reasons of erroneous data collection are very varied and nuanced.

3 PROBLEM DESCRIPTION

In this work we collaborate with the NGO Khushi Baby (http://khushibaby.org), whose main mission is to motivate and monitor the health of new mothers and children. They have deployed their platform to track the health of over 40,000 mothers and infants in Rajasthan, India in partnership with the district government. The NGO employs on-field employees with decades of experience on working with CHWs; these employees work closely with a data science team of the NGO.

The health workers, called ANMs (Auxiliary Nurse Midwives), are *state government employees* and are responsible for screening pregnant women and infants at these village-based health camps. Each ANM provides antenatal care checkups and infant checkups on 4 to 5 fixed camp sessions, each in a separate village. On average, an ANM caters to approximately 200 beneficiaries across these villages. The data collection task is difficult and was paper based (see Fig. 1). The NGO, in collaboration with the state government, has enabled a streamlined digital data collection service for ANMs to collect data from patients in health camps, reducing some of the burden of the ANMs. The NGO, *with approval from the state government*, has made available datasets that correspond to camps that have been conducted in the past by ANMs.

The NGO's data science team has developed in collaboration with on-field experts a set of 11 heuristic rules and associated thresholds, which is what they were using to assess the quality of data collected—whether some health records are indicative of lack of diligence on the part of the ANM. Examples of non-diligence include: filling in test results in the absence of conducting the test, rounding results when more precise data is available, failing to record data fields that are non-mandatory, and recording test results for certain patients but reporting that the same test equipment is not available for other patients in the same camp (note that equipment failure is distinct from equipment not available). Analysis of the state's database (ANMs send data to state database either via the NGO's system or via paper records) and observations from the field have shown major gaps in data fidelity. *Less than 50 percent of expected maternal and child deaths are reported in the state's database.* Frequencies of blood pressure values in the state database show evidence of rounding, with low reporting of hypertension, which is expected at a certain baseline frequency during pregnancy. Even the number of actual blood pressure examinations done has been called into question [9].

Dataset: The data from the NGO is for 74 ANMs. Every ANM conducts camps at regular intervals in her (ANMs are females) jurisdiction. Pregnant women are advised to get at least four check-ups in the course of the pregnancy and they visit the nearest camp, when due for a check-up. The data recorded is quite extensive ranging from basic health parameters such as blood pressure and sugar levels to pregnancy-specific parameters such as fundal height and baby position. All data about the beneficiaries who visit the health camp was de-identified and anonymized. The data provided was initially from 01 January 2020 to 31 January 2021. Maternal and child health camps were suspended in April 2020 due to the COVID-19 outbreak. A data drop from February 2021 to March 2021 was used for field test.

In our working with the data-set we iteratively discovered a number of issues with the data. The first main issue is that some ANM's data had few patients per health camp; however, it was not known if these ANMs really had few patients or were they not entering data into the system. During the course of the project, it was found out (by comparing with the state database) that these ANMs were not entering a large chunk of their data into the digital system and as a consequence we decided to filter out such ANMs.

Problem statement: The long term goal of Khushi Baby is to use the data collected for AI based support for better health care. However, the NGO knows from experience as well as confirmation from other NGOs that the data collected by ANMs is often not accurate. Thus, there is a need to identify non-diligent ANMs. Concretely, this requires solving three problems: (1) how to quantitatively measure non-diligence per ANM (2) how to explain the behavior of ANM over time and (3) how to predict non-diligence per ANM in order to enable apriori intervention.

The NGO also used (before this project started) a ranking of the ANMs based on a one-time unstructured input from the field team of the NGO. We attempted to treat this ranking as labels of diligence. However, we found: (1) the rankings often contradict with what the data says, for example, highly ranked ANMs were found to be entering all blood pressure values as same and (2) the ranking was fixed over time and not accounting for changing behavior of ANMs and hence could be the reason for point 1 above. Hence, we decided to not use this ranking.

4 METHODOLOGY

We describe our approach in three distinct parts: first is defining a vector non-diligence score, second is a scalar non-diligence score for each ANM and prediction of the same, and finally generation of behavior explanations. A summary view of the system is shown in Fig. 6 (Appendix D). Our code for the system is available publicly at [2].

4.1 Defining Non-diligence

Our first main idea is to obtain a probability of non-diligence per ANM corresponding to each of the domain-expert rules provided

(a) A mother and an infant, walking several kilometers on foot to visit a health camp

(b) Paper records filled by CHWs, which are later entered into a web portal manually

Figure 1: Challenging circumstances for data collection in the rural areas of Rajasthan, India

by the NGO using the data generated by each ANM. There are 11 rules as of now (March 2021) and additional rules can be added by defining them in the configuration file of the system. These 11 rules are of two types (1) rules that track known non-diligent phenomenon and (2) rules that track contradiction in data records. All rules specify a percentage of something and we know which extreme (0% or 100%) or a defined range corresponds to diligence. The full list of rules is in Appendix C.

Running example: *Known non-diligence rule*: percentage of blood pressure readings in a month that are 120/80 or 110/70. We know that higher percentage corresponds to non-diligence. We call this the BP rule.

