## PEER REVIEW HISTORY



## ARTICLE DETAILS

| | |
|---|---|
| **TITLE (PROVISIONAL)** | The creation of an electronic patient-reported outcome measure platform Voxe: A mixed methods study protocol in pediatric solid organ transplantation |
| **AUTHORS** | Anthony, Samantha; Pol, Sarah; Lin, Jia; Barwick, Melanie; Brudno, Michael; Manase, Dorin; Parekh, Rulan; Silva, Amanda; Stinson, Jennifer |

## VERSION 1 – REVIEW

| | |
|---|---|
| **REVIEWER** | Murugappan, Meena<br>FDA |
| **REVIEW RETURNED** | 10-Jun-2021 |

| | |
|---|---|
| **GENERAL COMMENTS** | Overall, I think the authors did a great job in defining the problem they are setting out to solve and describing how they plan to solve it. It would be helpful to define technical jargon early in the manuscript (e.g. scrum framework, wireframes, etc.). Lastly, the authors mentions they will collect data on participant compliance, time on task, successful task completion, frequency of critical and non-critical errors, and error-free rate. It would be helpful to add a sentence here regarding what sort of statistical analyses will be used for these data (i.e. descriptive/ summary statistics). |

| | |
|---|---|
| **REVIEWER** | Banerjee, Rahul<br>University of California San Francisco, Department of Medicine |
| **REVIEW RETURNED** | 27-Jun-2021 |

| | |
|---|---|
| **GENERAL COMMENTS** | In this piece, the authors describe a comprehensive strategy to develop, test, and implement an ePRO collection system for pediatric patients undergoing solid organ transplantation. The authors make a good case for the need for this intervention, both in the pediatric setting in general and specifically in the setting of transplantation and post-transplantation recovery.<br><br>My major concern with this manuscript as currently written is its lack of concrete objectives and endpoints to determine success. The authors describe general themes of outcomes for each phase: for example, adherence to ePROs / time on task / task completion / error frequency. But - both for their scientific protocol and for this manuscript (which is serving to set benchmarks for success transparently and a priori) - these need to be better defined. For each phase, I'd like to see a primary objective and endpoint defined. For example, "time on task" will almost certainly not be the primary endpoint with a cutoff determining success versus not. Assuming ePRO completion rate is the primary endpoint, for example for Phase 5, what is the minimum rate the authors hoping |

| | |
|---|---|
| | to see? At least in the adult setting that I am more familiar with, many studies have targeted anywhere above 50-70% ePRO completion. Even citing one of these prior studies and saying that they will use it for their current work is sufficient. If all the other objectives are exploratory, that is of course fine. But a threshold for feasibility (i.e., ePRO completion) needs to be set in advance.<br><br>A few minor points / suggestions:<br>1) Consider using the word "adherence" to ePRO assessments instead of "compliance," which has generally fallen out of favor.<br>2) What are "critical" versus "non-critical" errors?<br>3) Figure 1: I think it might be helpful to flesh this out to clearly demarcate that Phases 4-5 are being described in this protocol. Perhaps a "status update" for each Phase can be listed to show that, for example, Phase 1 is entirely complete while Phase 4 has begun recruiting. At times I found it difficult to follow which phase with which.<br>4) With the caveat that I'm more familiar with ePROs in the adult rather than the pediatric setting, I will ask - how involved are the parents/caregivers in this setting versus the 8-17-year-old patients themselves? I don't doubt that a 17-year-old can navigate to the Voxe website, remember their login information, and complete ePROs. But for an 8-year-old, I wonder whether parental assistance might be needed with some of these tasks - at least in the beginning. Can this be fleshed out more in the protocol? |

**VERSION 1 – AUTHOR RESPONSE**

Response to Reviewer 1:

