Schizophrenia Bulletin
https://doi.org/10.1093/schbul/sbac038

Improving the Applicability of AI for Psychiatric Applications through Human-in-the-loop Methodologies

Chelsea Chandler*[,1,2], Peter W. Foltz[2], and Brita Elvevåg[,3,4]**

[1]Department of Computer Science, University of Colorado Boulder, Boulder, CO, USA; [2]Institute of Cognitive Science, University of Colorado Boulder, Boulder, CO, USA; [3]Department of Clinical Medicine, University of Tromsø – the Arctic University of Norway, Tromsø, Norway; [4]Norwegian Centre for eHealth Research, University Hospital of North Norway, Tromsø, Norway

*To whom correspondence should be addressed; 430 UCB, 1111 Engineering Dr., Boulder, CO 80309, USA; tel: 703-895-4764, fax: 303-492-7177, e-mail: chelsea.chandler@colorado.edu

**To whom correspondence should be addressed; Postbox 6124, Tromsø 9291, Norway; e-mail: brita.elvevag@uit.no

ABSTRACT

Objectives: **Machine learning (ML) and natural language processing have great potential to improve efficiency and accuracy in diagnosis, treatment recommendations, predictive interventions, and scarce resource allocation within psychiatry. Researchers often conceptualize such an approach as operating in isolation without much need for human involvement, yet it remains crucial to harness human-in-the-loop practices when developing and implementing such techniques as their absence may be catastrophic. We advocate for building ML-based technologies that collaborate with experts within psychiatry in all stages of implementation and use to increase model performance while simultaneously increasing the practicality, robustness, and reliability of the process.**

Methods: **We showcase pitfalls of the traditional ML framework and explain how it can be improved with human-in-the-loop techniques. Specifically, we applied active learning strategies to the automatic scoring of a story recall task and compared the results to a traditional approach.**

Results: **Human-in-the-loop methodologies supplied a greater understanding of where the model was least confident or had knowledge gaps during training. As compared to the traditional framework, less than half of the training data were needed to reach a given accuracy.**

Conclusions: **Human-in-the-loop ML is an approach to data collection and model creation that harnesses active learning to select the most critical data needed to increase a model's accuracy and generalizability more efficiently than classic random sampling would otherwise allow. Such techniques may additionally operate as safeguards from spurious predictions and can aid in decreasing disparities that artificial intelligence systems otherwise propagate.**

Key words: machine learning/natural language processing/active learning/safeguards

INTRODUCTION

The notable success of machine learning (ML) and natural language processing (NLP) (see Table 1 for a glossary of technical terms used in this paper) in characterizing aspects of mental disorders has made an enormous impact in psychiatric research and speculations regarding the future of clinical decision making. Artificial intelligence (AI), which encompasses both ML and NLP, is capable of learning subtle and nuanced features and patterns of language and behavior. This is of particular interest in psychiatry where *what* patients say and *how* patients speak is a core component in clinical evaluation since symptoms and signs often emerge via speech. Deviations from "normal" speech (eg, less coherence, irregular part of speech use) are detectable with NLP methods. These techniques have been leveraged to predict diagnostic groups,[1–8] fluctuations in patient state,[9,10] patient affect,[11–14] thought disorder severity,[15–18] among others. While these methods can uncover important signals in language and behavior for mental health applications, the field is still in its infancy for clinical applications. A general lack of model generalizability, transparency, and explainability may further limit the field from achieving its full translational potential.[19]

Today, most ML applications use supervised learning techniques: models learn patterns from labeled data and generalize this knowledge to new data, similar to how clinicians might learn to associate symptom patterns to medical conditions.[20] This contrasts with expert rule-based AI systems, where models comprise a series