Handling Rules: We describe our approach here using the BP rule stated in the running example; rest of the rules are similar or flipped where 0% is diligent. For each month $t \in \{1, \ldots, T\}$ in the training data, we compute the percentages as stated in the BP rule for each ANM $i \in \{1, \ldots, n\}$. This gives us the list of percentages for the BP rules as $\{P_{i,t}\}_{i \in \{1,\ldots,n\}, t \in \{1,\ldots,T\}}$. Then, we filter out the percentages that are exactly 0% or 100% from this set $\{P_{i,t}\}$ - as these extremes are for sure diligent or not. We treat the remaining percentages in $\{P_{i,t}\}$ as samples from a probability distribution and plot a probability density distribution using Kernel Density Estimation (KDE). See Fig. 5 in Appendix D for an example KDE plot for the BP rule. Then, given the percentage for the BP rule, say $P_{i,t}$, in a month t for an ANM i, the probability of non-diligence p is the probability mass between $(0, P_{i,t})$. Clearly as $P_{i,t}$ increases, the probability of non-diligence is increasing and is exactly 1 when percentage is 100. For the rules where non-diligence with probability 1 is at the 0%, the probability of non-diligence p is the probability mass between $(P_{i,t}, 100)$.

Our measure of probability of non-diligence for a rule can be viewed as a measure of the current month's performance of an ANM relative to all the past performances of all the ANMs. This relative scoring is important as certain measurement tasks are inherently difficult due to local factors such as lack of equipment. Thus, an absolute threshold is meaningless without taking into account the difficulty of the data collection task. Our relative scoring is analogous to relative grading in university courses where the relativeness naturally takes into account the hardness of the underlying test.

Filtering of noisy ANMs: As stated earlier, we ignore ANMs with less data, according to two thresholds (both are configurable): (1) ANMs who recorded patients below a given threshold in each month and (2) ANMs who recorded a total number of patients below a given threshold in a year. Out of 74 ANMs in the data set, 8 ANMs were filtered out.

Vector valued score of ANM behavior: The above approach provides a 11 dimensional vector of probabilities $\vec{b}$ to measure each ANM's behavior at any given point in time by using past one month of data.

4.2 Non-diligence Score and Prediction

We obtain a single scalar valued non-diligence score for each ANM, which is required for the purpose of ranking the ANMs. The ideal behavior vector for an ANM is $\vec{0}$, meaning zero probability of non-diligence for every rule. Thus, given a vector $\vec{b}$, a natural score is its distance from $\vec{0}$, that is, $||\vec{b}||_2$. This score is interpretable and by definition explicitly gives equal importance to all rules. We define this as the *data collection diligence score*.

We aim to predict this diligence score of an ANM in the dataset. For prediction we use a simple linear regressor, since our data is limited. We use the diligence scores of the past six months for each ANM as features to predict the diligence score for the next month.

4.3 Clustering and Generating Interpretations

The past scores by themselves provide only a view of the relative diligence of an ANM. Further, these score do not identify broad patterns of non-diligence. In this section, we use clustering of *raw percentages* for every rule to obtain broad types of non-diligence and also interpret these clusters to generate description of the ANM's past behavior.

After obtaining the percentage vectors for every month and every ANM in the training data, we cluster these vectors into three clusters using k-means. We set the number of clusters as three, using the elbow method [27]. However, k is configurable in the system using the configuration file in anticipation of changed patterns in future. Observe that we use raw percentages when clustering and not the non-diligence probabilities. This is because the non-diligence probabilities are relative measures of performance and we wish our clusters to capture absolute ANM behavior in order to provide the correct targeted messages in the intervention stage. Once the clusters are determined using the current data, the clusters are fixed for n number of months, where n is configurable and reclustering is done only after n months. This is to ensure that we measure the future behavior vectors within a static frame.

Generating behavior interpretation: We generate interpretations for each cluster and call this as the *behavior interpretation.* Note that we do not have labels for ANM's behavior, but the nature of our constructed features (percentages for each rule) readily allow us to interpret the meaning of the clusters. Denote the 11 dimensional cluster center of cluster k as $\langle P_k^1, \ldots, P_k^{11} \rangle$. We convert the cluster centers (which is in term of percentages) to non diligence probabilities $\langle p_k^1, \ldots, p_k^{11} \rangle$ using the KDE approach we described in Section 4.1. We call the non diligence probabilities of the cluster centers as *cluster center diligence vector.*

We use the cluster center diligence vector to generate interpretations at different granularity levels. At lower levels, we provide information at a higher abstraction. For example, at level 0, we output whether the cluster is generally good or bad. More precisely, we compare the average ($\sum_{r=1}^{11} p_k^r/11$) non-diligence probability indicated by the cluster center; based on whether this average for a cluster is clearly greater than that of the another cluster, we tag the clusters as non-diligent and diligent.

For a more fine-grained interpretation of clusters, we partition the rules by computing the standard deviation of the percentages $\{P_{i,t}\}$ of all ANMs and time windows for each of the 11 rules. We create 3 partitions of the rules based on these 11 standard deviation values and call these the most (highest stddev), less, and least (lowest stddev) important rules; the thresholds for standard deviation are chosen by tuning. The least important rules are those where (mostly) percentages for all months and all ANMs are very close, i.e., all ANMs perform roughly the same, either all diligently or all non-diligently in these least important rules. This also reveals why using non-diligence probability is *not correct* for clustering as the probability score will assign relative scores from 0 to 1 to the ANMs even with small difference in the percentage measured for these rules. While this is fine for relative scoring, this is not fine for understanding behavior patterns, for example, ANMs with very similar performance on a rule might receive very different messages of how to improve if just the scores are used for this rules. The most important rules show higher variance of behavior and are considered important for intervention as the high variation in these rules inform us that some ANMs perform better than other ANMs with respect to these rules, and hence other ANMs can be encouraged to do better.

At level 1, for each cluster we output diligent or not in *each of the most important rules.* We do so by saying that if the cluster k center non diligence probability for a rule r is more than the average ($\sum_{k=1}^{3} p_k^r/3$) cluster center non diligence probability then this cluster k is interpreted as capturing non-diligent behavior for rule r. At level 2, we provide the same type of explanation for each of the less important rules. We do not provide interpretation for the least important rules, since there is no significant difference in the performance of the ANMs for these rules among the three clusters. However, we still output the least important rules for the NGO to carefully analyze why all ANMs perform the same for these rules, especially if the performance is non-diligent.

