Co-Designing Effective Human-Algorithm Collaborations for Suicide Risk Assessment with Clinicians, Teens, and Families

ELENA SWECKER[*], Carnegie Mellon University, Human-Computer Interaction Institute

Early suicide risk detection is critical in suicide prevention, as the majority of suicide deaths between the ages of 10 and 24 occur from the first attempt [1]. Given the challenges of assessing suicide risk based on patient self-reporting, the use of algorithmic decision-support systems (ADS) to support clinicians' diagnostic capabilities has seen growing interest over the past several years [4]. However, knowledge surrounding the effective and ethical implementation of modern ADS into clinical practice remains sparse. The purpose of this study is to inform the design of new ADS interfaces and training modules that support responsible and ethical use of suicide risk algorithms in clinical practice.

We conducted a series of semi-structured interviews and participatory design activities with individual participants, in which they were asked to generate and evaluate new ideas for how these kinds of clinical decision support systems should (or should not) be designed. A preliminary thematic analysis revealed that most participants feel very strongly about patient-focused scenarios and emphasized patient privacy, trust, and involvement in the decision-making. We have also gained insight into methods of supporting patients and families in processing the results of the ADS, and what input data would be meaningful to consider. These preliminary and exploratory results will lead to further investigation into the design and implementation of these tools, especially with a focus on trust and privacy.

1 INTRODUCTION

Early suicide risk detection is critical in suicide prevention, as the majority of suicide deaths between the ages of 10 and 24 occur from the first attempt [1]. Unfortunately, given the isolation of teens during the COVID-19 pandemic and the limited transparency in patients' descriptions of their suicidal ideation, mental health clinicians are provided with insufficient information. To combat these complications, the use of data-driven ADS to support clinicians' diagnostic capabilities has seen growing interest over the past several years [4].

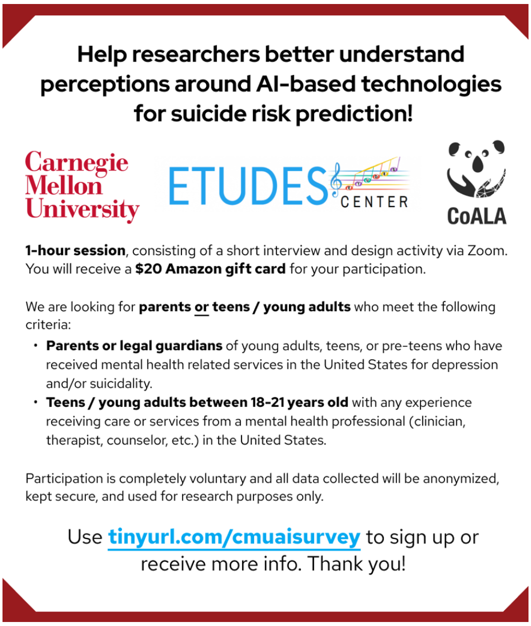
However, knowledge surrounding the effective and ethical implementation of modern ADS into clinical practice remains sparse. The purpose of this study is to inform the design of new ADS interfaces and training modules that support responsible and ethical use of suicide risk algorithms in clinical practice. Without this careful design, the use of risk assessment algorithms may be perceived by teens as threatening their autonomy and privacy, reducing trust between patients and clinicians [3]. Additionally, when the algorithm's predictions contradict with the patient's claimed psychological state, communication by clinicians of these predictions must be handled carefully to again avoid damaging trust in the clinician-patient relationship. Overall, by engaging participants in actively shaping and reflecting upon what human-algorithm collaboration in suicide risk assessment should look like, these ADS can be more carefully designed for ethical and effective use that maintains trust between teen patients and clinicians.

1.1 Approach

We conducted a series of semi-structured interviews and participatory design activities with individual participants, in which they were asked to generate and evaluate new ideas for how these kinds of clinical decision support systems should (or should not) be designed.

