Leveraging Community Health Workers and Social Determinants of Health for Predicting Emergency Department Readmissions

1st Kate Karam
School of Computing
DePaul University
Chicago, IL, USA
kburns10@depaul.edu

2nd Kelly MacCabe
Sinai Urban Health Institute
Sinai Health Systems
Chicago, IL, USA
kelly.mccabe@sinai.org

3rd Roselyne Tchoua
School of Computing
DePaul University
Chicago, IL, USA
rtchoua@depaul.edu

2023 Fifth International Conference on Transdisciplinary AI (TransAI) | 979-8-3503-5801-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/TRANSAI60598.2023.00014

***Abstract*—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 in decreasing the 30-day unplanned hospital ED readmissions at Sinai Health Systems. Using data from the Sinai Urban Health Institute, we compare predicting the readmissions of patients with and without data pertaining to Social Determinants of Health (SDoH), characterize the improvement in predictions and discuss lessons learned in the process. We show that when patients are simply *engaged* by CHWs, regardless of the content of those conversations, we can increase the predictive accuracy of our classifier by 5%. We use this result to make recommendations for improving patient care and discuss limitations and future work. Importantly our work points directly to the human connection between patients and CHWs as an important feature in the readmission rate.**

***Index Terms*—health informatics, data science, heath equity, emergency department readmissions, social determinants of health, community health workers, random forest**

I. Introduction

As more healthcare researchers and professionals study the impact of social factors on health outcomes, in an effort to decrease health disparities [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 disparities in health-related outcomes [8]. At Sinai Chicago (Sinai), Illinois' largest private safety net health care system, the Sinai Urban Health Institute (SUHI) was an early adopter of the Community Health Worker (CHW) model. CHWs form a liaison between the health care system and the community in an effort to address health inequities. CHWs are trained to identify and address barriers related to Social Determinants of Health (SDoH) [9], [10]. The 30-day readmission rate to the Emergency Department (ED) 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 logs and SDoH surveys, along with corresponding 30-day readmission patient data, provided us a unique opportunity to quantify the impact of the CHW program on an important health outcome.

The significance of this work is to highlight and document the positive impact of this program, which may be adopted in more Sinai clinics and in clinics at other hospitals. Our work points directly to the benefit of the CHWs: indeed, we show a 5% increase in predictive ability of our model for patients who are simply engaged with CHWs, regardless of the content of those interactions.

The contributions of this evidential work are: 1) a systematic comparison of predicting patient 30-day readmission with and without CHW data, including feature importance analysis, and 2) a summary of lessons learned along with a discussion of our findings and recommendations. The novelty resides in leveraging CHW logs and SDoH data to improve the prediction and prevention of 30-day readmission by incorporating social factors. Other predictive tools in common use for risk stratification and readmission reduction, such as the LACE[1] index as risk stratification as a readmission reduction strategy [11], which do not typically consider how social factors impact 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 outlines our methodology; Section 4 presents our results followed by a discussion in Section 5. We conclude in Sections 6.

[1]Lace is an acronym: L=Length of stay, A=Acuity of admission, C=Comorbidities, E=Emergency visits within the past 6 months

979-8-3503-5801-8/23/$31.00 ©2023 IEEE
DOI 10.1109/TransAI60598.2023.00014

II. Background

Research shows that inequalities in social conditions are fundamental causes of population health differences [8], [12], [13]. SUHI was an early adopter of the CHW model to address these disparities; the institute has been conducting community-engaged research aimed at understanding and addressing health disparities for over 20 years. Over 70% of Sinai patients are people of color from underinvested and overburdened West and Southwest Side neighborhoods in Chicago. These patients suffer disproportionately from a variety of maladies including diabetes, cardiovascular disease, gunshot wounds, violent crime, trauma, kidney disease, and breast cancer [14]. Importantly, among other endeavors, preliminary statistical reports have demonstrated that patients who are engaged with CHW have a 14% lower risk of 30-day readmissions. However, 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 readmission rate.