Given the cluster to which an ANM belongs in a given month, the field team can identify the diligence of the ANM for every rule using our behavior descriptions at different granularity levels. This also aids in designing future interventions accordingly. The history of clusters of an ANM also provides an idea about her behavior trend over time.

Table 1: Mean values of non-diligence scores of the ANMs tagged non-diligent and tagged diligent by the anomaly detector

	December 2020	January 2021
Tagged Non-diligent ANMs	1.8499	1.8391
Tagged Diligent ANMs	1.8537	1.7898

5 RESULTS

We present our experimental results in this section. The data-set was already described in Section 3. We process the data as we mentioned earlier. We split the data set we have into training and test data, where the split was according to time due to the sequential nature of data. Our training data was from January 2020 to November 2020 and test data from December 2020 to January 2021. An on-field observation exercise was initiated by field monitors of the NGO in February-March 2021, which is used to validate our results. We start by presenting results from a baseline method.

5.1 Baseline Methods Results

We tried two baselines. The first is a simple heuristic baseline that uses the 11 rules provided by the NGO with fixed thresholds percentage. For example, for the BP rule a threshold of 70% of all blood pressure readings in a health camp being exactly same, was used to tag ANMs as diligent or not. Note that these rules do not provide a score, so are not very desirable to start with. Moreover, these rules performed very poorly—the outcome we obtained was that either (1) no ANM was tagged as non-diligent when the rules were used in an AND manner and (2) all ANMs were tagged as non-diligent when the rules were used in an OR manner.

The other baseline that we try is anomaly detection. We used a popular variational auto-encoder based anomaly detector [1]. We pre-processed our training data over time similarly as for clustering by using one month time window and processing all rules as percentages. We did not convert the raw percentage to probabilities. We trained the anomaly detector on the training data and then tested how the non-diligence score differs for ANMs tagged non-diligent vs ANMs tagged diligent by the anomaly detector in the test data. The results for the two months in test data is shown in Table 1. ANMs tagged as non-diligent should have a higher non-diligence score than ANMs who are tagged diligent, which is not the case here. As stated earlier, this is likely because non-diligence is not an outlier and unsupervised anomaly detection is based on the outlier assumption.

5.2 Clustering Results

As described earlier our clustering done on the raw percentages provided three clusters, for which we present the three cluster centers (in percentages) in Table 2. The set of rules we used are given in Appendix C. We obtain the diligence probabilities of the cluster centers and at level 0 of interpretation the cluster with higher average probability of non-diligence is the non-diligent behavior cluster. Generally, cluster 0 is good in all rules; cluster 1 is bad in most known non-diligent rules and good in contradiction rules; and cluster 2 is bad in contradiction rules, but good in some known non-diligent rules like BP rule. Thus, at level 0 interpretation both cluster 1 and cluster 2 are non-diligent clusters. More levels of cluster interpretations are generated as explained in Section 4.3. In particular, four rules were found to be least important: rules 2,6,7,8, for which no explanations are generated. Rule 3 and 4 barely crossed the threshold for being least important and on closer inspection we do find ANMs with good behavior for these rules in cluster 0.

5.3 Prediction Results

We possess only small amount of data spanning about one year and 66 valid ANMs. We have used a simple linear regressor and the results are presented in Table 3. We do not observe very high R2 scores or perfect correlation, but the scores when paired with cluster explanations provide a good measure of diligence, which is validated in the field.

5.4 Field Test Results

We validated our results using the observations made in *a highly structured field monitoring exercise* carried out by the NGO in February-March 2021. The field exercise was conducted with approval from the government. The field monitors worked with *randomly chosen ANMs* for one week each for a total of 37 observations. Each chosen ANM was observed for one week by a monitor and the whole exercise spanned four weeks. The monitors are NGO employees with years of experience with the health camp data collection process. The monitors were provided a list of questions relating to the ANM performance (a summary of the questions is in Appendix B) and they recorded responses to these questions based on their observation. These questions provide a structured insight into the data collection diligence of ANMs as opposed to the informal ranking that was used before this project. Our NGO partner infer whether an ANM is diligent in a week based on the questionnaire, which is what we compare with the output of our AI system.

For this field test, we need to put the AI system's scores (predicted data collection diligence scores) into discrete buckets of diligent and non-diligent. We treat top 30% ANMs (top 20) as diligent, bottom 55% ANMs (bottom 36) as non-diligent and the middle 15% (10 ANMs) as ANMs whose performance is determined by past trend in score and past clusters. Among the middle 10 ANMs, we mark those as diligent who have a non-increasing trend in score and belong to good clusters in past months. We deliberately use our clusters for putting medium score ANMs into buckets in order to extract the maximal information from the AI system as possible. Moreover, in the planned intervention in future we plan to be conservative and commend the top 10 ANMs and encourage the bottom 20 ANMs to perform better. The clusters will help in targeted encouragement about which aspects to improve on.

Our NGO partners find that our AI system's results are quite useful. Results are provided in Table 4 for the 36 observations that we considered finally. We see a 67% agreement rate with the monitors' observations. Moreover, for 17% cases when AI predicts non-diligent, we can analyze the actual performance after-the-fact with the data collected by the ANM during the observation period. We find that the AI was correct in all these cases as there is evidence of non-diligence in the data collected during the observation period (see Appendix A.4 for these details). Thus, AI's success rate is about 84%. We note here that the field monitors are not able to capture these 17% non-diligence because their observation is only for a limited time and it is simply infeasible for a field monitor to monitor all health camps round the clock in all four weeks. Further, for other cases in which AI outputs diligent (but ground truth is non-diligent), we find that some of these cases of non-diligence could not have been detected from given data. For an example, ANM filling up all data for urine test from past experience without actually conducting any test.