Table 1. Glossary of technical terms

Artificial Intelligence (AI)	AI is a general term for computer systems that exhibit intelligent behavior and are able to learn, explain, and advise their users.
Machine Learning (ML)	ML is a subset of AI that harnesses statistical algorithms to learn *features* of data and the associated importance of each (or simply the associated importance of user-defined features). Once the features and their weights, as well as other *hyperparameters,* are set, the model can predict some outcome or clinical classification on new, unseen data.
Natural Language Processing (NLP)	NLP is another type of AI that incorporates both statistical and linguistic knowledge to understand human language.
Features	A measurable property of the data (eg, the number of words spoken is a *feature* that can be measured from natural language data).
Hyperparameters	Parameters that are set before the final training of a model rather than learned during the process (eg, the number of iterations of training used to train a model).
Classification	A category of ML models that are trained to predict a category. It can be binary (e.g., mentally ill or healthy) or multiclass (eg, classifying speech as "schizophrenic-like", "manic-like", or non-disordered).
Regression	A category of ML models that are trained to predict a numerical, continuous-valued output (eg, a clinical rating).
Neural Network	A type of ML model that is a system of nodes, composed in layers. Each node learns some nonlinear equation on a subset of training data and when all are combined, a categorical or real valued output can be predicted. Modern neural networks have hundreds to thousands of nodes and layers and are trained on large datasets.
K-Means Clustering	An unsupervised method of partitioning a dataset into K clusters with the goal of minimizing within-cluster variance. Each data point belongs to a single cluster determined by its nearest mean or centroid.
Edge Case	Any situation that occurs near a decision boundary, at the extremes of the inputs, an exception to a learned rule, or anything that may require additional or special handling.
Overfitting	A situation where a model too closely fits its training data. This is an issue because it may learn to fit to spurious correlations in the data rather than learning a generalized solution to the problem itself.

of rules to produce labels for new data. Since supervised learning models are trained on human-derived labels (such as diagnoses or clinician ratings), the humans' role is primarily in writing the code that finds the best models and parameters to maximize predictive performance (see Refs. [21,22] for an overview). The black box nature and lack of human involvement in many of these techniques increases risk of spurious predictions and societal biases that inflate disparities in diagnosis, monitoring, and treatment.[23–26] Without additional considerations and measures, these models can range from highly limited to critically dangerous in implementation.

Previously, we have followed traditional ML approaches to assessment in two distinct studies of the automatic scoring of story recall in healthy participants, and patients with (1) affective disorders[27] and (2) serious mental illness[28] (see Refs. [27,28] for details of the modeling process). Starting with fully annotated datasets (N =1177 and N = 846 story recall responses, respectively), NLP features which measured the amount and relevance of semantic content produced in the recalls were extracted and statistical analyses selected the most predictive ones. K-fold cross-validation was used to iteratively train and test regression models. Once the best features, model types, and parameters were determined, accuracy statistics were reported and published (average Pearson correlation with human ratings = 0.88 and 0.83, respectively), and the experiments ended. Despite being successful in terms of correlations obtained, the path to implementation remained unclear. Five key stages of the traditional ML framework applied to story recall scoring are shown in the top row of figure 1; the bottom row includes additional stages we advocate for.

This common ML framework has been applied widely in clinical assessment research (eg, [1–18]) and has proved useful for understanding how language features combine to model clinical judgments. However, several pitfalls can lead to incorrect model decisions. First, it does not allow for sufficient human understanding of the patterns of input data which may lead to model uncertainty, or how the model will handle such data. Second, effort is often spent on collecting and labeling fairly homogeneous data rather than the more diverse data needed to create a robust model.[29] Third, only rarely is an evaluation set put aside and not touched until a final model is trained and tuned, especially in the case of small datasets.[30] Thus, measures reported are likely biased to numbers generated *for* those datasets, such that models optimized for the entire dataset will be inflated when tested on the same data. Finally, it is assumed that models trained on specific locations and demographic groups can be applied elsewhere with similar results, yet ML models rarely transfer to new data as expected. This traditional framework may be sufficient for applications where there are no life-altering decisions involved, but

