**leveraging community health workers for predicting emergency department readmissions**

Nuvia Hernandez Kate Karam Nicole Baugh

Lewis University DePaul University**\*\*** North Carolina State University  
Romeoville, Illinois, United States Chicago, Illinois, United States Raleigh, North Carolina, United States  
 kburns10@depaul.edu nuviahernandez@lewisu.edu ncbaugh@ncsu.edu

Shilpa Musale Aditiya Patrick Moses Daniela Raicu   
jsmusale@depaul.edu apatri5@depaul.edu draicu@depaul.edu

Jacob Furst Kelly McCabe Roselyne Tchoua

Sinai Urban Health Institute, Sinai Chicago Chicago, Illinois, United States  
 jfurst@depaul.edu Kelly.McCabe@sinai.org rtchoua@depaul.edu

Received (Day Month Year)  
Revised (Day Month Year)  
Accepted (Day Month Year)

As the capabilities of machine learning have developed, more researchers and health care providers are beginning to consider applications for health informatics to improve health care outcomes and address issues of health equity. The Centers for Medicare and Medicaid Services considers 30-day readmission rates to the Emergency Department (ED) to be an “outcome of care” measure. Such measures show how well a hospital is doing in preventing complications, educating patients about their care needs, and helping patients make a smooth transition from the hospital to home or other care facilities. While certain readmissions are medically necessary, hospitals usually aim to decrease the rate of 30-day ED readmissions by decreasing the number of avoidable unplanned revisits. This work is an evidential study that demonstrates the positive impact of integrating Community Health Workers (CHWs) and Social Determinants of Health (SDoH) in decreasing the 30-day unplanned hospital ED readmissions at Sinai Chicago. Using data from the Sinai Urban Health Institute, we compare predicting the readmissions of patients with and without data pertaining to SDoH, characterize the improvement in predictions, and discuss lessons learned in the process. We show that when CHWs engage with patients, the predictive accuracy of the classifier increases by 5%. Importantly, we show that the features related to the CHWs are important to the classification, pointing to the importance of the program. We use this result to make recommendations for improving patient care and discuss limitations and future work. Notably, our work points directly to the human connection between patients and CHWs as an important feature predictive of readmission rate.

*Keywords*: data science; health equity; community health workers; social determinants of health.

# Introduction

As more healthcare researchers and professionals study the impact of social factors on health outcomes, in an effort to decrease health inequities [1]–[3], more opportunities are also emerging in health informatics (i.e., the intersection of Machine Learning (ML) and healthcare [4], [5]). There is evidence that the impact of Artificial Intelligence (AI) on minority health and health inequities has been largely understudied [6], [7] despite studies that show bias and racial inequities in health-related outcomes [8]. At Sinai Chicago (Sinai), Illinois’ largest private safety net health care system, over 70% of patients are people of color from underinvested and overburdened West and Southwest Side neighborhoods in Chicago. Sinai patients disproportionately suffer from a variety of maladies including diabetes, cardiovascular disease, gunshot wounds, violent crime, trauma, kidney disease, and breast cancer [9]. The Sinai Urban Health Institute (SUHI), the community-engaged research arm of Sinai, was an early adopter of the Community Health Worker (CHW) model, with over 23 years of experience developing and researching CHW interventions. Community Health Workers (CHWs) are trusted frontline public health workers who come from, or have a close understanding of, the communities they serve [10]. CHWs serve as liaisons between health care systems and communities, connecting personally with community members and assisting with health care outreach and navigation in an effort to address health inequities. CHWs are trained to identify and address barriers related to Social Determinants of Health (SDoH) [10], [11]. The 30-day readmission rate to the Emergency Department (ED), hereinafter referred to as the 30-day ED readmission, is an important performance metric for hospitals. This measure shows what happens after patients with certain conditions receive care at a medical center, and is one way to judge how well hospitals are delivering quality care. SUHI data, consisting of CHW patient conversation logs and SDoH surveys, along with corresponding patient 30-day ED readmission data, provided us a unique opportunity to quantify the impact of its CHW program on an important health outcome. Our work significantly highlights and documents the positive impact of SUHI’s CHW program, which has the potential to be scaled across more Sinai locations and in hospitals across the country. Our work points directly to the benefit of the CHWs: indeed, we show a 5% increase in predictive ability of classifiers for patients who are engaged with CHWs (versus those who are not) and investigate the reasons for this improvement. The contributions of this evidential work are: 1) a systematic comparison of predicting patient 30-day ED readmission for high-risk patients (engaged or not), with and without CHW data for engaged patients, and 2) a summary of lessons learned along with a discussion of our findings and recommendations. For engaged patients, we expand our analysis to include in-depth feature importance and sensitivity analysis. The novelty resides in leveraging CHW logs (e.g., calls made, time spent talking to patients or looking for resources) and SDoH data to improve the predictive capabilities of 30-day ED readmissions. The rest of this paper is organized as follows: Section 2 provides a brief background on SUHI, CHWs and the SDoH data they collect; Section 3 presents related work, Section 4 outlines our methodology including the comparison between three classifiers in predicting readmissions using baseline demographics data for all patients and non-engaged patients, then using CHW logs and SDoH data for engaged patients; Section 5 presents our results followed by a discussion in Section 6. We conclude in Section 7.

# Sinai Urban Health Institute’s Community Health Worker Program

Sinai Urban Health Institute (SUHI), the community-engaged research arm of Sinai Chicago, is a leader in developing and testing CHW centered health interventions, with over 23 years of expertise in conducting community engaged research aimed at understanding and addressing health inequities. SUHI CHWs serve as liaisons between the Sinai Chicago safety net hospital system, and the minoritized communities on Chicago’s South and West Sides. CHWs screen for SDoH that comprise a multitude of social, economic, political, and other non-clinical factors that drive patient health. Research shows that inequities in social conditions are fundamental causes of population health differences [8], [12], [13]. As such, SUHI’s efforts towards integrating CHWs as part of patient care teams aims to address underlying social needs and risk factors that impair patient health. SUHI has previously demonstrated CHW effectiveness towards improving patient care and outcomes in a variety of health areas including asthma [14,15], cancer [16, 17], and complex care management [18, 19]. Although preliminary reports have demonstrated that patients who are engaged with CHW have a 35% lower risk of 30-day ED readmissions [20], CHW data has yet to be fully leveraged and analyzed to systematically document the positive impact of CHWs on measures such as the 30-day ED readmission rate.

