What Data is Needed to Accurately Determine Someone's Mental Health?

As the demand for mental healthcare increases, more people have turned to mental health chatbots and other AI-powered teletherapy options. While these chatbot applications expand access to care that is often expensive and stigmatized, they also pose some cybersecurity risks. Some chatbot applications use phone sensors and other device data to make predictions about the severity of a patient's mental illness. While this makes these predictions more accurate, they might compromise user trust. In this paper, we are trying to find which data we should collect from the user's phone to minimize the amount of data being collected but also collecting adequate information to predict their mental health score. This is important because users are not fully trusting the mobile apps since they are afraid the apps might track all of their personal information. If we determine which data is actually necessary, we can build better user trust while maintaining the efficacy of a chatbot. The overall approach was to first collect the data that was deemed useful, such as education, sensing, survey, and Ecological Momentary Assessment (EMA) data, and then found how the accuracy of the scores would decrease when each category of data collected was removed to determine the most important category to collect, as well as what combination of categories yielded the highest accuracy. The most significant result was that when we didn't include the data from the surveys, the accuracy went down significantly. The major conclusions are that not every detail from a user's phone will need to be collected in order to yield accurate results, and in fact, simply asking users about their daily experiences allows for far more accurate results than on-device measures.

The question we were trying to answer was what kind of data from someone's phone is needed in order to accurately determine the state of someone's mental health. With this information, we could be more transparent about what is being collected when someone downloads a mental health app. Past studies have shown that people have overall positive perceptions of mental health chatbots due to their convenience and the lack of stigma. Coupled with the fact that there is currently a shortage of mental healthcare professionals, mental health apps have gained popularity over the years. Currently, mental health apps use data from mobile phone sensors, which could be seen as an invasion of privacy for many users. However, the more tech-literate populations still tend to trust AI therapists more. In the past, there has been research done to determine the effectiveness of mental health apps as well as how accurate data could be if everything is collected from a user's phone. The results showed that they were effective for the more common mental illnesses such as anxiety and depression, but for the more complex mental illnesses, the accuracy is often questioned. Another study has also shown that the overall accuracy of mental health apps have been relatively low in the past.

The dataset that we used in order to determine how device information can and should be used to predict user wellbeing was the Student Life dataset from Dartmouth that was collected with the intention to assess depression scores of the users. 59 Dartmouth undergraduates enrolled in a computer science course answered a variety of surveys at the beginning and end of their semesters, and answered a few questions about sleep and other daily habits each morning. The rest of the data were automatically collected from their devices. The data was in the form of .csv and .json files. It was split into user info, sensing, EMA, education, and survey data. In the end, we only used some of the data, such as activity

inference, conversation time, and phone lock time from sensing, sleep, stress, and exercise time in EMA, flourishing scale, PHQ9 score, stress scale, and also loneliness scale from survey, and for education, we used their GPA and total time spent on Piazza to calculate their scores. However not all students had data for all of these parts, and in the end, around 21 students had all of the data.

The ecological momentary assessment data that was collected was about sleep, stress, and exercise. Each morning, participants indicated how many hours they had slept the night before. We took the median number of hours as a metric of how much an individual user was sleeping over the semester. For stress levels, they reported their stress level on a scale, and we took the average stress level for each user. As for exercise, we decided to take the median amount of times they spent walking and another median for the times they were actually exercising.

Sensing data was collected directly from the user's phone. The first subsection we used was their average physical activity inference. The data was already put on a scale of 0-3, where the higher the number meant the more rigorous it was, and we only had to find the average for each user. Next, we calculated the amount of time each person spent conversing with another person throughout the entire semester. The data was stored using a Unix timestamp, so we only had to subtract the ending time from the starting time and then find the sum. The final subsection was their average phone lock time. Similar to conversation time, the data used a Unix timestamp, so the process of manipulating the data was similar, except this time we used the average amount of time and not the total amount of time.

For educational data, we decided to use their GPA at the end of the semester as well as the total time they spent on Piazza. The data for GPA included a user's cumulative GPA with different scales, and we decided to use the 4.0 GPA scale since we were most familiar with it. Next, for the Piazza data, we only decided to take the total amount of time the user was on Piazza and how many times they asked questions or made contributions.