- Overall, I think the authors did a great job in defining the problem they are setting out to solve and describing how they plan to solve it. It would be helpful to define technical jargon early in the manuscript (e.g. scrum framework, wireframes, etc.).
  - Response: Thank you for your positive feedback. We understand that additional definitions for technical jargon will provide clarity for readers. We have added the following definitions to the manuscript:
    - "This evidence-based approach will consider the needs of Voxe users at each design phase and will allow for iterative modification of wireframes, which are the static, two-dimensional visual representation or layout of Voxe, to best meet their identified needs." pg. 9
    - "A scrum framework is a project management process within a hybrid software development model that applies a flexible development process and places the needs of system end-users at the forefront to ensure that Voxe is both useful and usable. The principles of iterative feedback, incremental development and continual stakeholder involvement are central to this dynamic approach." pg. 13

- Lastly, the authors mention they will collect data on participant compliance, time on task, successful task completion, frequency of critical and non-critical errors, and error-free rate. It would be helpful to add a sentence here regarding what sort of statistical analyses will be used for these data (i.e. descriptive/ summary statistics).
  - Response: Thank you for your suggestion to include additional details regarding statistical analyses. We will be producing descriptive statistics and have added this detail to the manuscript.
    - "Quantitative data including time on task, successful task completion, frequency of critical and non-critical errors, and error-free rate will be analyzed and descriptive statistics will be produced." pg. 15

Response to Reviewer 2:

- In this piece, the authors describe a comprehensive strategy to develop, test, and implement an ePRO collection system for pediatric patients undergoing solid organ transplantation. The authors make a good case for the need for this intervention, both in the pediatric setting in general and specifically in the setting of transplantation and post-transplantation recovery.
  - Response: Thank you for your positive consideration.

- My major concern with this manuscript as currently written is its lack of concrete objectives and endpoints to determine success. The authors describe general themes of outcomes for each phase: for example, adherence to ePROs / time on task / task completion / error frequency. But - both for their scientific protocol and for this manuscript (which is serving to set benchmarks for success transparently and a priori) - these need to be better defined. For each phase, I'd like to see a primary objective and endpoint defined. For example, "time on task" will almost certainly not be the primary endpoint with a cutoff determining success versus not. Assuming ePRO completion rate is the primary endpoint, for example for Phase 5, what is the minimum rate the authors hoping to see? At least in the adult setting that I am more familiar with, many studies have targeted anywhere above 50-70% ePRO completion. Even citing one of these prior studies and saying that they will use it for their current work is sufficient. If all the other objectives are exploratory, that is of course fine. But a threshold for feasibility (i.e., ePRO completion) needs to be set in advance.
  - Response: Thank you for sharing your concern. We appreciate your feedback. Upon reflection, we have decided to remove the metric participant compliance or participant adherence from our protocol as we feel this metric would be better suited for Phase 6 of our research program. Establishing ePRO completion rate will be relevant for Phase 6 as it is an implementation-effectiveness evaluation.
  - The objective of the research outlined in this study protocol is to design user interfaces (Phase 4) and develop an ePROM platform (Phase 5) that is effective, efficient and satisfactory according to end-users.1-3
  - To clarify, all objectives are exploratory at this stage of research; findings from the two phases described in this study protocol (Phases 4 and 5) will inform the selection of primary and secondary objectives for our future implementation-effectiveness trial (Phase 6).
  - Additional information has been provided to the manuscript around some of the objective measures:
    - “Objective measures to be collected include: (1) time on task – the time it takes to complete each task, (2) successful task completion (fidelity) – when the end-user achieves the end goal of the task, (3) frequency of critical errors – a high severity error that could prevent an end-user from being able to complete a task, (4) frequency of non-critical errors – a low severity error that could decrease the efficiency with which an end-user completes a task, and (5) error-free rate – the percentage of task completions that occurred without any errors.” pg. 15
  - References

1. International Organization for Standardization. ISO 9000:2015(en) Quality management systems — Fundamentals and vocabulary 2015 [Available from: https://www.iso.org/obp/ui/#iso:std:iso:9000:ed-4:v1:en.
2. Barnum C. Chapter 6: Preparing for usability testing. In: Merken S, editor. Usability testing essentials: ready, set test! 2 ed. Amsterdam: Elsevier; 2020. p. 197-248.
3. Benedek J, Miner T. Measuring desirability: New methods for evaluating desirability in a usability lab setting. Usability Professionals Association. 2002;2003(8-12):57.