SUHI’s CHW program workflow begins with a patient referral from within Sinai departments such as Social Work, among others. Sinai departments typically refer patients based on a short, point-based screener that aims to detect medium and/or high-risk patients. Once a CHW is assigned a patient, they reach out to the patient and provide resources as needed, connecting patients to community-based resources, helping with discharge instructions, scheduling follow-up appointments, and providing emotional support. Part of this process involves notating patient conversations, collecting contact logs, documenting patient answers to an SDoH survey(which helps detect patient needs both healthcare related and social needs), and providing referrals for associated resources (e.g., questions about food insecurity are followed with options for food resources). When the CHW identifies an unmet social need, they provide tailored resources and referrals and serve as a navigator between patients and needed resources. SUHI’s CHWs are trained in motivational interviewing, cultural competence, SDoH, and health education (among other topics), and are able to connect with patients on a personal level in a manner that other clinical team members sometimes cannot. CHWs continue to work with their patients over time, conducting follow-up check-ins for an average of one month.

# Related Work

Here we describe related work in predicting 30-day ED readmissions in general and using SDoH data after providing some general background.

### Hospital 30-day ED Readmissions

Hospital patient 30-day ED readmission is one of the most critical quality outcome measurements alongside mortality and complication rates [21]. Patient 30-day ED readmissions is a quality of care as well as a performance metric for hospitals, often associated with a significant cost, sometimes considered a waste when readmissions could have been avoided. For example, the annual cost of readmissions to the healthcare system is estimated to be 17.4 billion for Medicare [22]. Hospitals are further invested in reducing their rate of avoidable ED readmissions as government agencies link their payment to such metrics. For instance, beginning October 1, 2012, under the Patient Protection and Affordable Care Act, the United States Department of Health and Human Services and Centers for Medicare & Medicaid Services established Hospital Readmissions Reduction Program (HRRP) to reduce high hospital patient readmissions rates [23], linking payments to quality of hospital care, in turn motivating hospitals to improve their ED rates.

### Predicting Hospital 30-day ED Readmissions

One way to reduce the rate of 30-day ED readmissions is to anticipate and prevent them. Hence, there is great interest in accurately predicting readmissions. Hospitals commonly use rule-based (point-based) screening tools like the international HOSPITAL score as a clinical predictor of patients at high risk of being readmitted within 30 days [24]. Several studies have endorsed this tool over the similar LACE index for large cohorts [22, 25]. A recent survey shows that ML models, particularly tree-based models, are promising techniques to improve predictive ability in patient ED readmissions [26]. Logistic regression is the most commonly used approach in this survey. The study also concludes that model performance varies with target populations and that state-of-the-art deep learning models do not necessarily outperform other models and exhibits a black box nature that mitigates the possibility of broad adoption [26]. Importantly, the authors point out that while some features, such as social factors, have been associated with a high risk of readmission, these data are not readily available in health institutions.

### Predicting Hospital 30-day ED Readmissions with SDoH

SDoH data contribute extensively to health inequities, which are “the unfair and avoidable differences in health status” seen within and between groups [27] and can be difficult to study [28]. Recognizing this, the National Academy of Medicine (NAM) recommended the inclusion of social environment measures in Electronic Medical Records (EMR)s to reduce bias [29]. There are several studies showing that a wide range of SDOH is strongly correlated with certain health outcomes such as cardiovascular diseases, diabetes and more [30-33]. Yet most risk prediction models do not include comprehensive study of the impact of SDoH on accuracy due to the lack of quality data [34]. The challenges to collecting quality data include the lack of consensus on standardized SDoH screening questions [35], need for consent, lack of trust and privacy concerns [36]). Due to this lack of readily available data, researchers attempt to extract SDoH from unstructured data (e.g. clinical notes [37] or infer community-level SDoH data from EHRs (e.g. smoking, alcohol use). Due to the limited availability of individual-level SDoH data, many studies are limited to community-level SDoH data (unemployment rate, access to public transportation, air pollution levels) [34].

Keeping in mind the challenges around collecting quality SDoH data, most previous studies demonstrate that (mostly community-level) SDoH do not lead to improvement in model performance [34,38-40]. However, several of these works also found that for some subgroups of patients, SDoH did improve the predictive accuracy of models. In [40], the authors attempted to add SDoH to the LACE index and concluded that some vulnerable populations such as Black patients and the elderly may benefit from the inclusion of SDoH. In a recent study, the authors find that the inclusion of SDoH improves the C-statistics of ED readmission predictive models for specific subgroups including the elderly and the obese patients [41]. In a comprehensive study of the impact of individual and community-level SDoH data on common EHR prediction tasks in the intensive care unit, authors also find that adding SDoH to EHR does not generally improve the accuracy of classifiers when compared to using EHR data only [34]. However, the same authors note that including SDoH can lead to better-calibrated and fairer models in specific subgroups, with varying levels of contribution depending on the population and predictive task. For example, models trained on female patients with diabetes, incorporating these data improves model performance when combined with discharge notes or all EHR data. This is important as researchers are also aware of algorithmic bias in ML models [42].

# Methodology

We frame the problem statement as a classification to predict which patients will be readmitted; in building such a model, we can learn important features and whether some of these are uniquely related to the CHW program. Our dataset includes patients who were engaged with CHWs (interacted with via phone calls and in some cases have replied to the SDoH survey) and patients who were not engaged. Patients who are not engaged are patients whom CHWs have not been able to contact, or have yet to contact or patients who have opted out of the program. Therefore, for about half of the patients, we have mainly demographics data and referral information; while for the other half we have additional features including logs of interaction with CHWs and SDoH data. The dataset provides the opportunity to compare classification with demographics only and the classification for patients with all data available. We use a Random Forest (RF) Classifier due to its ensemble nature, which helps with limited size datasets, and its use of Decision Trees, which readily provide feature importance. We also use two more classifiers to predict readmissions for engaged patients, a Logistic Regression, which is the main classifier used in related work and a Neural Network in order to explore any possible increase in accuracy. Figure 1 summarizes the structure of our data and the need to study the predictive differences between engaged and non-engaged patients as well as the impact of additional data for engaged patients. Once we identify the best performing classifier, we focus on optimizing accuracy for the patients who interact with CHWs.