The final section we used was survey data. To create the base comparison, we decided to take their PHQ9 test responses as an accurate analysis of their mental health. For all of the survey data, they answered a form with the responses such as "not at all," "sometimes," or "all the time." By looping through the data and converting all the responses from a scale of 0-3, we were able to convert all the data into numbers. From there, we found the sum of each user's response and used that number for all of the surveys: stress, flourishing, and loneliness scale.

This data we collected was used to create a baseline prediction for the user's depression score, with the actual score being the PHQ9 score from the user. We then created a baseline prediction accuracy with all of the data included. Since there was going to be some redundant information in all of the measures, losing one section individually would not decrease the accuracy a lot. However, if losing a single section or factor leads to a huge decrease in accuracy, it means that that specific data is important for the overall accuracy and should be collected. After removing each data subsection individually, we then removed the data based on the category that it was in, either EMA, sensor, or survey data. Our prediction here was that the survey/self-report data would be the most important while the EMA and sensor data would lead to a slight decrease in accuracy, but not enough accuracy to justify collecting all of the data from a user. To determine the accuracy of each try, we used linear regression from sklearn.

Our results showed that not all the data that was collected in our dataset was necessary to determine a user's overall depression score. However, they also showed that survey data affected the overall accuracy the most, while the sensor data affected it slightly, so we can afford to not collect as much data from a user's phone while also yielding accurate results.
After finding the accuracy for all the data, we separated the data into sensor, education, EMA, and survey data and removed them individually to see how it would affect the accuracy score. The sensor data included the average conversation time, average phone lock time, and the average activity inference. When removing the sensor data, we found that it did not have a significant impact on the overall accuracy score (84.6% accuracy, as opposed to our baseline of 87.9% accuracy). The education data included the user's GPA at the end of the semester as well as the total amount of time a user spent on Piazza asking questions throughout the entire semester. When we removed the education data, it had a greater effect on the accuracy score than the sensor data, but it still was not significant (83.9% compared to 87.9%). The Ecological Momentary Assessment (EMA) data included the mean number of sleep the user got throughout the semester, their average stress levels, and median amount of time exercising and walking. When we removed the EMA data, it had a more significant impact on the overall accuracy score than the removal of the sensor and education data. However, none of their removals affected the overall accuracy score more than the removal of survey data. The surveys that were used were the overall stress levels, loneliness scores, and the flourishing score.

Data used	Accuracy score
All data used	0.879
All - Sensor	0.846
All - Education	0.839
All - EMA	0.751
All - Surveys	0.657

Table 1. *Accuracy scores for a linear regression model trained to predict end-of-semester depression scores (PHQ-9) with different subsets of student data*