6 DISCUSSION AND LESSONS LEARNED

Our final product has been through many iterations and prior failed field trails before the success reported in this work. Here we distill a list of key lesson learned based on our experience.

- **Lesson 1: get continuous feedback from the ground**: Our initial product was not considered useful by the field team of the NGO. The reasons for these are: (1) We had used 3 additional rules provided by the NGO, all of which required the patients to get certain tests done at main health centers which are outside the ANMs' jurisdiction and report to ANMs. An ethical review at the NGO concluded that these 3 rules might not be a fair proxy to measure ANMs' diligence. Thus, these 3 ambiguous rules were dropped in the final product. (2) We had used a binary classifier to classify diligent and non-diligent ANMs, which had more than 90% accuracy. However, the field team required a more fine-grained score in order to do the interventions, which we provide now.
- **Lesson 2: blackbox explanations are futile**: Our initial attempt at interpreting ANM behavior was by assigning importance weights to the rules (a common explainablity approach). These blackbox explanations were useless to the NGO team and hence futile in our context. Thus, we developed easier explanation using interpretation of clusters centers at different levels of descriptiveness for different perceptive skills of the stakeholders.
- **Lesson 3: use techniques relevant for the problem and not vice versa**: In an intermediate version of our tool, we used a LSTM network to predict fuzzy cmeans cluster scores generated with two clusters. Although it had high accuracy and R2 scores, the initial testing proved to be unconvincing to the field monitors' and fuzzy cmeans provided poor explanations. In fact, our linear regressor provides close performance and our strict clusters provide much better explanation of ANM's behavior.
- **Lesson 4: be ready for production environment to change**: Our experience with dropping rules and the discovery of newer non-diligence based on field monitor observations reinforce the fact that the production environment is ever evolving. For this precise reason, our AI system has the flexibility to add/change/delete rules as well as a number of other configurable parameters such as number of clusters.

Table 2: Cluster centers (Rules are numbered 1 to 11)

Cluster	1	2	3	4	5	6	7	8	9	10	11
Cluster 0	10.50	0.94	97.43	86.04	12.67	0.15	0.72	0.29	1.46	7.33	2.31
Cluster 1	70.36	0.12	99.90	88.54	3.75	2.20	2.02	1.56	8.09	4.40	4.90
Cluster 2	11.27	1.02	99.19	89.97	10.44	8.11	2.24	7.20	59.97	20.04	27.44

Table 3: MSE and R2 score of the predictor for test data

Metric	December 2020	January 2021
MSE	0.0401	0.0296
R2	0.2575	0.2624

Table 4: Comparison of the number of ANMs identified by diligent or non-diligent by the AI and the field monitors

		Monitors' observations	
		Diligent	Non-diligent
AI's result	Diligent	11	6
	Non-diligent	6	13

7 SUMMARY AND FUTURE WORK

Listing a few known limitations in broadly applying our technique, we identify that the model itself will need to be tuned to new data and new CHWs. New domain specific insights from the field may further improve the clustering and prediction; our approach is flexible to take these into account. Our work uses rules designed by domain experts for India. These rules might need to be adapted for local context when applied to elsewhere in the world.

Our work is in final stages of deployment and will continuously generate behavior explanations and predict scores every month for the ANMs as well as measure their performance in hindsight. Moreover, the data format for tracking pregnancy health used in Rajasthan is consistent with National Health Mission Guidelines followed across India. Thus, this work has the potential to be broadly applied across a health workforce of 250,000 ANMs who care for 25 million pregnant women annually. Over a longer term, high quality public health data is vital for proper allocation of limited resource in under-resourced parts of the world and we hope that this work is a step in this direction.

ACKNOWLEDGMENTS

This work was supported by Google Research's AI for Social Good Faculty Award (2019) and Singapore Management University's Lee Kong Chian Fellowship (2020). We thank the lead of the field team at Khushi Baby, Hamid Abdullah, for crucial assistance in the field test conducted for this study.

REFERENCES

[1] Jinwon An and S. Cho. 2015. Variational Autoencoder based Anomaly Detection using Reconstruction Probability. Special Lecture on IE, 2, 1-18..

[2] Anonymous. 2021. Healthworker Diligence Measurement and Explanation System. https://anonymous.4open.science/r/1762326b-a4a6-415b-997b-994fdadc1327.

[3] Amna Batool, Kentaro Toyama, Tiffany Veinot, Beenish Fatima, and Mustafa Naseem. 2021. *Detecting Data Falsification by Front-Line Development Workers: A Case Study of Vaccination in Pakistan.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445630

[4] Battista Biggio and Fabio Roli. 2018. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition* 84 (2018), 317–331.

[5] Benjamin Birnbaum, Gaetano Borriello, Abraham D Flaxman, Brian DeRenzi, and Anna R Karlin. 2013. Using behavioral data to identify interviewer fabrication in surveys. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, 2911–2920.

[6] Qiang Cao, Xiaowei Yang, Jieqi Yu, and Christopher Palow. 2014. Uncovering large groups of active malicious accounts in online social networks. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security.* Association for Computing Machinery, New York, NY, USA, 477–488.

[7] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)* 41, 3 (2009), 1–58.

[8] Sanjoy Dasgupta, Nave Frost, Michal Moshkovitz, and Cyrus Rashtchian. 2020. Explainable k-means and k-medians clustering. *arXiv preprint arXiv:2002.12538* (2020).