## Understanding the data

Here, we provide some insight into the type of data we leverage in our classification model.

### Leveraging CHW data

In addition to demographics and referral information collected before they take on a case, CHWs enter information and notes about their interactions with patients, up to ten contact attempts. We set out to clean, process, and transform this data into useful features for ML models. This is challenging as there are human factors involved. For example, patients do not always answer all questions in the screener. Sometimes this is because CHWs, using their prior knowledge and ability to “read the room”, may not ask all follow up questions. In addition, careful inferences need to be made in interpreting communications with CHWs to decode a “maybe” as a “yes” in certain cases and appropriately interpret a “refuse to answer” entry.

### Leveraging SDoH data

In our work, CHWs asked a series of questions from an extensive SDoH survey. Note that this is an attempt to collect rare and valuable individual-level SDoH data. Challenges here included the length of the questionnaire, which patients do not always want to/are able to complete, resulting in sparse data. Moreover, certain questions have multiple possible responses, some of which are not well represented in the data.

## Prediction Readmissions

We define a classification problem of high-risk patients who are readmitted to the ED. Since the goal of the project is to determine the extent to which CHWs help predict the rate of these readmissions, it is important to be able to use an explainable model, which will allow feature importance inspection. We hypothesize that one way to demonstrate the positive impact of CHW is to identify CHW-related features as important for the classification using a Random Forest (RF) classifier. Recognizing that the main model previously used in related works is Logistic Regression (LR), we also use this classifier on the engaged patients as well as a Neural Network classifier. Neural Networks typically require large amounts of data as they are more complex classifiers, however in some cases they have also been shown to achieve higher accuracy and are increasingly being leveraged in health informatics [5, 26].

## Experiment Design

Our initial dataset includes a set of patients, about half of which have engaged with CHWs. This potentially presents a challenge as half of the patients do not have values for CHW contacts nor SDoH data. On the other hand, it provides the opportunity to compare and contrast similar populations of patients as they were all classified as high risk and were thus referred to CHWs, but CHWs have not been able to connect to all. Therefore, we can compare and contrast readmission rate predictions excluding and including this data when available.

Our baseline data includes demographic and referral features for all patients (as many as possible). We then compare and contrast how the classifiers perform with data from patients who did not connect with CHWs (using only their baseline data) as opposed to the classifiers which leverage data from patients who did connect with CHWs including the additionally available CHW and SDoH data for these patients. Once we identify the best performing classifier, we focus on optimizing the performance for this later set of patients only and detail our analysis.

# Results

Following our methodology, we present results about the data preprocessing and the classification results. We start with describing the data in more detail.

## Data

We started this project with an anonymized dataset containing records for 1,634 patients and a dictionary of codes explaining questions or features for our purposes, and corresponding possible answers. Each patient record includes a total of 315 original features (characteristics) including log entries of time spent with CHWs and answers to the SDoH survey. The features were both categorical (e.g., whether the CHW “spoke to the patient” as a result of a contact attempt) and some continuous (e.g., how long they spoke on the phone during the first contact). There were also some dates in the data, such as date referred, and the date the case was closed. In summary, each patient record represents an instance in which a patient was referred to a CHW for follow up. One can see the initial dataset in three parts: 1) Referral and demographics data, which include age, race, gender, type of referral (e.g., high-risk), and common comorbidities (e.g., diabetes), 2) Contact logs containing information about a total of ten contact attempts and 3) SDoH questions and answers. Other important considerations include:

### Class Label

An important part of the data description is the class label that identifies whether a patient is readmitted within 30 days or not; this is the feature we are attempting to predict. This column is the day readmit variable which is 0 (false) or 1 (true). In terms of class labels distribution, 67.5% of the patients were not readmitted (vs. 32.5%). Amongst the patients who were engaged with a CHW, the proportions are slightly different with 21% of engaged patients being readmitted.

### Engaged Patients

We designed a new variable called ‘engaged’ to indicate that a patient was in contact with a

CHW, however minimally. This was defined as a patient who responded to a single contact attempt or answered at least one SDoH question.

### Recurring Patients

We note that a patient can have multiple records, because a patient may be referred to the CHW program more than once. Therefore, a natural pre-processing step involves identifying unique patients (and creating a feature accounting for recurring referrals). There were 1,381 unique patients.

## Data Cleaning and Preprocessing

The first data cleaning step is to check for missing values. The most common way to replace missing values is to use the mode (most common value) for categorical variables and the average for the continuous variables. Other preprocessing steps involved removing variables which are not relevant to the task, or do not contain enough information, and creating new features which aggregate or combine original variables.

### Recurring Patients

Whether a patient is new to the hospital is not as relevant as to whether they are new to

the CHW program. Therefore, we instead check the number of times a patient appears in the dataset and create a *NewPatientCount* variable to account for this. The intuition behind this variable is that perhaps recurring patients may be at a higher risk to be readmitted. We have 1,201 new patients (or 87.0%), 128 patients who were referred twice (9.3%) and 52 patients (3.7%) who were referred three times. We replace the initial NewPatient variable by our *NewPatientCount* variable. We note that while in our subset of data, the *NewPatientCount* is more useful, it is not perfect as some patients may be on their second visit but are in their first in our dataset, similarly there may be patients who would be recurring patients if the period of time was extended.

### Duration of Intervention

There are a few date fields: referral date, referral month. Rather than using dates, which are not as meaningful on their own or as readily usable along with other types of features, we use the differences between dates as features (most notably, we consider the time spent in days with a CHW, that is the difference between the date the case is closed and the date a patient was assigned to a CHW).