- Consider using the word "adherence" to ePRO assessments instead of "compliance," which has generally fallen out of favor.

o Response: Thank you for bringing this to our attention. We have since removed participant compliance from our protocol.

• What are "critical" versus "non-critical" errors?
o Response: Thank you for your question. A critical error is defined as a high severity error that could prevent an end-user from being able to complete a task. For example, if an end-user navigated to a screen that does not allow them to navigate back to a screen which allows them to complete the task. A non-critical error is defined as a low severity error that could decrease the efficiency with which an end-user completes a task. For example, if there are multiple paths to reach a specific screen and the end-user navigates on the path with more clicks (i.e., the less efficient path), but still has the ability to complete the task.
o The definitions of critical and non-critical errors have been added to the manuscript as noted in a response above.

• Figure 1: I think it might be helpful to flesh this out to clearly demarcate that Phases 4-5 are being described in this protocol. Perhaps a "status update" for each Phase can be listed to show that, for example, Phase 1 is entirely complete while Phase 4 has begun recruiting. At times I found it difficult to follow which phase with which.
o Response: Thank you for your suggestion to amend Figure 1. We agree that including a status update for each Phase would be helpful for readers. We have included this information in the revised Figure 1.

• With the caveat that I'm more familiar with ePROs in the adult rather than the pediatric setting, I will ask - how involved are the parents/caregivers in this setting versus the 8-17-year-old patients themselves? I don't doubt that a 17-year-old can navigate to the Voxe website, remember their login information, and complete ePROs. But for an 8-year-old, I wonder whether parental assistance might be needed with some of these tasks - at least in the beginning. Can this be fleshed out more in the protocol?
o Response: Thank you for your comment and for sharing your experience. You have clearly articulated our motivation for developing Voxe. It is critical to us that Voxe is pediatric patient friendly, and a lot of effort has been put into every stage of this research program to maximize Voxe's ease of use for this specific demographic. For example, we have actively and meaningfully engaged patients throughout this research and will be recruiting patient participants as young as 8 years of age (in Phase 5) to elicit their feedback on Voxe.
o It is very important to us that we evaluate learnability, which is "the extent to which something can be learned efficiently".1 Learnability along with parental involvement will be explored qualitatively.
o In future phases of this research program when pediatric patients will be completing ePROMs via Voxe, all patients will receive thorough and extensive instructions and an orientation to the platform itself. The purpose of this is to mitigate some of the concerns noted in your comment.
o References
1. Tullis TA, Bill. Chapter 4 – Performance Metrics. Measuring the User Experience. Second ed: Elsevier; 2013. p. 63–97.

**VERSION 2 – REVIEW**

| | |
|---|---|
| **REVIEWER** | Banerjee, Rahul<br>University of California San Francisco, Department of Medicine |
| **REVIEW RETURNED** | 04-Sep-2021 |

| | |
|---|---|
| **GENERAL COMMENTS** | The authors have made several modifications to their manuscript based on the suggestions by the other reviewer and me, and this revision reads much more strongly and clearly.<br><br>My two minor new suggestions - and I will defer to the editor whether to request a editorial-desk-only revision versus disregard these and accept the manuscript as is - would be:<br><br>1) Abstract - I suggest including the ages of patients (10-17 for "first" phase = Phase 4, 8-17 for "second" phase = Phase 5) in the abstract. In their response letter, the authors highlighted this fact to underscore that the patients themselves will be filling these out and never their parents/caregivers - which I hadn't completely conceptualized and which adds to the importance of the Voxe platform. I think including the ages in the abstract will underscore this for the reader as well.<br><br>2) Phase 5 - Even for feasibility studies, a pre-defined threshold of task completion / adherence / acceptability is key to define success. (For Phase 4, this isn't applicable since data being collected are qualitative stakeholder input and the end product is the Voxe platform). But for Phase 5, the authors define success as Voxe being "considered acceptable to participating end-users with no further refinements identified." What does this mean specifically? Will there be additional stakeholder interviews to define acceptabiilty? Or - with regard to the objective measures being collected in this phase - will success be extrapolated from the fact that, say, at least 50% of tasks are being completed and that no more than 10% of patients encountered a critical error? It's fine to set the bar low, but I do think identifying some sort of a quantitative bar a priori is critical to any research that involves objective endpoints (even non-pharmacologic ones like this). |

**VERSION 2 – AUTHOR RESPONSE**

Response to Reviewer 2:

- The authors have made several modifications to their manuscript based on the suggestions by the other reviewer and me, and this revision reads much more strongly and clearly.

o Response: Thank you for your positive feedback.