At a high level, the interviews helped us to gain insight into a participant's level of familiarity with ADS and experience with mental health professionals. From there, the central aspect of these

[*] In collaboration with Connie Chau, Nick Su, and Professor Ken Holstein



Help researchers better understand perceptions around AI-based technologies for suicide risk prediction!

Carnegie Mellon University
ETUDES CENTER
CoALA

1-hour session, consisting of a short interview and design activity via Zoom. You will receive a **$20 Amazon gift card** for your participation.

We are looking for **parents or teens / young adults** who meet the following criteria:

- **Parents or legal guardians** of young adults, teens, or pre-teens who have received mental health related services in the United States for depression and/or suicidality.
- **Teens / young adults between 18-21 years old** with any experience receiving care or services from a mental health professional (clinician, therapist, counselor, etc.) in the United States.

Participation is completely voluntary and all data collected will be anonymized, kept secure, and used for research purposes only.

Use **tinyurl.com/cmuaisurvey** to sign up or receive more info. Thank you!

Fig. 1. An example recruitment poster

session is conducting iterative participatory design activities, using the Participatory Speed Dating (PSD) method [2], with the goal of engaging participants in generating, evaluating, and revising new technology concepts. The general procedure is to show a participant a series of storyboards that each illustrate a different hypothetical scenario involving this technology. For each, participants are asked to discuss their immediate reactions and suggest any modifications that they might find more realistic or interesting.

A Participatory Speed Dating approach can lead to “the discovery of unexpected design opportunities, when unanticipated needs are uncovered or when anticipated boundaries are discovered to not exist" [2]. Further, these sessions help to generate new insights, whether they be problems, solutions, or considerations, which are crucial for a preliminary study such as this.

1.2 Key Contributions

Firstly, this research has generated insights into the design of ADS interfaces for suicide risk algorithms. Beyond that, a key contribution of this research is the success of applying the PSD method to a new field within Human-Computer Interaction and human-centered design. More specifically, some modifications to the previous PSD approach, such as a reverse storyboarding process (as described below), have allowed for additional valuable discussion and are a novel development in this category of study technique.

2 METHODOLOGY

2.1 Recruitment

During the first phase of this study, we focused on recruiting parents and teens. An example of the recruitment poster for these stakeholder groups can be seen in Figure 1, which includes some restrictions in accordance with our Institutional Review Board approval. For parents, this additional criteria was that they must have teenage children who have received mental health services specifically for depression or suicidality.

Further, we were not able to conduct interviews with our target group, younger teens, but we were able to recruit from the older teen/young adult category (i.e. 18 to 21). Additionally, we could not ask specifically that they have experience with depression or suicidality, but we did require that they have some experience with any sort of mental health professional. Then, during the preliminary part of the interview, they could elect to reveal any addition details as they felt comfortable.

Initial recruiting took place via Reddit subthreads (such as r/samplesize and r/askparents). However, a series of abnormal and illegitimate sign-ups led us to take down those initial posts and try again using platforms such as NextDoor and relevant Facebook groups (where we saw no additional issues).

Although we attempted to recruit from both of these stakeholder groups, there were limited parental sign-ups, so the following analysis and protocol description will reflect only the young adult interviews conducted so far (as were the vast majority).

2.2 Semi-Structured Interviews

During a session, the first component was a semi-structured interview, in which participants were asked to answer a series of questions (at their comfort level, with however much detail they desired, or not at all). These gave us insight into relevant background information such as:

- Their level of familiarity with “clinical decision support technologies for suicide risk prediction" (or other support technologies)
- Their experience, comfort level, and trust working with mental health professionals
- Their experience with depression and/or suicidality (personally or with people they know)

2.3 Iterative Design Activity (Participatory Speed Dating)

As mentioned earlier, the general procedure for this second part of the session is to show a participant a series of storyboards that each illustrate a different hypothetical scenario involving the ADS (specifically, an AI algorithm to predict someone’s risk for suicide). We discuss their first impressions and immediate reactions, which give insight into how realistic, preferable, etc. each situation may be. At any time, participants may suggest modifications to a storyboard to investigate new ideas, illustrate more realistic situations, demonstrate a potentially more favorable outcome, etc. Some storyboards are incomplete to explicitly encourage participants to provide their own suggestions.