### Type of referral

This is an important feature as the analysis will show; this variable indicates the type of referral including High-risk readmit, Repeat return (to the ED), COVID, Behavioral health, ED, and Other.

### Demographics

We combined the original demographic categories due to data imbalance (e.g. 0.07% were American Indian, 0.07% were Native Hawaiian or other Pacific Islander, 0.2% were Asian); the remaining and new categories after filling in the mode for 8% of the data missing are Black (1) 65.0%, Latinx (9) 22% and Other (7) 13%. We keep and preprocess language, sex and age as features. After filling in missing values (4% of the data) with the mode, the distribution of the language feature is as follows: English (86.0%) vs. Not English (14.0%). About 45.5% of patients are female, while 54.5% are male. The mean age amongst the patients is 58.

### Insurance

For some variables such as insurance, which are important but include so many possibilities that some are very sparse, we combined the original values into fewer that retained the important meaning of these variables. In this particular case, original values are: 1 = Uninsured; 2 = Medicaid; 3 = Medicare; 7 = Medicare & Medicaid; 8 = Medicare & Private Insurance; 4 = Other public insurance; 5 = Private Insurance; 6 = Not listed or not sure, with some categories being underrepresented. Therefore, we combined them into new categories: 1 = Uninsured; 2 = Public; 3 = Private, and -1 = Not listed or NA. Note that 26% of this column was missing and needed to be replaced with the mode.

### Comorbidities

This new feature checks if a patient has at least one comorbidity from a combined list of diabetes, asthma, and hypertension. The intuition behind this variable is that the more comorbidities, the greater the risk of readmission. We found that 701 (50.8%) patients did not have any of these conditions, 456 (33.0%) had at least one, 203 (14.7%) had two and 21 patients (1.5%) had 3 of these conditions.

### Contact Log Summary

There are important logs in the data about contacts between the CHWs and the patients. Details about contacts (type of contact, length of contact, response from the patients etc.) are recorded for a total of ten visits. One straightforward and meaningful way to combine these logs is to sum them. Some of these contact features are more nuanced. For example, we can sum the number of attempts to contact a patient, but a high number of contact attempts may still result in no actual contact. Therefore, we carefully process the outcome of each contact to check if any outcome resulted in “spoke to patient” according to the dictionary of codes, which was essential in making data preprocessing decisions. By the end of this process, we have some continuous and categorical variables summarized (continuous: e.g., summed minutes spent on the phone or categorical: e.g., outcome is marked as “spoke to patient” as a combined outcome for ten contacts).

### Engaged

This new variable, which indicates that a patient had contact with a CHW, can be used to fill missing values for the SDoH questions. For example, we know that patients who were not engaged were never asked the questions; therefore, for the SDoH missing answers, we can look at the mode or most common answer amongst only the engaged patients, to infer answers for other patients when appropriate. We find that 45.8% were engaged in our dataset.

### SDoH Features

We note that some of the answers to the SDoH questions were very sparse and it is important to understand that a CHW may talk to a patient, but a patient may refuse to answer the survey questions or a CHW may not ask all questions if the patient is in a hurry, weak, etc. Either way, from this point on, we attempt to infer the missing values as much as possible using the mode of the “Engaged” patients and clearly attribute Non-Applicable (NA) for patients who are not engaged with one distinction. We replace the “refused to answer” with NA as well, as if these patients were never asked this question. In collaboration with domain experts, we also carefully infer a “yes” for “unsure” in the case of sensitive questions such as “do you have housing insecurity?” The rationale behind this decision is that we deem it reasonable to treat someone as if they have housing insecurity if they are unsure. Some aggregation examples include several questions about food insecurities that were aggregated to detect any food insecurities. As a result of this particular aggregation,

we found that 26.0% of the patients reported having food insecurity. Several SDoH variables were removed as they were mostly empty. For example, 2% of patients answered the question and responded yes around needing additional health education. We note some redundancy in questions about comorbidities (asthma, diabetes etc.) and insurance for example.

### Correlation Analysis

At this point we expect correlations. Indeed, since for many categorical variables, the NA category after transforming all multi-valued features into dummy variables, simply reflects patients who were never engaged by CHWs, such column will be similar for most of the CHW features as well as the SDoH features. We use a correlation analysis to drop redundant features. In summary, after preprocessing the data, we have 22 demographics, referral, and other point-based features (e.g., asthma, diabetes) for all patients, we have engaged patients with contact log information (12 features) and additional information about which patients answered a series of SDoH questions (36) for a total of 70 features. Finally, we normalize all continuous variables using min-max normalization, i.e., the maximum value is mapped to a 1, while the minimum value is mapped to a 0.

## Predicting Readmissions

At this point we are ready to predict readmission using the remaining features. We also seek to get the feature importance in predicting readmissions. We summarize the three classifiers performances (Rf, LR and NN) and discuss the feature importance for the RF for engaged vs. not- engaged patients. We show the classifiers’ accuracies, specificity sensitivity and Area Under the receiver operating characteristic Curve (AUC). The accuracy is the correctness of the classifier in classifying engaged vs not-engaged, however in some cases accuracy is not enough to determine if a classifier is learning as even a rule classifying every patient as readmitted will correctly classify the readmitted patients while misclassifying every remaining patient. Therefore we also look at sensitivity (true positive rate) and specificity (true negative rate), each indicating the extent to which the classifier is correctly identifying each patient (readmitted vs. not readmitted). Finally, the AUC provides an aggregate measure of performance across all possible classification thresholds. The AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative. It is a measure of the discrimination capability of the classifiers. The AUC is the most common measure of discrimination for classifiers. Its value varies from 0 to 1, the closer to 1, the better the classifier. Results for the classification using baseline data (i.e. demographics and referral information), and with additional data for engaged patients are shown in Table 1.

The main result is that the classifiers are not effectively learning without CHW/SDoH data as the AUCs for all classifiers are under the no-skill value of 0.5. The next interesting result is that the NN achieves the highest accuracy and AUC but displays the lowest sensitivity for engaged patients. The RF on the other hand achieves a higher sensitivity with respect to both the NN and the LR. Importantly, towards our goal of understanding the data, the RF provides feature importance based on their ability to discriminate between engaged and not-engaged patients. While further exploring alternate methods for feature importance for neural networks such as visualization, we select the RF for further in-depth analysis of the engaged patients.