[9] Dinesh Songara, Akanksha Goyal, Pankaj Suthar. [n. d.]. Approach Paper, Rajasthan State. https://globalcenters.columbia.edu/sites/default/files/content/Mumbai/Publications/Approach%20Paper%20Rajasthan.pdf. Online; accessed 29 Oct 2020.

[10] Eleazar Eskin. 2000. Anomaly Detection over Noisy Data using Learned Probability Distributions. In *Proceedings of the Seventeenth International Conference on Machine Learning.* Morgan Kaufmann Publishers Inc., Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 255–262.

[11] Nave Frost, Michal Moshkovitz, and Cyrus Rashtchian. 2020. ExKMC: Expanding Explainable k-Means Clustering. *arXiv preprint arXiv:2006.02399* (2020).

[12] Nico Görnitz, Marius Kloft, Konrad Rieck, and Ulf Brefeld. 2013. Toward supervised anomaly detection. *Journal of Artificial Intelligence Research* 46 (2013), 235–262.

[13] Zhen Hu, Genevieve B Melton, Elliot G Arsoniadis, Yan Wang, Mary R Kwaan, and Gyorgy J Simon. 2017. Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. *Journal of biomedical informatics* 68 (2017), 112–120.

[14] Azra Ismail and Neha Kumar. 2019. Empowerment on the margins: The online experiences of community health workers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, 1–15.

[15] Neha Kumar, Trevor Perrier, Michelle Desmond, Kiersten Israel-Ballard, Vikrant Kumar, Sudip Mahapatra, Anil Mishra, Shreya Agarwal, Rikin Gandhi, Pallavi Lal, et al. 2015. Projecting health: community-led video education for maternal health. In *Proceedings of the Seventh International Conference on Information and Communication Technologies and Development.* Association for Computing Machinery, New York, NY, USA, 1–10.

[16] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. 2017. Identifying unknown unknowns in the open world: representations and policies for guided exploration. In *Proceedings of the AAAI.* AAAI Press, 2124–2132.

[17] Qi Liu and Miklos Vasarhelyi. 2013. Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information. In *29th world continuous auditing and reporting symposium (29WCARS), Brisbane, Australia.*

[18] Andrew Mao, Ece Kamar, Yiling Chen, Eric Horvitz, Megan Schwamb, Chris Lintott, and Arfon Smith. 2013. Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 1. AAAI.

[19] Michela Carlotta Massi, Francesca Ieva, and Emanuele Lettieri. 2020. Data mining application to healthcare fraud detection: a two-step unsupervised clustering method for outlier detection with administrative databases. *BMC medical informatics and decision making* 20, 1 (2020), 1–11.

[20] Aditya Mate, Andrew Perrault, and Milind Tambe. 2021. Risk-Aware Interventions in Public Health: Planning with Restless Multi-Armed Bandits. In *20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS).* London, UK.

[21] Ted McCarthy, Brian DeRenzi, Joshua Blumenstock, and Emma Brunskill. 2013. Towards operationalizing outlier detection in community health programs. In *Proceedings of the Sixth International Conference on Information and Communications Technologies and Development: Notes-Volume 2*. 88–91.

[22] Krishnan Naidoo and Vukosi Marivate. 2020. Unsupervised Anomaly Detection of Healthcare Providers Using Generative Adversarial Networks. In *Conference on e-Business, e-Services and e-Society*. Springer, 419–430.

[23] World Health Organization et al. 2013. *Using lay health workers to improve access to key maternal and newborn health interventions in sexual and reproductive health*. Technical Report. World Health Organization.

[24] Joyojeet Pal, Anjuli Dasika, Ahmad Hasan, Jackie Wolf, Nick Reid, Vaishnav Kameswaran, Purva Yardi, Allyson Mackay, Abram Wagner, Bhramar Mukherjee, et al. 2017. Changing data practices for community health workers: Introducing digital data collection in West Bengal, India. In *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*. 1–12.

[25] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. 2016. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814* (2016).

[26] Goran Radanovic, Boi Faltings, and Radu Jurca. 2016. Incentives for effort in crowdsourcing using the peer truth serum. *ACM Transactions on Intelligent Systems and Technology (TIST)* 7, 4 (2016), 1–28.

[27] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan. 2011. Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*. 166–171. https://doi.org/10.1109/ICDCSW.2011.20

[28] Nihar Shah, Dengyong Zhou, and Yuval Peres. 2015. Approval voting and incentives in crowdsourcing. In *International conference on machine learning*. PMLR, 10–19.

[29] Reetu Sharma, Premila Webster, and Sanghita Bhattacharyya. 2014. Factors affecting the performance of community health workers in India: a multi-stakeholder perspective. *Global health action* 7, 1 (2014), 25352.

[30] Simon J Sheather and Michael C Jones. 1991. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)* 53, 3 (1991), 683–690.

[31] Adish Singla and Andreas Krause. 2013. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *Proceedings of the 22nd international conference on World Wide Web*. 1167–1178.

[32] Tian Tian, Jun Zhu, Fen Xia, Xin Zhuang, and Tong Zhang. 2015. Crowd fraud detection in internet advertising. In *Proceedings of the 24th International Conference on World Wide Web*. 1100–1110.

[33] Kristen J Wells, John S Luque, Branko Miladinovic, Natalia Vargas, Yasmin Asvat, Richard G Roetzheim, and Ambuj Kumar. 2011. Do community health worker interventions improve rates of screening mammography in the United States? A systematic review. *Cancer Epidemiology and Prevention Biomarkers* 20, 8 (2011), 1580–1598.