The storyboards also reflect a variety of scenarios that each have different stakeholders (i.e. parents, clinicians, or teens) as the central focus of the situation. However, we decided to show all participants all of the storyboards, since one stakeholder group could have valuable insights into situations that focus on another stakeholder group even if they may not have as much personal experience with that scenario.

A basic storyboard example can be seen in Figure 2. In this example, a parent is attending a clinician visit with their child. The AI algorithm suggests that the child has a high risk for suicide, but the parent disagrees. However, after discussing the mechanisms of the algorithm with the clinician, the parent is convinced that they should make a plan. This one storyboard raises several conflicting perspectives on the persuasiveness of simply understanding the algorithm’s specifics: some believed that this was realistic and a parent would be reasonably convinced, and others strongly disagreed, noting that it would just seem like their child was not being individually considered (and rather being viewed as the same as past patients). This was the structure of the basic

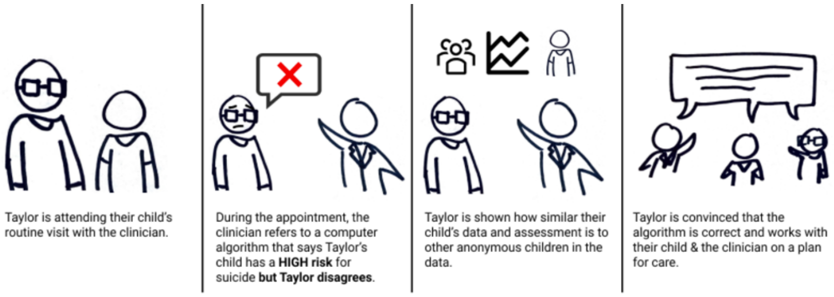

Taylor is attending their child's routine visit with the clinician.
During the appointment, the clinician refers to a computer algorithm that says Taylor's child has a **HIGH risk** for suicide **but Taylor disagrees**.
Taylor is shown how similar their child's data and assessment is to other anonymous children in the data.
Taylor is convinced that the algorithm is correct and works with their child & the clinician on a plan for care.

Fig. 2. An example of a basic storyboard

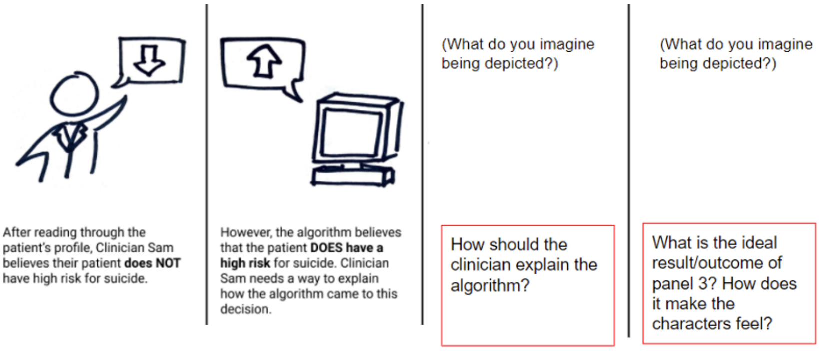

After reading through the patient's profile, Clinician Sam believes their patient **does NOT** have high risk for suicide.
However, the algorithm believes that the patient **DOES have a high risk** for suicide. Clinician Sam needs a way to explain how the algorithm came to this decision.
(What do you imagine being depicted?)
How should the clinician explain the algorithm?
(What do you imagine being depicted?)
What is the ideal result/outcome of panel 3? How does it make the characters feel?