Table 1.  Classifier accuracies for *All*, *Engaged* and *Engaged* patients only.

| Classifier (data)a | Accuracy | Specificity | Sensitivity | AUC |
| --- | --- | --- | --- | --- |
| RF (baseline features - not engaged patients) | 0.668 +/- 0.074 | 0.779 +/- 0.106 | 0.454 +/- 0.137 | 0.454 |
| RF (all features - engaged patients) | **0.798 +/- 0.063** | 0.920 +/- 0.071 | **0.362 +/- 0.132** | **0.641** |
| LR (baseline features - not engaged patients) | 0.710 +/- 0.029 | 0.957 +/- 0.031 | 0.236 +/- 0.097 | 0.236 |
| LR (all features - engaged patients) | 0.792 +/- 0.058 | 0.936 +/- 0.061 | 0.276 +/- 0.156 | 0.606 |
| NN (baseline features - not engaged patients) | 0.708 +/- 0.051 | 0.981+/-  0.022 | 0.182 +/- 0.154 | 0.182 |
| NN (all features - engaged patients) | **0.811 +/- 0.033** | **0.990 +/- 0.017** | 0.171 +/- 0.154 | **0.760** |
| *Note*: Every experiment is run three times to provide a confidence interval on the results. Individual measurements are measured using 10-fold cross-validation. | | | | |  |
|  |  |  |  |  |  |  |
|  | | |  |  |  |  |

Experiment 5: Best classifier with important features (15 most important features including some demographics and CHWs)

Experiment 6: Vary threshold and find best recall/precision combo

Table 3: showing Differences between engaged and non-engaged:

### Random Forest Classification – Baseline Features

Here we use only the 22 referral and demographics features for the classification. After tuning and using cross-validation to select the best parameters, we achieve a testing accuracy of 74.4±2.6%. The fact that the accuracy is slightly down seems to point towards the fact that adding CHW questions improves the model’s discrimination capability slightly. More interestingly, when looking at the feature importance shown on Figure 2, age appears as the dominant feature once again. Comparatively, however, all other features are significantly less important almost by a factor of 10 with the next features being the type of referral, gender, new patient and number of comorbidities.

### Random Forest Classification – All data

We train and tune our RF using ten-fold cross-validation using 80% of the data for training and report a testing accuracy to 75.5±3.7%. The most important/useful feature used by the model to predict readmission is age (See Figure 1 for the top 15). Given the initial population, (generally older), this result is not necessarily surprising. However, it is important and encouraging that the next most important variable has to do with the CHWs, i.e., total duration of “engagement” with a CHW even when about half of the patients are yet to be “engaged”. Then comes the type of referral and again the total time spent talking to a CHW. The main insight from this result is that CHWs are important in predicting the 30-day readmission, showing up in 4 of the first 12 important features. One perhaps more initially disappointing finding is that the first SDoH question shows up only in 26th position with noticeably less importance.

These results are interesting and indicate that contact with CHWs is important in predicting readmission, therefore we setup follow-up experiments to confirm and further assess the important of CHWs: 1) comparing the performance of the classifier with and without CHW data, 2) classifying engaged patients only, and 3) characterizing this importance of CHW in the prediction of 30-day readmissions.

### Random Forest Classification – Engaged Patients Only

In order to dig deeper into these results, we then look at predictions within the Engaged group only. Ultimately, the goal is for all patients to be engaged, therefore we investigate predicting the rate of 30-day readmissions amongst engaged patients only. Indeed, after tuning, prediction of 30-day readmission amongst engaged patients only increases by 5 points to 79.8±4.8%. Interestingly, age is dislodged from the most important feature position to the 3rd (see Figure 3) with duration of engagement with CHW becoming the most important feature, followed by whether the patient is recurring, and the total time spent in minutes talking to a CHW. This is a significant result as it seems to indicate that engaging the patient leads to an increase in the ability to predict readmissions (see characterization next) by five percent. Also importantly, as we focused on engaged patients, we can dig deeper into important SDoH features: race shows up in 18th position, alcohol use in 21st and food insecurity in 24th.

## Characterization of Feature Importance

Look at features importance figures from Experiments 1-3 and discuss that instead. Now we look at the importance for the model using all of the data in more detail.

## Optimize Prediction for Engaged Patients

Table 2.  Classifier accuracies for *Engaged* patients only with and without CHW and SDoH data.

| Classifier (data) | Accuracy | Specificity | Sensitivity | AUC |
| --- | --- | --- | --- | --- |
| RF (demographics + referral) | 0.794 +/- 0.041 | 0.963 +/-  0.033 | 0.190 +/-0.154 | 0.577 |
| RF (demographic/referral + CHW logs) | 0.797 +/- 0.066 | 0.935 +/-  0.052 | 0.305 +/-0.187 | 0.620 |
| RF (demographics/referral + CHW logs + SDoH data) | **0.798 +/- 0.063** | 0.920 +/- 0.071 | **0.362 +/-0.132** | **0.641** |
| RF (using 15 most important features) |  |  |  |  |
| *Note*: Table notes. aAdd 10-fold cross-validation accuracy. bExplain recall.  cHighlight statistically significant results. |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  | | |  |  |  |  |

## Differences between Engaged and Not-Engaged Patients

Table of statistical differences between patients?

Discussion of implicit bias?

## Summary of Findings, Recommendations and Discussion

This systematic analysis has provided evidence of the increase in predictive capabilities of classifiers when patients are engaged with the CHWs. **A major and unexpected takeaway from our analysis is that feature importance points directly to the human connection between CHWs and patients** rather than specific SDoH questions and answers (duration of “engagement”, number of contact attempts and time spent finding resources). Furthermore, adding CHW data improves the prediction of 30-day readmissions to the ED by 5% over the baseline using only referral and demographic data.

