Quantitative and Qualitative Evaluation of Provider Use of a Novel Machine Learning Model for Favorable Outcome Prediction

Elisabeth Yang[1], MA, Yin Aphinyanaphongs[2], MD, Paawan V. Punjabi[2], MD, Jonathan Austrian[2], MD, Batia Wiesenfeld[3], PhD

[1]Yale School of Management, New Haven, CT;[2]NYU Langone Health, New York, NY; [3]NYU Stern School of Business, New York, NY

ABSTRACT

Predictive models may be particularly beneficial to clinicians when they face uncertainty and seek to develop a mental model of disease progression, but we know little about the post-implementation effects of predictive models on clinicians' experience of their work. Combining survey and interview methods, we found that providers using a predictive algorithm reported being significantly less uncertain and better able to anticipate, plan and prepare for patient discharge than non-users. The tool helped hospitalists form and develop confidence in their mental models of a novel disease (Covid-19). Yet providers' attention to the predictive tool declined as their confidence in their own mental models grew. Predictive algorithms that not only offer data but also provide feedback on decisions, thus supporting providers' motivation for continuous learning, hold promise for more sustained provider attention and cognition augmentation.

Introduction

Artificial intelligence and machine learning aim to revolutionize healthcare. Success is predicated on integration of artificial intelligence tools into decision making. An excellent model may fail to deliver value to patients and providers if it is not used consistently and effectively. While much prior research has explored the creation and validation of predictive analytics tools,[1] and some research has evaluated the effect of these tools on patient care and providers' efficiency and accuracy,[1] we know little about how predictive algorithms influence providers' experience of their work, and particularly how these tools shape the stress and burden of making important decisions under pressure and uncertainty.

Machine learning-based predictive tools operate in a social context where a human decision maker accepts or rejects recommendations from a model. They may be especially valuable in stressful, complex and uncertain conditions in which providers lack knowledge-based representations of patient conditions. In such cases, predictive models can serve as the basis for developing advanced decision support tools to improve medical decision making by offering timely and useful information to providers. Understanding the provider experience of working with these models and incorporating them into clinical decision making is a crucial and necessary step to inform future implementation and adopt best practices.

This novel study focuses on the human experience of consuming a model in production and identifying the circumstances under which providers can find value. The target of our study is the clinical deployment of a COVID-19 favorable outcome model into clinical practice. As a global healthcare crisis and extreme event, COVID-19 drew public attention to the stressful, complex and uncertain conditions under which providers make medical treatment decisions. At the onset of the pandemic, providers lacked a knowledge-based representation of the clinical trajectories of patients with COVID-19, complicating their patient care decision-making and thus serving as an ideal context for our study.

In this context, the predictive analytics unit of one large academic healthcare system with five hospitals in New York City, the epicenter of the first wave of the pandemic, developed a predictive scoring tool to help providers identify patients at low risk of adverse events within 96 hours, who could be safely discharged or transitioned to a lower level of care. [2] The tool was made available for providers at the beginning of May to voluntarily add to their patient list or reference in a COVID report and use in making discharge decisions.

Methods and Materials

To study the effect on providers of implementing the algorithm-based predictive scoring tool, we employed a sequential explanatory design, with a phase one quantitative data collection followed by a phase two qualitative data collection. [3] We integrated both survey (quantitative) and interview (qualitative) data so that we could combine the strengths of each modality to answer the 'what' and 'why' questions and derive mutually illuminating findings. [4,5] Integration took the form of analyzing the survey data to inform the subsequent development of the interview protocol and the identification of participants to interview as well as triangulating the findings to explore the reasons behind key relationships identified in the survey. The survey and semi-structured interview protocol were approved by the IRB of both New York University (F2020-3779) and NYU Langone Health (i20-00982).

Phase 1 field survey study design and sample.

In mid-July 2020, we distributed a brief, anonymous online survey to the set of 633 providers who were recruited to treat COVID-19 patients during the first wave of the pandemic. Recruited participants included physicians (making up two thirds of the cohort), nurses, and a small number of social workers and care managers. Respondents included many physicians from a variety of different specialties who do not generally work in the hospital but who were recruited on a short-term basis to help provide care for the huge onslaught of ill COVID-19 patients. Responses were received from 180 hospital staff (28% response rate), nearly all of whom (175) were providers who treated COVID-19 patients. Most (59%) of our respondents were attendings. Respondents self-reported (N=159) whether they used the tool in their clinical decision-making, leading us to classify 38 respondents as tool users (24%).