Fig. 3. An example of a storyboard including blank panels

storyboard, which made up about 3/4 of our about 16 total storyboards. Additionally, participants could suggest modifications to these storyboards, and many altered storyboards were also used in later sessions.

For a couple storyboards, panels were intentionally left blank so that participants must generate a likely outcome to a given situation. In the example in Figure 3, the clinician and the algorithm have conflicting analyses on the patient's suicide risk, and the participant must discuss how the clinician should communicate these results and weight the opposing assessments. The use of missing panels directly generated discussion regarding communication about the AI algorithm, as desired.

Finally, over the course of this first phase of the study, we designed a novel modification to the PSD approach, conversationally referred to as “reverse storyboard completion." Similar to the previous example, reverse storyboard completion involves removing the first couple panels of a storyboard, rather than the last couple, so that participants generate the problem that a given solution could solve (rather than visa versa). This approach is intended to provide insight into the main problems with suicide risk assessment technology that stand out to participants.

Figure 4 shows a final storyboard example, this time employing the reverse storyboard completion approach. In this storyboard, participants would be asked to discuss what kind of situation an explicit step-by-step guide to explaining the algorithm would be useful in. Primarily, this raised issues regarding disagreements of opinion between various stakeholders and discussions on how the guide could be used to resolve them. These threads would not have been as directly examined if not for the reverse storyboard completion technique, which encourages problem ideation.

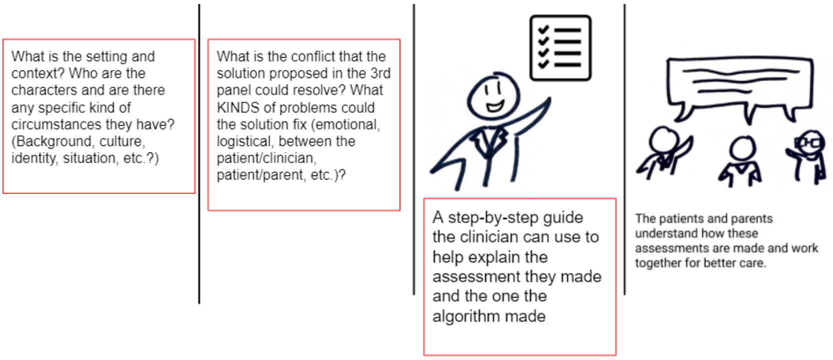

What is the setting and context? Who are the characters and are there any specific kind of circumstances they have? (Background, culture, identity, situation, etc.?)
What is the conflict that the solution proposed in the 3rd panel could resolve? What KINDS of problems could the solution fix (emotional, logistical, between the patient/clinician, patient/parent, etc.)?
A step-by-step guide the clinician can use to help explain the assessment they made and the one the algorithm made
The patients and parents understand how these assessments are made and work together for better care.

Fig. 4. An example of a storyboard involving the reverse storyboard completion technique

3 PRELIMINARY RESULTS

Because we are still in the preliminary stages of this research, no data-driven analysis has been completed on our current results. However, a broader thematic analysis has revealed some key observations and trends across the sessions completed so far.

First of all, most participants feel strongly about patient-focused scenarios, emphasizing patient privacy, trust, and involvement in decision-making as the most important elements of inter-stakeholder relationships and the introduction of these AI technologies. Even beyond the storyboards that focused on patients, participants highlight that transparency and patient privacy are among their chief concerns. For example, all data used by the algorithm should be explicitly cleared by the patients, and any assessments by clinician or algorithm should be transparently discussed with the patients.

Additionally, most participants have not heard of suicide risk assessment technologies before but stress that it should be used principally as a tool (i.e. more to support clinician assessment and less as a direct diagnostic resource, so the clinician should remain very involved in these assessments). This is clear from consistent agreement with clinician opinion regardless of the algorithm's contribution. In situations where the algorithm's predictions are considered, it's only secondary to the clinician's perspective (in the eyes of most participants) and a cause for concern but not necessarily action.