The current process for data collection and intervention starts with a patient referral from a social worker or other hospital staff. They refer patients based on a short, point-based screener that aims to detect *high-risk* patients. Once a CHW takes on a case, their goal is to communicate with the patient and provide resources as needed. Part of this process involves collecting contact logs and answers to a SDoH survey, which also helps detect healthcare needs and describe associated resources (e.g., questions about food insecurity are followed with options for adequate resources).

Therefore, one way to measure the positive impact of CHWs is equivalent to first leveraging this additional data in our analysis of ED readmissions, second to use the data analysis to prevent readmissions and improve the program. Such outcome would fulfill the promise of ensuring that "machine learning is fair, not only on ethical grounds but also on strong operational and business grounds" [15].

III. Methodology

We frame the problem statement as a classification to predict which patients will be readmitted; in building such model, we can learn important features and whether some of these are related to the CHW program. We use a Random Forest (RF) Classifier due to its ensemble nature, which helps with limited datasets, and its use of Decision Trees, which readily provide feature importance.

A. Understanding the data

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

1) Leveraging CHW data: Defined as "a frontline public health worker who is a trusted member of and/or has an unusually close understanding of the community served" by the American Public Health Association , CHWs deliver health education, conduct home assessments, and assist patients in navigating the complex health care system. In addition to demographics and referral information, collected before they take on a case, CHWs enter information and notes about their interactions with patients, over 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.

2) Leveraging SDoH data: SDoH contribute extensively to health inequities, which are "the unfair and avoidable differences in health status" seen within and between groups [16] and can be difficult to study [17]. Recognizing this, the National Academy of Medicine (NAM) recommended the inclusion of social environment measures in Electronic Medical Records (EMR)s to reduce bias [18]. In our work, CHWs asked a series of questions from an extensive SDoH survey. Challenges here included the length of the questionnaire, which patients do not always want to, or are not always able to participate in, resulting in sparse data. Moreover, certain questions take multiple possibilities, some of which are not well represented in the data.

B. Predicting readmissions

As previously mentioned, 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 reduce 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 RF classification.

C. 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 referred to CHWs, but CHWs simply have yet to reach out to them. Therefore, we can compare and contrast readmission rate predictions excluding and including this data when available. We first build a classifier using all the data available and compare it with a baseline model using only baseline features including referrals information and demographics. Finally we narrow in only on patients for whom we have CHW logs and SDoH data. For each experiment, we explore and compare feature importance.

IV. Results

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

A. Data

We start 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 by a social worker 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 conditions (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:

1) 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.

2) 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.

3) Recurring patients: We note that a patient can have multiple record, 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.

B. 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.

1) 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 at their second visit but are in their first in our dataset, similarly there maybe patients who would be recurring patients if the period of time was extended.

2) Duration of Intervention: There are a few date fields: *sw_date*, *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).

3) 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.

4) Demographics : We combined the original demographic categories due to data imbalance (just 0.07% were American Indian, 0.07% were Native Hawaiian or other Pacific Islander, 0.2% were Asian for example); the remaining and new categories after filling in the mode for 8% of the data missing are Black (1) 65.0%, Latino (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 women, while 54.5% are male. The mean age amongst the patients is 58.

5) 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.

6) Co-morbidity and co-morbidity count: 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 co-morbidities, 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.

7) 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).

8) 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.

9) 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 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. **These maybe omitted from the survey in the future.**

10) Correlation analysis to eliminate redundancies: 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 features 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.

C. 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. Recall that our hypothesis is that the CHW program (and the SDoH questions) will appear as important features during this process, highlighting their positive impact on the 30-day readmission rate. We separate the analysis in 1) the classification and identification of important features, and 2) the characterization of the importance of CHWs.