[34] Liang Xiong, Barnabás Póczos, Jeff Schneider, Andrew Connolly, and Jake VanderPlas. 2011. Hierarchical probabilistic models for group anomaly detection. In *Proceedings of the 14th international conference on Artificial Intelligence and Statistics*. 789–797.

[35] Liang Xiong, Barnabás Póczos, and Jeff G Schneider. 2011. Group anomaly detection using flexible genre models. In *Advances in neural information processing systems*. 1071–1079.

[36] Deepika Yadav, Pushpendra Singh, Kyle Montague, Vijay Kumar, Deepak Sood, Madeline Balaam, Drishti Sharma, Mona Duggal, Tom Bartindale, Delvin Varghese, et al. 2017. Sangoshthi: Empowering community health workers through peer learning in rural india. In *Proceedings of the 26th International Conference on World Wide Web*. 499–508.

[37] Rose Yu, Xinran He, and Yan Liu. 2015. Glad: group anomaly detection in social media analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 10, 2 (2015), 1–22.

Table 5: Percentages of non-diligent ANMs tagged using variational auto encoder based anomaly detector

	December 2020	January 2020
Percentage	11.67	10.53

A ADDITIONAL RESULTS

All our code is written in python and R and ran on a cloud infrastructure. For KDE, we use the kde1d library in R (invoked from python) with the an adaptive Gaussian kernel bandwidth [30].

A.1 Additional Results for Baselines

As stated in Section 5, we used a variational auto-encoder based anomaly detector as our second baseline to detect non diligent ANMs. Log reconstruction plots for the two months in test data set are available in Fig. 2. The ANMs with lower reconstruction log probabilities correspond to anomalies. We chose 0 as the threshold to separate the diligent and non-diligent ANMs. Table 5 shows the percentages of ANMs tagged as non-diligent by the anomaly detector in the test set. These percentages are considered small by our NGO partners.

A.2 Additional results: Non-diligence score variation

We observed the non-diligence score variation of the training and test data from January 2020 to January 2021 (excluding April 2020) in Fig. 3.

A.3 Additional results for the predictor

We have observed that the performance of ANM varies with time (in Appendix A.2). Hence, it is important not to rank the ANMs based only on the scalar non-diligence scores of the last month. A prediction model is required to capture the past performance as well as trends and predict the scalar non-diligence score for the next month using the past scores. For an example, if an ANM had good scores at first but the performance is dropping over the past few months, it is important to identify that trend and motivate the ANM.

The distribution of the predicted scores and true non-diligence scores of the test data is shown in Fig. 4 and correlation (Pearson) of the prediction and true score is on average 0.54 (correlations can be between [-1,1] with 0 denoting no correlation).

Time window choice of one month: The size of the time window is an important consideration in formulating the non-diligence score. The ANM behavior varies a lot from one health camp to another, which is why we do not choose a time window as one health camp (one week) for non-diligence score calculation. We found that a time window of one month (four to five health camps) exhibits more stable behavior than a window of one week.

A.4 Additional results for the field test

Field monitors observations provided an independent view of ANM's diligence. Monitors are known to the ANMs as they regularly help the ANMs with the digital equipment as and when required. Our

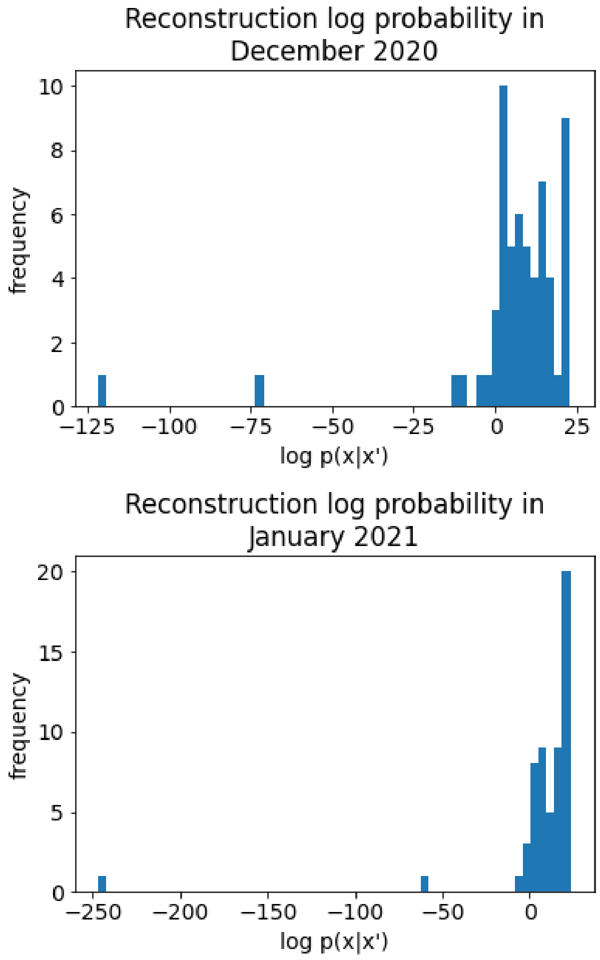

Figure 2: Log reconstruction plots using the variational autoencoder

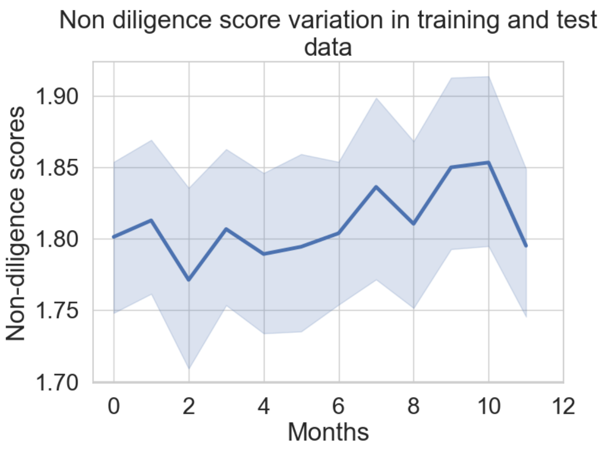

Figure 3: Non-diligence score variation in training and test data sets from January 2020 to January 2021 excluding April 2020