Respondents reported their perceived uncertainty, discharge planning ability, and confidence in a safe discharge in relation to their COVID-19 patients (see Appendix 1). Predictive scoring tool users and non-users were asked the same questions, allowing us to use analysis of variance to compare the perceptions of users to non-users. This design enabled us to identify the effect of using the tool on providers' work experiences, attitudes and perceptions whether or not providers attribute those effects to their use of the tool. We tested for mediation of the effect of tool use on confidence in a safe discharge using the bootstrapping method developed by Preacher and Hayes, [6] using 5,000 bootstrap resamples.

Phase 2 Semi-structured interview design and sample.

We contacted hospitalists across the facility with the goal of recruiting those with experience caring for COVID-19 patients when the predictive model was available. We employed purposeful sampling to identify and select cases that were "information rich" (Patton, 2002) and to ensure that we recruited both users and non-users of the predictive model. We contacted potential participants by e-mail and requested their participation via telephone or video conferencing interview. We also developed a semi-structured interview guide that included questions on interviewees' perceptions of uncertainty about COVID-19 treatments and disposition (factors considered when making patient discharge decisions), and use or non-use of the predictive scoring tool (see Appendix 2). Under this guide, the interviewer could also explore what informants wanted to discuss. After the 8thth interview, we reached *a priori* thematic saturation, [7] in which we recognized that the previously determined conceptual categories and themes were adequately represented in the data. We continued data collection for two more interviews to confirm that no new themes were emerging. We thus conducted a total of ten interviews, including five predictive model users and five non-users, between October 2020 and February 2021. All interviews lasted between 27 and 48 minutes, with an average time of 37 minutes. All interviews were recorded with the participant's consent, and transcribed verbatim.

We conducted a thematic analysis of interview transcripts at two levels: within each case and across cases. The transcript content was coded and analyzed using NVivo (Version 12), a qualitative data analysis software program. In the first phase of analysis, the transcripts were assigned codes based on the topic areas and questions that comprised the *a priori* interview protocol. In the second, we read all interview transcripts and identified novel themes that

emerged in the data, such as providers' strong desire for learning. We then iteratively refined the inductively derived codes. In a third phase of analysis, we used the process of comparing and contrasting[8] to classify all codes into themes.

Results

In the survey overall sample, predictive scoring tool users experienced significantly lower uncertainty (*mean*=1.92 for users, *mean*=2.34 for non-users) about caring for COVID-19 patients than non-users ($F(1, 148) = 4.155$; $p = .045$), and reported marginally greater discharge planning ability (i.e., the ability to anticipate, plan, and prepare in advance for patient discharge; *mean*=2.75 among users, *mean*=2.36 among non-users) ($F(1, 148) = 3.785$; $p = .054$).

In the subsample of attendings, the same pattern emerged but the effects were stronger. Attendings who used the tool experienced significantly lower uncertainty (*mean*=1.89 for users, *mean*=2.32 for non-users) ($F(1, 77) = 5.081$; $p = .027$), and significantly greater discharge planning ability (*mean*=2.88 for users, *mean*=2.21 for non-users) ($F(1, 77) = 9.419$; $p = .003$).

There was no significant difference between users and non-users on confidence in discharge safety. However, among attendings, confidence in discharge safety was correlated with discharge planning (i.e., ability to anticipate, plan, and prepare in advance for patient discharge) ($r = .298$; $p = .007$). Our mediation analyses using 5000 bootstrapped resamples[6] found a significant indirect effect of tool use on confidence in discharge safety via discharge planning, as indicated by the confidence interval not including zero (*LLCI, ULCI*: -.4967, -.0389). Uncertainty did not significantly mediate the effect of tool use on confidence in discharge safety. In sum, we found that users of the predictive scoring tool had significantly reduced uncertainty and significantly increased perceived ability to anticipate, plan or prepare for patient discharge in advance, and that discharge planning ability explained the relationship between tool use and increased confidence in safe discharge.