1) Random Forest Classification: We use a Random Forest (RF) Classifier. RFs are ideal in cases when there are many features without a priori knowledge on which might be more important. Using this ensemble of decision tree also helps with overfitting, making RF a ubiquitous, generally robust classifier. We train and tune our RF using ten-fold cross-validation using 80% of the data for training and report a testing accuracy to $75.5 \pm 3.7\%$.

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.

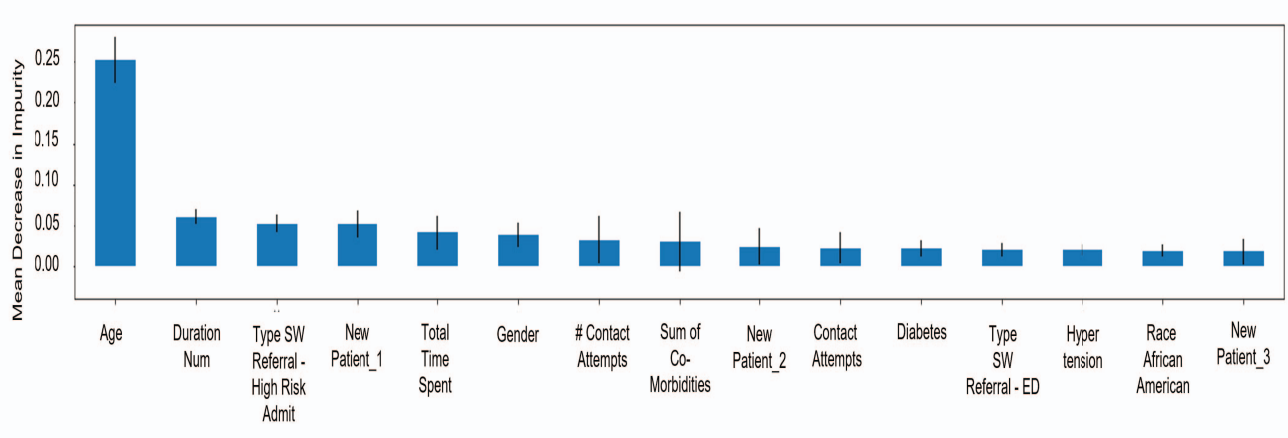

Fig. 1. Feature importance using all the data (patients with and without CHW and SDoH data) showing the 15 most important features in the RF classification.

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.

2) Random Forest Classification —Baseline referral and demographics: Here we remove all CHW and SDoH features and are left with 22 referral and demographics features and repeat the RF 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.

3) 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 abuse in 21st and food insecurity in 24th.

D. Characterization of importance

Now we look at the importance for the model using all of the data in more detail. We characterize this importance by looking at readmission rates in the test set to further investigate the patterns learned by the classifier and disaggregating the rate per values of the features. The testing data had a readmission rate of 10.8% for the 20% of total data used for testing our tuned model (the model was tuned using cross-validation and 80% of the data for training).

1) Age: This is one of the most important features, especially without CHW/SDoH features. When focusing on age, patients under 50 were generally not readmitted. Interestingly, there is a spike of predicted risk in patients in the 50-59 years old range (recall the average is 58 years old). The rate falls to 7% for patients between 60 and 69, 6% for those between 70-79, then back up to 8% for patients older than 80%. The main takeaway is that the rate is high for the "average" patient (somewhat expectedly), low below this number, however it increases for patients 70 and older.

2) Duration: This was the second most important feature. The average duration for a patient was 56 days. When there was no time spent with a patient, the readmission rate was 12.6%, higher than the average of 10.5%. This is a good indicator that no intervention is associated with higher readmission. As a patient became engaged and the duration of time was between 1—90 days, the readmission rate was 0%. This also seems to highlight that CHW interventions help reduce readmissions. After 90+ days of working with a patient, the readmission rate went up high to 42.9% (see Figure 4). This may seem counterintuitive since we are