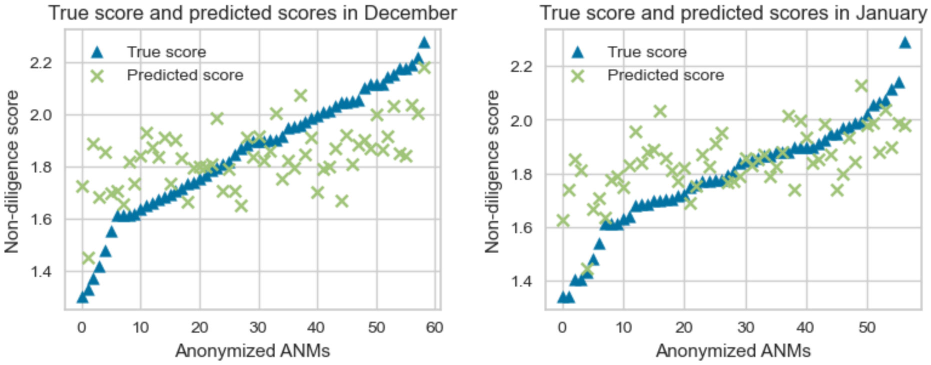

Figure 4: True scores and predicted scores per ANM distribution in December 2020 and January 2021 (x-axis are anonymized ANM ids)

Table 6: Analysis of the ANMs who are predicted as non-diligent by the ANM, but observed as diligent by the field monitors. (ANM IDs are anonymized)

Week	ANM ID	Non-diligent rule/s
2	a	Non-diligent in two short term rules (4th and 5th rules)
3	b	Non-diligent in two short term rules (3rd and 5th rules)
3	c	Non-diligent in three short term rules (3rd, 4th and 5th rules)
3	d	Non-diligent in two short term rules (4th and 5th rules)
3	e	Non-diligent in two short term rules (4th and 5th rules)
4	f	Non-diligent in one short term rule (4th rule)

NGO partner has conducted this field test after approval from the state government, specifically the Reproductive and Child Health Officer. The field monitoring task is itself very arduous as it involves travelling to many remote parts. Earlier monitoring exercise (before this project) had some implicit bias. But this exercise was specifically designed to exclude such biases. However, we note here that the field monitors are experts who have years of experience with the health camp data collection process, yet sometimes the field monitors' observations may not be perfect due to human factors involved such as mistakes. In our evaluation of the monitors recording, we discard monitor's observations for one ANM due to contradictory answers for the observation questions.

In the field test we see an agreement rate of 67%. For another 17% of ANMs, where the AI predicts as non-diligent but the field monitors observe as diligent, we analyze the actual non-diligence probability vector (corresponding to 11 rules) of each ANM in the observed period. We find that these ANMs were actually non-diligent w.r.t certain rules as shown in the Table 6, which proves that the AI was correct in predicting these ANMs as non-diligent. Hence, our AI system's accuracy is about 84%.

For further robustness checks, we also analyzed whether there was any effect on the ANMs behavior when being observed versus when they were not. We used the data of ANMs who were observed and who were not observed in 4 weeks from February 2021. Overall, we see that there is not a significant effect from the observers, as shown in the Table 7.

B QUESTIONNAIRE

The field monitors were provided a list of questions relating to the ANM performance (a summary of the questionnaire is given below) and they recorded responses based on their observation.

- Whether equipment needed for each test is available.
- Whether the ANM conduct each test properly or fills in values without checking or manipulate checked data corresponding to each of the rules. Few example questions related to BP rule recording are as below.
  - Does the ANM properly check BP for all women?
  - How many women had BP, 120/80?
  - Does the ANM enter data without the checking?
  - Does the ANM enter rounded-off BP values?
  - Does she record data for few patients and enter "No equipment available" for few other patients?
- Whether infrastructure facilities necessary for the health camp are available

C RULES USED FOR NON DILIGENCE DETECTION

We use rules provided to us by the NGO, these rules are prescribed by a team of public health and medical experts who serve as nodal technical support partners to the state department of health. In this setup we used the 11 rules mentioned below. Neither the thresholds of these rules nor the rules themselves are fixed. The rules can be easily configured using the configuration file and they can be

Table 7: Observer effect analysis

	With observers	Without observers
New score (Avg. non-diligence score in the 4 weeks of monitoring exercise)	1.6797	1.7092
Mean of the difference between new score and avg. past score	-0.1590	-0.1051
Std. Dev. of the difference between new score and avg. past score	0.2558	0.2150

adapted to another context, if the system is to be used elsewhere. We track the proportion (percentages) for *every conducted health camp* by an ANM.

(1) Proportion of 120/80, 110/70 blood pressure readings being very high is suspicious.
(2) Proportion of hypertension blood pressure readings being very low is suspicious.
(3) Proportion of urine readings as absent being very high is suspicious.
(4) Proportion of beneficiaries with anaemia (haemoglobin readings below 11) being at extremes is suspicious.
(5) Proportion of beneficiaries with severe anaemia (haemoglobin readings below 8) being very low is suspicious.
(6) Recording "no equipment available" for some patients while recording blood pressure data for other patients is suspicious.
(7) Recording "no equipment available" for some patients while recording weight data for other patients is suspicious.
(8) Recording "no equipment available" for some patients while recording haemoglobin data for other patients is suspicious.
(9) Recording "no equipment available" for some patients while recording blood sugar data for other patients is suspicious.
(10) Recording "no equipment available" for some patients while recording fetal heart rate data for other patients is suspicious.
(11) Recording "no equipment available" for some patients while recording urine data for other patients is suspicious.