The themes emerging from the interviews included the uncertainty faced by providers, the predictive scoring tool's impact on providers, reasons for non-use or discontinuation of use of the predictive scoring tool, and providers' desire for learning. In addition to explaining survey results obtained in the first phase, analysis of the qualitative data complemented those findings by exploring informants' views in more depth and elaborating on seemingly contradictory findings, such as why divergence between the model's prediction and the provider's judgment led to greater interest in the model in the beginning, but decreased reliance on the model over time.

Theme 1: Uncertainty associated with COVID-19 and its impact on providers.

Providers experienced a great deal of uncertainty about diagnosing, treating, and discharging patients with COVID-19, especially at the beginning of the pandemic, and uncertainty had cognitive, emotional, and behavioral implications. In particular, providers initially found it difficult to anticipate patient trajectory and determine adequate treatment regimens for this novel disease. Treating these patients unsettled providers' established mental models for treatment decision-making (theme 1a), and making medical decisions under these circumstances generated a feeling of psychological and emotional strain (theme 1b; see Table 1 for representative quotes for all themes). At the same time, providers felt pressure to discharge recovering patients to make room for the large number of potentially sicker patients not yet admitted, exacerbating the experience of uncertainty and strain (theme 1c).

Theme 2: Predictive scoring tool's impact on providers.

The predictive scoring tool helped providers identify patients at low risk of adverse events by providing a numeric score with higher numbers (colored red) indicating higher risk of adverse events, and lower numbers (orange for moderate, green for low) indicating lower risk. The tool impacted providers in two ways. First, when the tool confirmed their intuition, it helped reduce their uncertainty and stress (theme 2a). Second, when the tool was not aligned with their clinical intuition, the predictive model made providers pause, think deeper, and investigate further, but also led providers to discount the advice of the tool (theme 2b). The majority of the providers interviewed explained that the model's discrepant predictions sparked further investigation and triggered action, such as reassessing patient data or talking to the patient, particularly when the providers judged the patients to be safe for discharge but saw a red flag on the model that predicted the patient to be at high risk of deterioration. Discounting of model advice was reported, primarily when the providers predicted that the patient was neither safe nor ready to be discharged, but the

model registered "safe to discharge." In these cases, the providers often thought that their clinical judgment was better, or that the model overlooked some information.

Theme 3: Reasons for discontinued use, or non-use, of the predictive scoring tool.

Despite the potential value offered by the predictive scoring tool, reasons for non-use included lack of awareness of the tool or information about its validity. Among tool users, our interviews indicated that their use of the tool declined over time. Several factors drove this decline. As providers gained more personal experience in treating the novel disease, they used the model less frequently in discharge decisions because they were finally able to develop a structured mental model of the disease, get a better sense of a typical infected patient's trajectory, and feel more confident about discharge. Reduced tool use also stemmed from paradoxical perceptions. On one hand, they perceived that the tool aligned with their clinical intuition too much, and thus did not provide new and useful information. At the same time, they perceived that the tool disagreed with their intuition too much, leading them to discount the tool's advice.

Theme 4: Driving model use by tapping into provider desire for learning.

The interview findings did uncover a motivation of providers that neither the tool nor any other sources effectively addressed – the desire for feedback and learning. The providers we spoke to made clear that they actively seek feedback to improve their clinical decision-making, and that such actions are viewed as an important part of their professional practice.

Discussion

While many scholars have called for post-implementation studies of the use of predictive analytics tools in healthcare, there is still a paucity of this work.[1] The present research begins to address this gap by exploring how the implementation of a predictive algorithm influenced providers' coping with the challenges of caring for patients during the COVID-19 pandemic. Predictive model post-implementation studies also enable exploration of a neglected target outcome of these models: providers' psychological experience of their work, which is our focus. Our findings suggest that using the predictive tool early in the pandemic alleviated providers' uncertainty and increased planning ability and confidence, thus addressing sources of provider stress.

Advice discounting and algorithmic aversion[8,9] are common expert responses to algorithmic advice and we found some evidence of them in our study, but the uncertainty associated with the novel virus may have reduced these general tendencies at the onset of the pandemic. Our findings suggest that predictive algorithms can be useful in helping providers *create and confirm new mental models*, thus allowing them to cope more effectively with novelty. Our informants seem to have used the predictive model's explainability features to develop a mental model of the risk of adverse events from COVID-19, but did not reliably attribute their mental model development to the model.