### Recommendations

The more-in-depth importance characterization points to several recommendations:

* Patients between the age of 50-59 are an important higher-risk population and so are patients over the age of 70. Risks can be mitigated for other patients.
* Repeated U-shaped results during characterization of the CHWs impact points to diminishing returns beyond a certain amount of contact and should raise flags that a patient may be in need of a different intervention (e.g., medical vs social).
* Initial referral information is useful to predicting readmissions. More information should be gathered from social workers on how they determine “high-risk.” Perhaps all referral sources should administer the simple point-base screener.
* Aggregating comorbidities may be a new automatically derived variable, and patients with more than one could automatically be considered high-risk.
* Similarly, recurring patients are particularly high-risk and necessitate a more focused intervention and outreach strategy.
* **More resources should be assigned to CHWs, as focusing on engaged patients significantly affects predictive capabilities.**
* As a follow up, when looking at important features for engaged patients only, SDoH impact factors start to emerge. Here we recommend collecting more data.

### Increasing Recall

CHWs may opt to “flag” more patients as high-risk, at the risk of over-diagnosing other patients. In other words, with this model and more CHWs, we can prioritize recall over precision in an effort to “retrieve” and treat more high-risk patients at the cost of talking to some patients who are less likely to be readmitted. On the precision-recall curve shown in Figure 7, we detect the inflection point (maximum F—1 score). Compared to the previous 50% of probability (typical for a 2-class problem), if we lower the threshold required to classify a patient as “readmitted” to 30.0%, we can achieve a recall of 72% at the cost of the low precision of 45%.

### Adding CHW/SDoH to EHRs

Some hospitals already use discharge tools such as the Re-Engineered Discharge (RED) toolkit, a nationwide discharge program that has been established to be effective against all 30-day readmissions [43]. Applying RED involved following a discharge plan including, educating patients about diagnosis and medicine for example. However, the successful implementation of RED requires a comprehensive framework and participation from all stakeholders [43]. A recent study of HRRP as another potential solution to the ED readmission problem presents its successes and challenges [44]. Some of the highlighted limitations and corresponding solutions involve adjusting for socioeconomic status. Some researchers further proposed adding the development of community approaches to healthcare, recognizing that social factors, patient compliance and access to care are barriers to reducing ED rates [45].

Our work complements the existing literature on 30-day patient readmissions, in that it presents evidence-informed strategies for addressing the impact of SDoH on readmissions. Some patients with important needs are willing and able to connect to CHWs; the readmission rate amongst them is more predictable and presents a path forward for further investigation and strategies to decrease readmissions. Notably, CHW-related features are important in the prediction. Our results also align with current related work in that we find that the addition of SDoH data does not yet generally significantly increase the predictive accuracy of classifiers when compared to a baseline, in part due to scarce data. It is important to note that our relative success when compared to other studies which only found improvement of integrating SDoH in prediction for subgroups is due to the fact that our patients are somewhat already a subgroup (having all been referred to CHWs as higher-risk). We note that an important limitation to this work is that we do not have EHR data, however with Sinai’s recent transition to EPIC, patient EHR data may have more potential to be utilized in work such as this.

Therefore, one main recommendation, knowing that not all readmissions will be prevented, is to leverage more information from EHRs (contain clinical information about patients, such as medical history, vital signs, laboratory data, immunizations, and medications). We know for example that comorbidities are important, therefore more EHR data may yield similar impact on classification. Conversely, since we show that CHWs have an impact on some readmissions due to social rather than medical intervention, combining the two elements would be beneficial and constitute a comprehensive effort to prevent any type—medical or social—of avoidable readmissions.

# Conclusion

The systematic analysis presented in this work has provided evidence of the positive impact of the Community Health Worker (CHW) program on an important health outcome: the 30-day Emergency Department (ED) readmission rate. **A major and unexpected takeaway from our analysis is that feature importance pointed directly to the human connection between CHWs and patients rather than specific SDoH questions and answers.** While this may be due in part to data scarcity, important features in classification of 30-day readmissions currently include duration of engagement with CHWs, number of contact attempts and time spent finding resources. Furthermore, focusing on engaged patients improves the prediction of 30-day readmissions to the ED by 5% over the baseline using only referral and demographics data to 79.8±4.8%. Our work provides characterization of feature importance and results in a series of findings and recommendations for the CHW program in the form of red flags for patients who may be in need of more drastic or different types of intervention. Finally, having demonstrated the positive impact of CHWs on ED readmission, our work points to several exciting future investigation avenues, including early flagging of high-risk patients to prevent readmissions and the integration of SDoH in EMR data for more complete analysis of how social factors impact health outcomes in a broader range of patient types.

# Figures

* Figures 1: Diagram of methodology/experiment design
* Figures 2, 3, 4: Feature importance for RF for each experiment
* Figure 5 of precision-recall curve with most important features

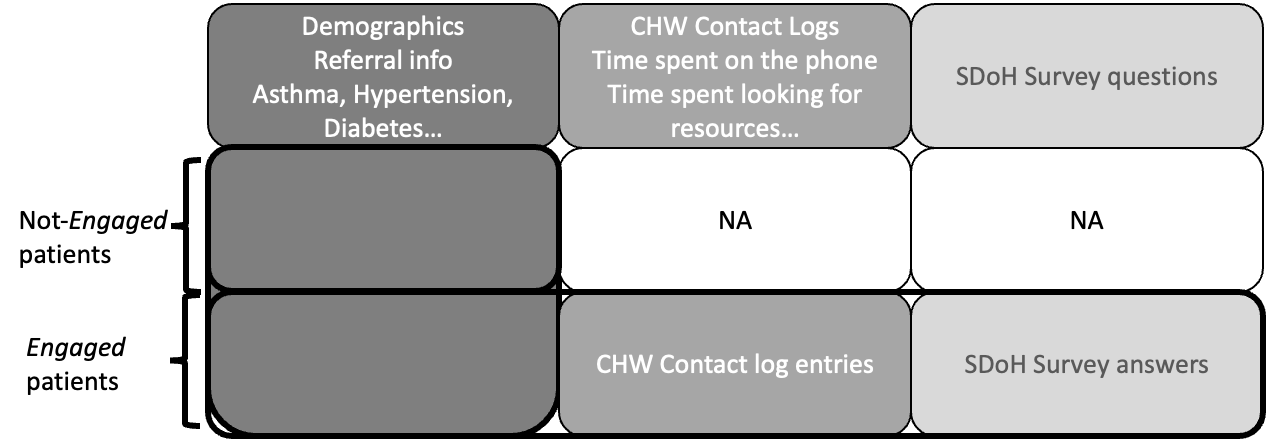


Fig. 1.  Structure of patient data including basic demographics and referral information, CHW contact logs data and SDoH survey answers when available (for engaged patients only).