These 11 rules are of two types (1) rules that track known non-diligent phenomenon (1-5) and (2) rules that track contradiction in data records (6-11). We explained a running example for the first type of rules in 4.1 and below is an example for the second type of rules.

Contradiction rule: When an ANM has entered “No equipment available” for few patients yet records fetal heart rate readings for other patients in the same health camp, it is a contradiction in fetal heart rate reading. We get the average over a month percentage of “No equipment available” records out of the total number of records in a health camp. 0% means no contradiction. Thus, unlike the BP rule, lower percentage means more diligent and higher percentages are non-diligent readings in this rule. This rule captures the contradictions of the no equipment entries. Note that *having no equipment for all patients does not trigger this rule.* But recording “No equipment available” for some patients while recording values for some patients in the same camp indicates non-diligence in this rule. In order to avoid any misinterpretation of using the contradiction rules (6-11), if the equipment breaks after use in a camp, then the “Equipment not working” option must be selected by the CHWs, rather than equipment not available. Training to ensure that CHWs understand this key difference is communicated by the NGO partners.

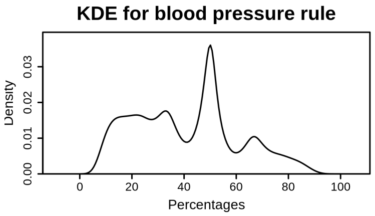

Figure 5: Kernel Density Estimation plot for the blood pressure rule after removing extremes

D ADDITIONAL FIGURES

A summary view of the system is shown in the Fig. 6 and Fig. 5 is an example KDE plot for the BP rule, which was drawn after removing the extreme data points.

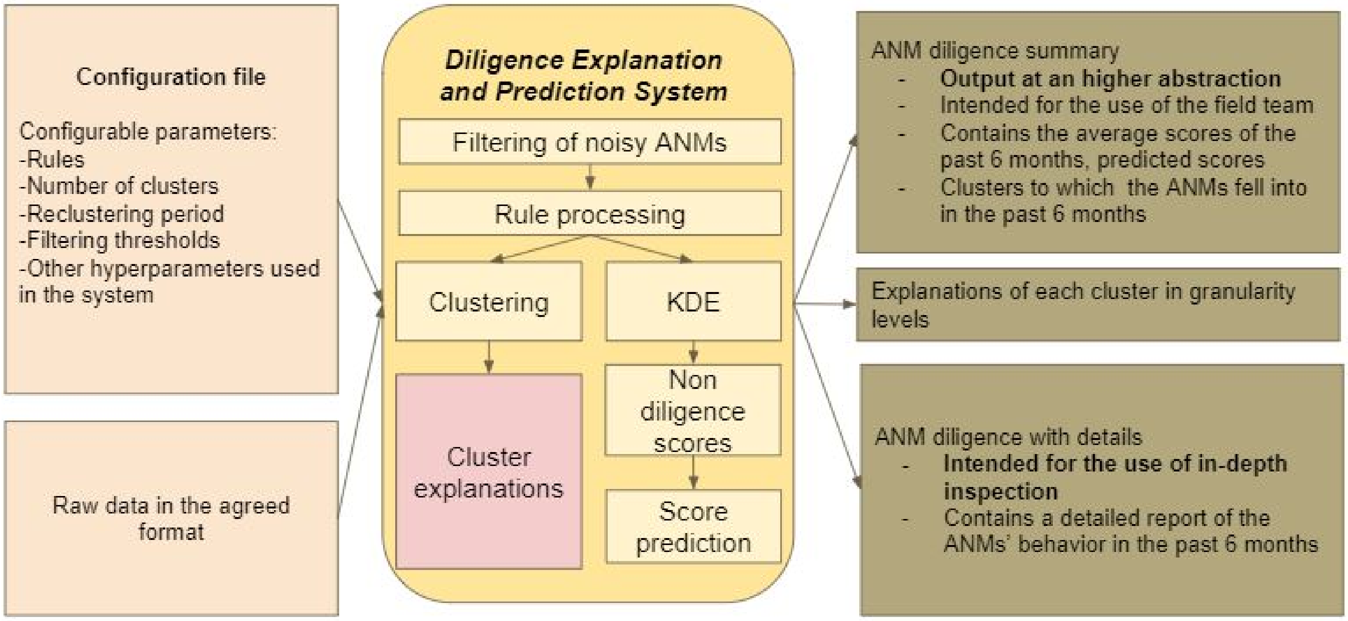
Configuration file
Configurable parameters:
-Rules
-Number of clusters
-Reclustering period
-Filtering thresholds
-Other hyperparameters used in the system
Raw data in the agreed format
Diligence Explanation and Prediction System
Filtering of noisy ANMs
Rule processing
Clustering
KDE
Cluster explanations
Non diligence scores
Score prediction
ANM diligence summary
- Output at an higher abstraction
- Intended for the use of the field team
- Contains the average scores of the past 6 months, predicted scores
- Clusters to which the ANMs fell into in the past 6 months
Explanations of each cluster in granularity levels
ANM diligence with details
- Intended for the use of in-depth inspection
- Contains a detailed report of the ANMs' behavior in the past 6 months

Figure 6: System diagram