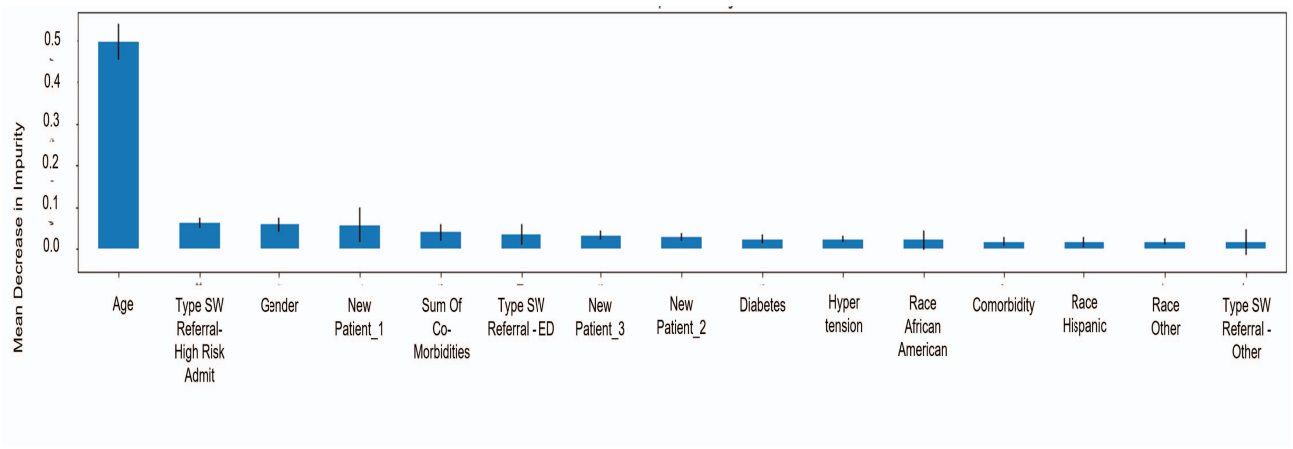

Fig. 2. Feature importance at baseline without CHW and SDoH data) showing the 15 most important features in the RF classification.

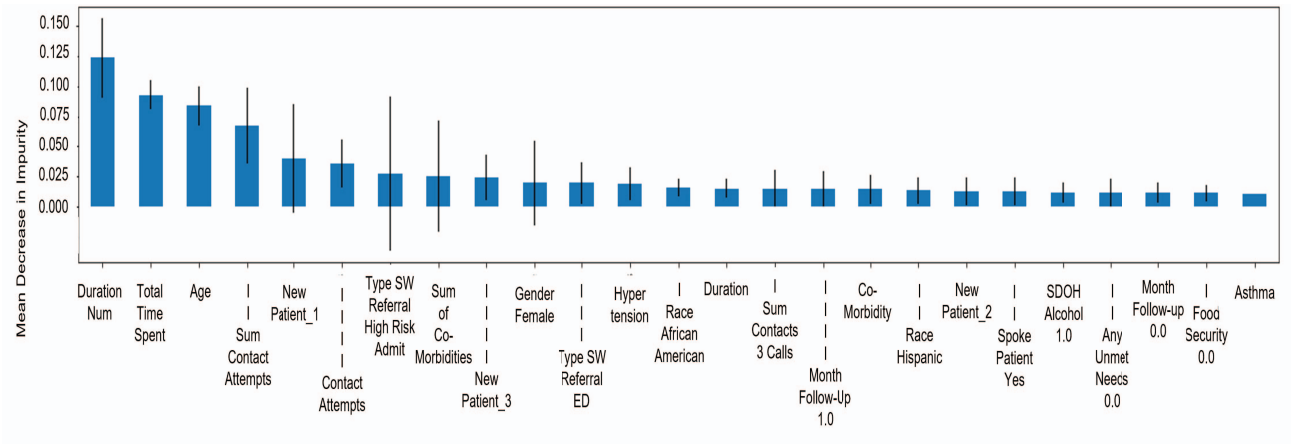

Fig. 3. Feature importance for engaged patients only showing the 25 most important features in the RF classification.

predicting the 30-day readmission, however this is only one healthcare outcome, the program aims to help improve patients' health overall. Hence, this finding mainly indicates that the members that were enrolled in the program the longest, who had many touchpoints with a CHW, are more likely to be higher risk.