The qualitative findings suggest a potential paradox: providers expect the algorithm to agree with their intuition in the vast majority of instances, and confirmation of their clinical intuition is essential to their development of trust in the algorithm. However, when the predictive model is experienced as merely providing information that providers feel they already know, as when it generally confirms their prior clinical judgment, consulting the model is less likely to seem worthwhile. Confirmation bias may subjectively bias providers to more readily recall instances in which the tool affirmed, rather than conflicted with, their prior beliefs. A tool that merely confirms their mental model may not be viewed as worthy of sustained attention given high workloads and competing demands.

At the same time, discrepancies drive attention to the tool, leading providers to feel professionally obligated to invest time and effort to explain the discrepancy and thus reconcile the tool's recommendation with their own intuition. Some have expressed concern that providers may blindly accept the recommendations of predictive algorithms, resulting in algorithms replacing rather than augmenting provider expertise. [10] Our findings conflict with this prediction – our informants report first developing their own judgment independent of the algorithm, and only then comparing their judgment to the algorithmic advice. When discrepancies emerged between the two, they sought additional information. Notably, this pattern preserves clinical autonomy and judgment but can lead to information overload and reductions in provider efficiency. Provider attention was especially motivated when the predictive model suggested that patient

risk was unexpectedly *high,* so providers may have used the model more to determine which patients to keep in the hospital for monitoring, rather than the evaluation the model was designed for of identifying which patients are safe enough to discharge.

Discrepancies may increase both trust and distrust in the model, depending on how they are resolved. When discrepancies are resolved in favor of the model, then they can strengthen provider's trust in and attention to the model. However, this is unlikely for several psychological reasons unrelated to model sensitivity or specificity. First, advice discounting and algorithmic aversion suggest that people are more likely to resolve discrepancies in favor of their own intuition. [11] Our informants confirmed this prediction. The explainability features of the model made it easier for providers to reconcile discrepancies but not to keep score regarding whether they or the model were more likely to be correct. Also, feedback indicating that the model was right and they were wrong is less salient to providers because patients are often kept in the hospital without experiencing adverse events.

Providers were, however, strongly motivated to improve their mental models and thus to learn. Learning from predictive models may be more beneficial than learning from experience because algorithms more accurately incorporate, and thus reflect, the diversity of possible cases and can be updated as conditions shift (e.g., in the case of COVID-19, as new treatments came into use). Informants believed that they would be more likely to continue to reference the model intervention if they were also offered feedback (e.g., what happened to patients they discharged after discharge) that supported their learning as an accompanying intervention to complement the prediction from the model. Future research may be fruitfully directed toward understanding how interventions incorporating predictive tools could also facilitate continuous learning.

Outreach promoting the predictive scoring tool targeted attending providers, who were directly responsible for patient discharge decision-making, rather than the entire care team. Discharging a patient involves care managers and social workers who can play a critical role in preparing for, and sometimes even prompting consideration of, patient discharge. Failing to ensure broad tool use across roles prevented the model from serving as a coordination mechanism enabling care managers and social workers to anticipate and prepare for patient discharge or prompt providers to consider earlier discharges. Future research may consider how predictive model implementation shapes teamwork in addition to the behaviors and experiences of individual providers.