A final theme from this preliminary and exploratory analysis is that some participants note a power dynamic that may exist between patients and parents, emphasizing that clinicians should always remain on the patient's side. The discussed power dynamic ranges from light skepticism to explicit disagreement on the parents' parts, with many of the young adult participants citing their own experiences discussing mental health with their parents and the varying levels of success in those discussions.

4 CONCLUSIONS

The majority of the above results aligned with our general predictions regarding the initial reactions of hesitancy and distrust with having an algorithm generate assessments about such a sensitive and human-focused issue. However, as technology evolves on a larger scale, it seems that people are being more open to it being involved in new ways. This was reflected by participants communicating comfort with the algorithm at least being used as a tool for clinicians, and perhaps even as a strong factor (with a high enough confidence level).

However, with any human-centered study such as this, it is hard to quantify the nuanced and differentiating perspectives discussed across all sessions. As this study continues to grow (described in the next section), specific and valuable insights into the design of this algorithm will emerge that can begin to be used by our collaborators at the University of Pittsburgh in the implementation of these tools into practice.

5 FUTURE WORK

As mentioned above, the research up to this point has given us a preliminary and exploratory look at the perspectives that exist among one stakeholder group, but there are many areas for further investigation. First, we would like to recruit more participants from our less-represented stakeholder groups (i.e. parents and clinicians) to expand the dimensionality of perspectives we gather. We would also like to increase the number of participants in each stakeholder group to obtain a more distributed sample of perspectives.

Additionally, we will be reflecting upon the preliminary results discussed above to determine alterations to our current interview and participatory activity protocols and designs. This may involve examining more closely a particularly interesting dynamic that recurs across the sessions (though it is unclear at this point which dynamic this may be).

Even further down the line, we would like to consider expanding from individual participatory speed dating sessions to group sessions in order to observe multi-party interactions. And finally, we will be working with our partners at the University of Pittsburgh to determine how to incorporate our results into system design.

Overall, this study only begins to reflect the numerous and diverse considerations that must be accounted for in order to introduce these ADS in a responsible and ethical way. We hope that by continuing our investigation, along with the research of several other key groups, we can begin to introduce these tools that may even become invaluable to reducing the prevalence of teen suicide.

ACKNOWLEDGMENTS

A special thanks to Connie Chau for her incredible hard work and mentorship and for teaching me how a human-centered study is designed and executed from the ground up, to Nick Su and Prof. Ken Holstein for their valuable insight and feedback over the past few months, and to 15-400 Professors Jonathan Aldrich and Bogdan Vasilescu for their guidance and encouragement!

REFERENCES

[1] Alastair J S McKean et al. 2018. Rethinking Lethality in Youth Suicide Attempts: First Suicide Attempt Outcomes in Youth Ages 10 to 24. *Journal of the American Academy of Child and Adolescent Psychiatry* 57, 10 (2018), 786–791. https://doi.org/10.1016/j.jaac.2018.04.021

[2] Kenneth Holstein et al. 2019. Designing for Complementarity: Teacher and Student Needs for Orchestration Support in AI-enhanced Classrooms. *Lecture Notes in Computer Science* 11625 (June 2019). https://doi.org/10.1007/978-3-030-23204-7_14

[3] Lindsey C. McKernan et al. 2018. Protecting Life While Preserving Liberty: Ethical Recommendations for Suicide Prevention With Artificial Intelligence. *Frontiers in Psychiatry* 9, 650 (Dec 2018). https://doi.org/10.3389/fpsyt.2018.00650

[4] Tad Hirsch et al. 2017. Designing Contestability: Interaction Design, Machine Learning, and Mental Health. *DIS. Designing Interactive Systems (Conference)* 2017 (2017), 95–99. https://doi.org/10.1145/3064663.3064703