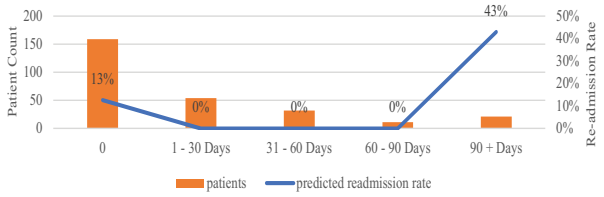

Fig. 4. Feature importance: Count of patients on the left axis and predicted percent readmissions for various durations of CHW interventions.

3) New (Recurring) Patient: This new engineered feature counts the number of times a patient is referred to the CHW program. As expected, the readmission rate jumps from minimal (near 0) to 59% for second-visit patients and to 90% for third-visit patient This result is important as a second and third referral should raise a red flag and perhaps be treated differently by the CHW, with the understanding that the risk of readmission is much higher than the average for these patients.

4) Type of referral: This important feature confirms that the scoring used to mark patients as high-risk is indeed important in predicting the readmission rate. For our test set, the predicted rate was higher (20.0%) for "high-risk" referrals than for patients coming from the ED (0%) and the aggregated "Other" (0%). All patients in the test set were high-risk, which is not entirely surprising since nearly half the patients (54.9%) were "high-risk".

5) Total time spent: This is equal to the time a CHW has spent talking to a patient since the first contact as well as the amount of time spent finding resources, researching, and any other work associated with the patient since their first

contact. The average total time spent was 29 minutes. Like duration, readmission rate is high when there is no time spent talking with a patient. There are also higher readmission rates at the higher total time spent, which may also be indicative of patients that are sicker and at higher risk (see Figure 5).

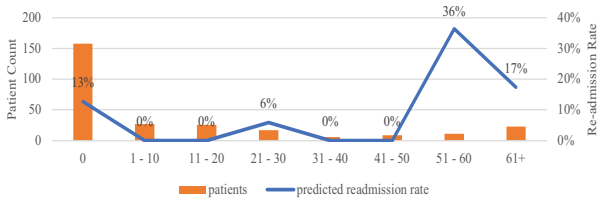

Fig. 5. Feature importance: count of patients and predicted percent readmissions for various amounts of total time spent looking for resources.

6) Gender: Within the total population, 55% were Male and 45% were Female. Males tended to have a higher readmission rate than Females. Interestingly, the females were the older population by around 3 years, at 59 years old on average. Females also had slightly more comorbidities on average than males, 0.70 compared to 0.59. Hence, this results requires further investigation.

7) Sum of contacts: This feature represents the total amount of time spent contacting a patient in minutes. The average time spent contacting a patient was around 7 minutes. The most important takeaway here is that the highest rate is again found when there is no contact to CHW. Similarly, to duration, beyond a certain limit, more time spent does not necessarily prevent readmissions. Figure 6 illustrates this behavior.

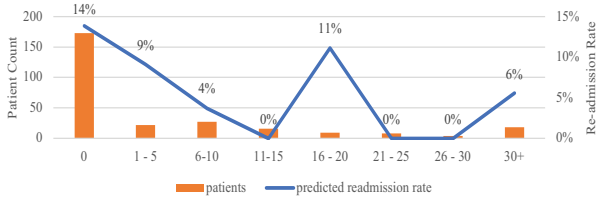

Fig. 6. Feature importance: count of patients and predicted percent readmissions for various total numbers of contacts.