- My two minor new suggestions - and I will defer to the editor whether to request a editorial-desk-only revision versus disregard these and accept the manuscript as is - would be: Abstract - I suggest including the ages of patients (10-17 for "first" phase = Phase 4, 8-17 for "second" phase = Phase 5) in the abstract. In their response letter, the authors highlighted this fact to underscore that the patients themselves will be filling these out and never their parents/caregivers - which I hadn't completely conceptualized and which adds to the importance of the Voxe platform. I think including the ages in the abstract will underscore this for the reader as well.

o Response: Thank you for your suggestion. We have included the ages of patients in the abstract.

 “Transplant recipients, aged 10 to 17, and healthcare providers will participate in three rounds of testing (24 participants total)… Transplant recipients, aged eight to 17, and healthcare providers will participate in four rounds of iterative testing (24 to 40 participants total).” pg. 2

• Phase 5 - Even for feasibility studies, a pre-defined threshold of task completion / adherence / acceptability is key to define success. (For Phase 4, this isn't applicable since data being collected are qualitative stakeholder input and the end product is the Voxe platform). But for Phase 5, the authors define success as Voxe being "considered acceptable to participating end-users with no further refinements identified." What does this mean specifically? Will there be additional stakeholder interviews to define acceptability? Or - with regard to the objective measures being collected in this phase - will success be extrapolated from the fact that, say, at least 50% of tasks are being completed and that no more than 10% of patients encountered a critical error? It's fine to set the bar low, but I do think identifying some sort of a quantitative bar a priori is critical to any research that involves objective endpoints (even non-pharmacologic ones like this).

o Response: Thank you for sharing your concern. Stakeholder interviews are being conducted to assess implementation outcomes. Additional information has been provided to the manuscript around objective endpoints:

 “…effectiveness – accuracy and completeness with which users achieve specific goals, displayed as a percentage of tasks successfully completed (average task completion rate is 78%; above average is considered successful task completion)…” pg. 10

 “The System Usability Scale is considered a reliable way to evaluate electronic platforms and a score of 68 is considered above average.” pg. 11

 “…successful task completion (fidelity) – when the end-user achieves the end goal of the task successfully (above the average task completion rate of 78% is considered successful)…” pg. 15

 “A score on the System Usability Scale that is greater than 68 is considered above average.” pg. 15

 “Patient participants will complete the Microsoft Desirability Toolkit by selecting words from a list of product reaction words to describe their attitude towards the Voxe platform.” pg. 15

o References

1. Sauro J. What Is A Good Task-Completion Rate? Secondary What Is A Good Task-Completion Rate? 2011 [Available from: https://measuringu.com/task-completion/.

2. Cho H, Powell D, Pichon A, Kuhns LM, Garofalo R, Schnall R. Eye-tracking retrospective think-aloud as a novel approach for a usability evaluation. International journal of medical informatics. 2019;129:366-73.

3. Haggerty T, Brabson L, Grogg KA, Herschell AD, Giacobbi Jr P, Sedney C, et al. Usability testing of an electronic health application for patient activation on weight management. Mhealth. 2021;7.

4. Brooke J. SUS-A quick and dirty usability scale. Usability evaluation in industry. 1996;189(194):4-7.

5. Barnum C. Chapter 6: Preparing for usability testing. In: Merken S, editor. Usability testing essentials: ready, set test! 2 ed. Amsterdam: Elsevier; 2020. p. 197-248.

6. Benedek J, Miner T. Measuring desirability: New methods for evaluating desirability in a usability lab setting. Usability Professionals Association. 2002;2003(8-12):57.