# Acknowledgments

Nuvia Hernandez and Nicole Baugh were supported by the NSF Medical Informatics (MedIX) REU Program Award No 1950894.

Sinai Urban Health Institute activities were supported by funding from:

Lloyd A. Fry Foundation

Google Health Equity Research Initiative

Sinai Chicago

SUHI Equity Research Fellowship Program

As well as the generous support and partnership with Sinai Chicago’s Social Work Department, and the entire SUHI Project team: Melinda Banks, René Bucio, Aaron Chestnut, Jannette Guzman, Stacy Ignoffo, Stephanie Jordan, Anubhav Kidambi, Alyce Roberson, Gloria Seals, Maria Guadalupe Villa, and Madeline Woodberry.

# References

1. B. Robson and R. Harris, “Hauora: M`aori standards of health iv. a study of the years 2000–2005,” Wellington: Te Ropu Rangahau Hauora a Eru Pomare, 2007.
2. D. Satcher, “Include a social determinants of health approach to reduce health inequities,” Public Health Reports, vol. 125, no. 4 suppl, pp. 6–7, 2010.
3. R. J. Lavizzo-Mourey, R. E. Besser, and D. R. Williams, “Understanding and mitigating health inequities—past, current, and future directions,” New England Journal of Medicine, vol. 384, no. 18, pp. 1681–1684, 2021.
4. R. Fang, S. Pouyanfar, Y. Yang, S.-C. Chen, and S. Iyengar, “Computational health informatics in the big data age: a survey,” ACM Computing Surveys (CSUR), vol. 49, no. 1, pp. 1–36, 2016.
5. S. Srivastava, S. Soman, A. Rai, and P. K. Srivastava, “Deep learning for health informatics: Recent trends and future directions,” in 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2017, pp. 1665–1670.
6. S. S. Oh, J. Galanter, N. Thakur, M. Pino-Yanes, N. E. Barcelo, M. J. White, D. M. de Bruin, R. M. Greenblatt, K. Bibbins-Domingo, A. H. Wu et al., “Diversity in clinical and biomedical research: a promise yet to be fulfilled,” PLoS medicine, vol. 12, no. 12, p. e1001918, 2015.
7. I. Y. Chen, P. Szolovits, and M. Ghassemi, “Can ai help reduce disparities in general medical and mental health care?” AMA journal of ethics, vol. 21, no. 2, pp. 167–179, 2019.
8. N. Priest and D. R. Williams, “Racial discrimination and racial disparities in health.” 2018.
9. J. Morita, Unequal cities: structural racism and the death gap in America’s Largest Cities. JHU Press, 2021.
10. A. P. H. Association et al., “Community health workers. american public health association website,” 2016.
11. A. Wennerstrom, C. G. Haywood, D. O. Smith, D. Jindal, C. Rush, and G. W. Wilkinson, “What are the roles of community health workers in medicaid managed care?
12. D. R. Williams, M. Costa, J. P. Leavell et al., “Race and mental health: Patterns and challenges,” A handbook for the study of mental health: Social contexts, theories, and systems, pp. 268–290, 2010.
13. B. Hollister and V. L. Bonham, “Should electronic health record-derived social and behavioral data be used in precision medicine research?” AMA Journal of Ethics, vol. 20, no. 9, pp. 873–880, 2018.
14. M. Gutierrez Kapheim, J. Ramsay, T. Schwindt, B. R. Hunt, & H. Margellos-Anast. Utilizing the Community Health Worker Model to communicate strategies for asthma self-management and self-advocacy among public housing residents. Journal of Communication in Healthcare, 8(2), 95-105, 2015.
15. H. Margellos-Anast, M. A. Gutierrez, & S. Whitman, S. Improving asthma management among African-American children via a community health worker model: findings from a Chicago-based pilot intervention. Journal of Asthma, 49(4), 380-389, 2012
16. B. R. Hunt, K. L. Allgood, J. M. Kanoon, & M. R. Benjamins. Keys to the successful implementation of community-based outreach and navigation: lessons from a breast health navigation program. Journal of Cancer Education, 32, 175-182, 2017.
17. K. L. Allgood, G. H. Rauscher, S. Whitman, G. Vasquez-Jones, & A. M. Shah. Validating self-reported mammography use in vulnerable communities: findings and recommendations. Cancer epidemiology, biomarkers & prevention, 23(8), 1649-1658, 2014.
18. S. Ignoffo, H. Margellos-Anast, M. Banks, R. Morris, & K. Jay. Clinical integration of Community Health Workers to reduce Health Inequities in Overburdened and under-resourced populations. Population Health Management, 25(2), 280-283, 2022.
19. T. W. Markossian, K. McCab, J. Ginn, Y. Galvan, M. Banks, & S. Ignoffo. (2022) Integrating Community Health Workers into Hospital Systems Through a Social Work Partnership: A Report from the Field. Journal of Health Care for the Poor and Underserved, 34(1), 478-495.
20. T. W. Markossian, K. McCab, J. Ginn, Y. Galvan, M. Banks, & S. Ignoffo. (2023). Integrating Community Health Workers into Hospital Systems Through a Social Work Partnership: A Report from the Field. Journal of Health Care for the Poor and Underserved, 34(1), 478-495.
21. N. Goldfield. (2010). Strategies to decrease the rate of preventable readmission to hospital. CMAJ, 182(6), 538-539.
22. J. D. Donzé, M. V. Williams, E. J. Robinson, E. Zimlichman, D. Aujesky, E. E. Vasilevskis, S. Kripalani, J. P. Metlay, T. Wallington, G. S. Fletcher, and A. D. Auerbach, 2016. International validity of the HOSPITAL score to predict 30-day potentially avoidable hospital readmissions. JAMA internal medicine, 176(4), pp.496-502.
23. Hospital Readmission Reduction Program. Centers for Medicare and MedicaidServices (CMS). (2020). https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/HRRP/Hospital-Readmission-Reduction-Program
24. R. E. Burke, J. L. Schnipper, M.V. Williams, E. J. Robinson, E. E. Vasilevskis, S. Kripalani, J. P. Metlay, G. S. Fletcher, A. D. Auerbach and J. D. Donzé, 2017. The HOSPITAL score predicts potentially preventable 30-day readmissions in conditions targeted by the hospital readmissions reduction program. Medical care, 55(3), p.285.
25. Robinson, R., & Hudali, T. (2017). The HOSPITAL score and LACE index as predictors of 30 day readmission in a retrospective study at a university-affiliated community hospital. PeerJ, 5, e3137.
26. K. Teo, C. W. Yong, J. H. Chuah, Y. C. Hum, Y. K. Tee, K. Xia, & K. W. Lai. (2023). Current trends in readmission prediction: an overview of approaches. Arabian journal for science and engineering, 48(8), 11117-11134.
27. C. for Disease Control and Prevention, “About social determinants of health (sdoh),” 2023.
28. J. M. McGinnis, P. Williams-Russo, and J. R. Knickman, “The case for more active policy attention to health promotion,” Health affairs, vol. 21, no. 2, pp. 78–93, 2002.
29. I. of Medicine, Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2. Washington, DC: The National Academies Press, 2014. [Online]. Available:

https://nap.nationalacademies.org/catalog/18951/capturing-socialand-

behavioral-domains-and-measures-in-electronic-health-records

1. B. Galobardes, G. D. Smith, and J. W. Lynch. 2006. Systematic Review of the Influence of Childhood Socioeconomic Circumstances on Risk for Cardiovascular Disease in Adulthood. Annals of Epidemiology 16, 2 (Feb. 2006), 91–104.
2. R. J. Walker, M. Gebregziabher, B. Martin-Harris, and L. E. Egede. 2014. Relationship between social determinants of health and processes and outcomes in adults with type 2 diabetes: validation of a conceptual framework. BMC Endocrine Disorders 14, 1 (Oct. 2014), 82.
3. F. Amrollahi, S. P. Shashikumar, A. Meier, L. Ohno-Machado, Shamim Nemati, and Gabriel Wardi. 2022. Inclusion of social determinants of health improves sepsis readmission prediction models. Journal of the American Medical Informatics Association 29, 7 (July 2022), 1263–1270.
4. J. Meddings, H. Reichert, S. N. Smith, T. J. Iwashyna, K. M. Langa, T. P. Hofer, and L. F. McMahon. 2017. The Impact of Disability and Social Determinants of Health on Condition-Specific Readmissions beyond Medicare Risk Adjustments: A Cohort Study. Journal of General Internal Medicine 32, 1 (Jan. 2017), 71–80.
5. M. Y. Yang, G. H. Kwak, T. Pollard, L. A. Celi, & M. Ghassemi. (2023, August). Evaluating the Impact of Social Determinants on Health Prediction in the Intensive Care Unit. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (pp. 333-350).
6. M. N. Cantor, & L. Thorpe. (2018). Integrating data on social determinants of health into electronic health records. Health Affairs, 37(4), 585-590.
7. D. McGraw. (2015). Privacy concerns related to inclusion of social and behavioral determinants of health in electronic health records. In Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2. National Academies Press (US).
8. A. S. Navathe, F. Zhong, V. J. Lei, F. Y. Chang, M. Sordo, M. Topaz, S. B. Navathe, R. A. Rocha, and L. Zhou, 2018. Hospital readmission and social risk factors identified from physician notes. Health services research, 53(2), pp.1110-1136.
9. M. Jamei, A. Nisnevich, E. Wetchler, S. Sudat, & E. Liu. (2017). Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. PloS one, 12(7), e0181173.
10. J. R. Vest, & O. Ben-Assuli. (2019). Prediction of emergency department revisits using area-level social determinants of health measures and health information exchange information. International journal of medical informatics, 129, 205-210.
11. A. Belouali, H. Bai, K. Raja, S. Liu, X. Ding, & H. Kharrazi. (2022). Impact of social determinants of health on improving the LACE index for 30-day unplanned readmission prediction. JAMIA open, 5(2), ooac046.
12. Y. Zhang, Y. Zhang, E. Sholle, S. Abedian, M. Sharko, M. R. Turchioe, Y. Wu, and J. S. Ancker, 2020. Assessing the impact of social determinants of health on predictive models for potentially avoidable 30-day readmission or death. PLoS One, 15(6), p.e0235064.
13. M. H. Chin, M. Afsar-Manesh, A. S. Bierman, C. Chang, C. J. Colón-Rodríguez, P. Dullabh, D. G. Duran, M. Fair, T. Hernandez-Boussard, M. Hightower, and A. Jain, 2023. Guiding Principles to Address the Impact of Algorithm Bias on Racial and Ethnic Disparities in Health and Health Care. JAMA Network Open, 6(12), pp.e2345050-e2345050.
14. S.E. Mitchell, J. Martin, S. Holmes, C. van Deusen Lukas, R. Cancino, M. Paasche-Orlow, C. Brach, and B. Jack, 2016. How hospitals reengineer their discharge processes to reduce readmissions. Journal for healthcare quality: official publication of the National Association for Healthcare Quality, 38(2), p.116.
15. E.G. Ferro, E.A. Secemsky, R.K. Wadhera, E. Choi, J.B. Strom, J.H. Wasfy, Y. Wang, C. Shen, and R.W. Yeh, 2019. Patient readmission rates for all insurance types after implementation of the hospital readmissions reduction program. Health Affairs, 38(4), pp.585-593.
16. S. J. Warchol, J. P. Monestime, R. W. Mayer, & W. W. Chien (2019). Strategies to reduce hospital readmission rates in a non-Medicaid-expansion state. Perspectives in health information management, 16(Summer).