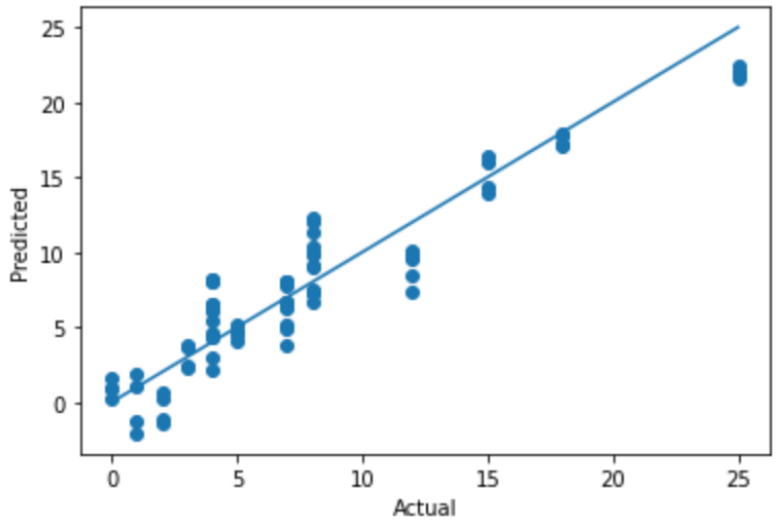

Figure: *predicted versus real for the version that includes all predictors*

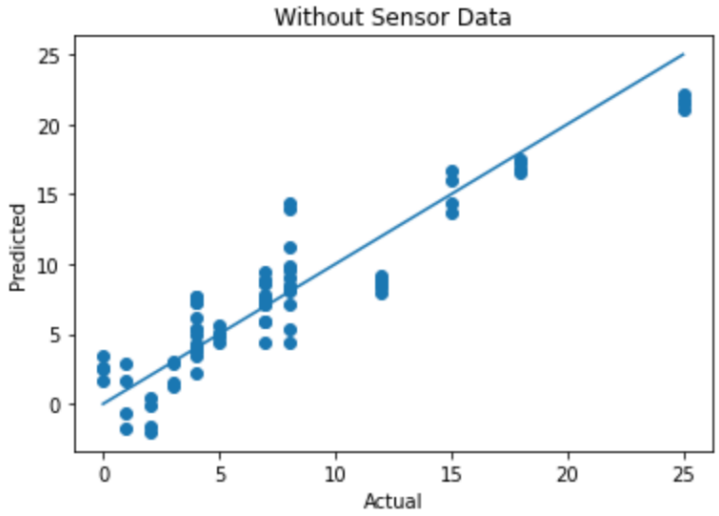

Figure: *predicted versus real without sensor data*

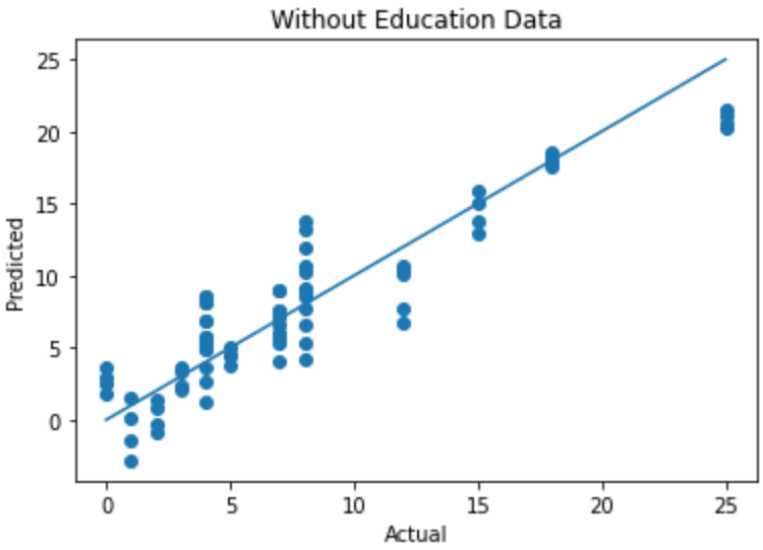

Figure: *predicted versus real without education data*

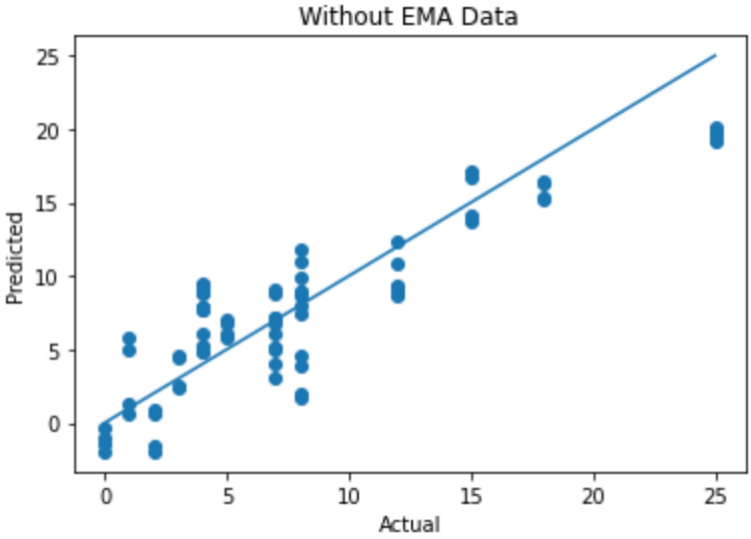

Figure: *predicted versus real without EMA data*

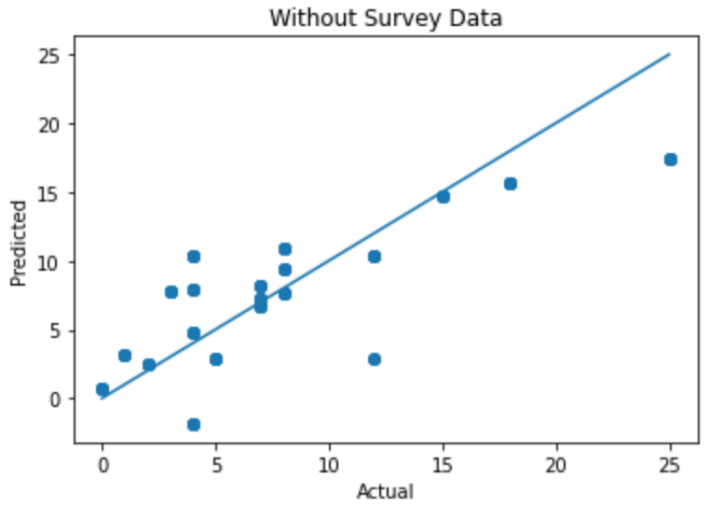

Figure: *predicted versus real without survey data*

We can conclude that by removing sensor and education data, the overall accuracy will not be affected significantly. This can apply to mental health chat bots because users will be able to know what kind of data is being collected from their phones, and the data being collected is also not very invasive of their privacy. Currently, sensor and education data are being collected since it is a passive way to collect data about someone, while survey and EMA data requires the user to answer questions. It is easier for the user to not have to answer surveys everyday, so chatbots collect the sensor and education data without active user input. Having users self-report their mental health status may also be inaccurate. In a previous study, it was shown that the results of someone self-reporting their improvement vs. the clinician-rated results were vastly different (Cuijpers et al., 2010).

With these results in mind, it would be useful to run a follow-up experiment in which we tested our code on a different dataset to see if we had similar results. If this follow-up dataset were more robust, we could feel more confident in our conclusion that mobile sensor data should not be collected by mental health chatbot apps because they compromise user trust and safety without providing better service. The dataset could be more robust in several ways: people from other age groups could be included, the total time could be longer than one semester, it might be collected in a different country, and the data might be more specific and go into detail about which app is open and how many times it was opened.