References

1. Wallace E, Uijen MJ, Clyne B, Zarabzadeh A, Keogh C, Galvin R, Smith SM, Fahey T. Impact analysis studies of clinical prediction rules relevant to primary care: a systematic review. BMJ open. 2016 Mar 1;6(3):e009957.
2. Razavian N, Major VJ, Sudarshan M, Burk-Rafel J, Stella P, Randhawa H, Bilaloglu S, Chen J, Nguy V, Wang W, Zhang H. A validated, real-time prediction model for favorable outcomes in hospitalized COVID-19 patients. NPJ digital medicine. 2020 Oct 6;3(1):1-3.
3. Ivankova NV, Creswell JW, Stick SL. Using mixed-methods sequential explanatory design: From theory to practice. Field methods. 2006 Feb;18(1):3-20.
4. Creswell JW, Klassen AC, Plano Clark VL, Smith KC. Best practices for mixed methods research in the health sciences. Bethesda (Maryland): National Institutes of Health. 2011 Aug 1;2013:541-5.
5. Woolley CM. Meeting the mixed methods challenge of integration in a sociological study of structure and agency. Journal of Mixed Methods Research. 2009 Jan;3(1):7-25.
6. Preacher KJ, Hayes AF. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behavior research methods. 2008 Aug;40(3):879-91.
7. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, Burroughs H, Jinks C. Saturation in qualitative research: exploring its conceptualization and operationalization. Quality & quantity. 2018 Jul;52(4):1893-907.
8. Corbin J, Strauss A. Strategies for qualitative data analysis. Basics of Qualitative Research. Techniques and procedures for developing grounded theory. 2008;3.
9. Harvey N, Fischer I. Taking advice: Accepting help, improving judgment, and sharing responsibility. Organizational behavior and human decision processes. 1997 May 1;70(2):117-33.

10. Önkal D, Goodwin P, Thomson M, Gönül S, Pollock A. The relative influence of advice from human experts and statistical methods on forecast adjustments. Journal of Behavioral Decision Making. 2009 Oct;22(4):390-409.
11. Rubin DL. Artificial intelligence in imaging: the radiologist's role. Journal of the American College of Radiology. 2019 Sep 1;16(9):1309-17.
12. Bonaccio S, Dalal RS. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. Organizational behavior and human decision processes. 2006 Nov 1;101(2):127-51.

Table 1. Representative quotes for all themes

Theme 1a: Physician uncertainty was exacerbated by the absence of mental models for decision-making about COVID-19 patients (cognitive effects)	
AP09	*In the first wave, we didn't have much in the way of treatment algorithms… We had a lot of theories running around as to what we could give. But we didn't have any specific evidence as to what would work.*
AP01	*It was difficult to know how to sometimes interpret inflammatory markers that weren't always heading in the same direction. Some might be going up and others might be going down. So it was hard to know who was necessarily actually getting better or worse.*
Theme 1b: Uncertainty associated with treating and discharging COVID-19 patients was psychologically taxing for providers (emotional effects)	
AP05	*There was so much that was unknown. It was a lot of uncertainty. We didn't necessarily know exactly which patients would be at risk for deterioration. . . . That was probably the hardest part.*
AP02	*I think the first week was pretty traumatizing with the fact that no one knew what was going on.*
AP01	*Especially early on, I felt pretty uncomfortable discharging patients.*
Theme 1c: Physicians felt pressure to discharge to serve other sick patients (behavioral effects)	
AP05	*If this person doesn't need to be here, then we could potentially use these resources, this space, for someone else who might need it.*
Theme 2a: The predictive scoring tool alleviated stress by frequently confirming providers' clinical intuition	
A08	*If they have a low score then I might feel better about discharging them.*
AP04	*I would use it to confirm what I already thought…[and] see if the color matched my thought.*
Theme 2b: Discrepancies between clinical intuition and the tool triggered further investigation and advice discounting	
AP02	*[the model] makes me mentally go through the reasons why I don't agree with it, and then kind of confirm that.. if it said they were high risk for a severe complication, I might pause and really talk to them and make sure they're really feeling well, maybe have them walk, and just really go through my mind of why I think my assessment doesn't correlate with that analytic.*
AP02	*If someone was doing bad and that [model prediction] was good, obviously I'm going to keep them in the hospital.*
Theme 3: Physicians who were confident in their mental model for treating COVID-19 patients discontinued use of the tool or never used it	
AP06	*If a model disagrees with you too much, especially for cases when it's really obvious, you would look at that and think, what in the world is this? It becomes meaningless after a while. If a model like that agrees with you too much, it's kind of the same. It becomes meaningless after a while.*
AP07	*By April [2020], I think I probably had a big enough sample to start to feel more confident, in discharging people a little more aggressively based on how they looked.*
Theme 4: Physicians' desire for feedback to learn would motivate sustained use of the tool	
AP01	*I think it probably would have been helpful for me to know if I had discharged somebody who had a red marker, a red indicator, and they ended up being readmitted. . . . I never was able to follow up with it to see if anyone got readmitted.*