8) Sum of comorbidities: The readmission rate does increase as expected with added morbidities (9%, 11%, 17%) from 0, 1, and 2 comorbidities. The risk fell to 0 for 3 morbidities, however this is more likely attributed to the very low number of such patients in the data set and the test set.

9) Number of Contact Attempts: The number of attempts made represents the amount of times a CHW reached out to a patient. Like for duration and total time spent, there is a high readmission rate for patients that did not have any contact with a CHW as well as for patients that had more contacts than the average 1.7 times, and slightly higher than average at 12% for patients who had more than six contact attempts, perhaps indicating patients whom CHWs are not able to reach.

10) Race and ethnicity: Interestingly here, the rate of readmission was lower for Black (7%) compared to 17% for non-Black patients. While this also points to more investigation that takes into account the demographics of the CHWs, we hypothesize that the close connection between CHWs and patients may be a factor.

E. Summary of Findings and Recommendations

This systematic analysis has provided evidence of the positive impact of the CHW program. 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 (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.

1) 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 1 could automatically be considered high-risk.
- Similarly, recurring patients are particularly high-risk and should be treated as such.
- 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 getting more data.

2) 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%.

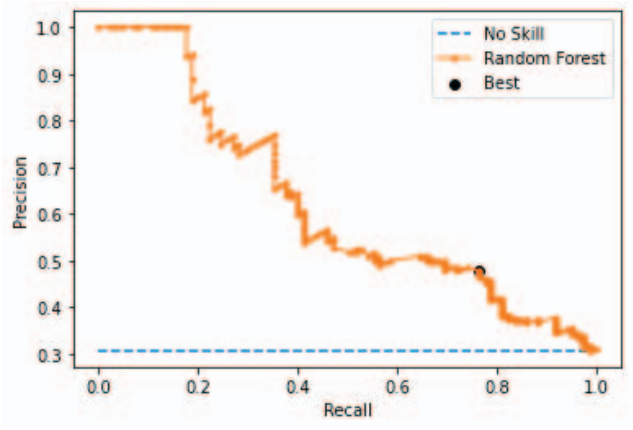

Fig. 7. Precision Recall Curve showing inflection point (maximum) F1 score.

3) Adding CHW/SDoH to EMRs: One recommendation, knowing that not all readmissions will be prevented is to leverage more information from EMRs. We know for example that comorbidities are important, therefore more EMR dada may yield similar impact on classification. Conversely, we show that some readmissions are avoidable for engaged patients, due to social rather than medical intervention. Therefore, combining the two types of data would be beneficial and more complete in an attempt to prevent any type—medical or social—of avoidable readmissions.

V. 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 $79.8 \pm 4.8\%$. 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 early to prevent readmissions and the integration of social determinants of health in EMR data for more complete analysis of how social factors impact health outcomes in a broader range of patient types.

References

[1] B. Robson and R. Harris, "Hauora: Màori 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] A. P. H. Association *et al.*, "Community health workers. american public health association website," 2016.

[10] 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? results from a national study," *Population Health Management*, 2022.

[11] C. Van Walraven, I. A. Dhalla, C. Bell, E. Etchells, I. G. Stiell, K. Zarnke, P. C. Austin, and A. J. Forster, "Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community," *Cmaj*, vol. 182, no. 6, pp. 551–557, 2010.

[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] J. Morita, *Unequal cities: structural racism and the death gap in America's Largest Cities*. JHU Press, 2021.

[15] S. S. Gervasi, I. Y. Chen, A. Smith-McLallen, D. Sontag, Z. Obermeyer, M. Vennera, and R. Chawla, "The potential for bias in machine learning and opportunities for health insurers to address it: Article examines the potential for bias in machine learning and opportunities for health insurers to address it." *Health Affairs*, vol. 41, no. 2, pp. 212–218, 2022.

[16] C. for Disease Control and Prevention, "About social determinants of health (sdoh)," 2023.

[17] 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.

[18] 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-social-and-behavioral-domains-and-measures-in-electronic-health-records