As chatbots grow in popularity, there have been a lot more privacy concerns as well. We decided to determine to what degree certain information is helpful for diagnosing mental health conditions, specifically depression. By using real behavioral data from students at Dartmouth through an online dataset, we tried to predict their depression score from different subsets of information. From the results, we can see that not all information that is currently being collected by mental health apps are needed to determine if a user has depression. With this information, there can be more transparency with the data that is being collected, and users will be able to trust the apps more.

References

Abd-Alrazaq, A. A., Rababeh, A., Alajlani, M., Bewick, B. M., & Househ, M. (2020). *Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis*. Journal of Medical Internet Research. Retrieved January 1, 2023, from https://www.jmir.org/2020/7/e16021/

Abd-Alrazaq, A., Alajlani, M., Ali, N., Denecke, K., Bewick, B., & Househ, M. (2021). *Perceptions and Opinions of Patients About Mental Health Chatbots: Scoping Review*. Journal of Medical Internet Research. Retrieved January 1, 2023, from https://www.jmir.org/2021/1/e17828/PDF

Hofmannd, S. G., Li, J., Cuijpers, P., & Andersson, G. (2010, June 18). *Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: A meta-analysis*. Clinical Psychology Review. Retrieved January 1, 2023, from https://www.sciencedirect.com/science/article/abs/pii/S0272735810000954

Jungmann, S. M., Klan, T., Kuhn, S., & Jungmann, F. (2019, October 29). *Accuracy of a chatbot (ADA) in the diagnosis of mental disorders: Comparative case study with Lay and expert users*. JMIR formative research. Retrieved January 1, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6914276/

Mohr, D., Zhang, M., & Schueller, S. (2019). *Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning*. HHS Public Access. Retrieved January 1, 2023, from https://europepmc.org/backend/ptpmcrender.fcgi?accid=PMC6902121&blobtype=pdf

Nadarzynski, T., Miles, O., Cowie, A., & Ridge, D. (2019). *Acceptability of Artificial Intelligence* . Digital Health. Retrieved January 1, 2023, from https://journals.sagepub.com/doi/10.1177/2055207619871808

Weiner, S. (2022, August 9). *A growing psychiatrist shortage and an enormous demand for mental health services*. AAMC. Retrieved January 1, 2023, from https://www.aamc.org/news-insights/growing-psychiatrist-shortage-enormous-demand-mental-health-services