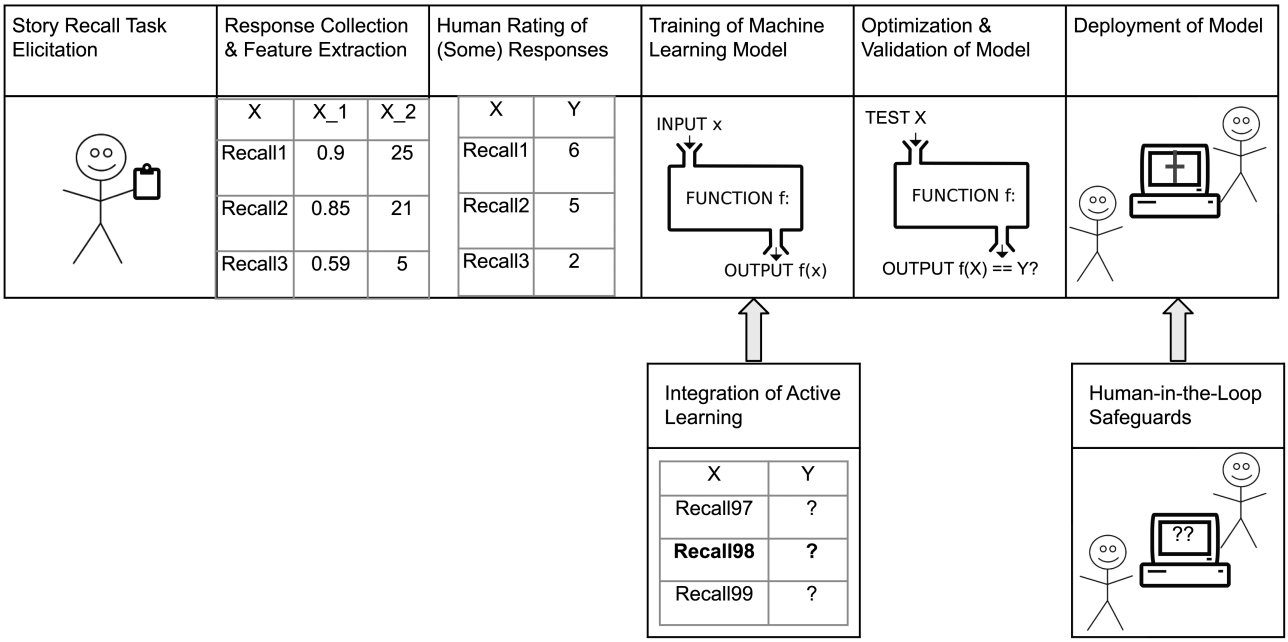

Fig. 1. Top row: stages of the traditional ML framework applied to story recall scoring with minimal human involvement. Bottom row: active learning and human-in-the-loop safeguards which must be incorporated for translational value.

for high stakes assessments that must be robust and reliable, this is inadequate.

We argue that traditional ML approaches must be supplemented to improve their generalizability, transparency, reliability, and applicability for the translation of clinical research into mainstream assessment. While there exist a range of methodologies that are applicable for solving this problem, we advocate specifically for the incorporation of a human-in-the-loop framework. These techniques involve humans in additional stages of the model development process, enabling understanding of how illness dimensions (eg, different stages, symptoms) impact model predictions. Furthermore, they are well-suited for modeling dynamic, continuous real-world data that comprise multiple channels. From the clinical perspective, this framework is appealing as it generates models that are as robust and reliable as the data allows, supports understanding gaps in model knowledge, and can alert to unusual inputs before spurious decisions are generated.[31]

Human-in-the-Loop AI

Human-in-the-loop is a general term for processes that enable collaboration between humans and machines. This paper focuses on the application of active learning, a technique employed during model training where the most essential data is chosen for human labeling as early as possible (see figure 2). It begins with the collection of labeled and unlabeled data, and a partially trained model that will improve with iterations of active learning. Two popular active learning sampling techniques—uncertainty and diversity sampling—allow us to label *only* data that are necessary to make a model more certain in areas of uncertainty and fill general knowledge gaps, thus improving efficiency (ie, requiring less labeled data to reach a given accuracy) and model robustness. These sampling techniques should be used in conjunction with random sampling to avoid biasing the model towards learning how to generate predictions on mostly rare examples. Once the chosen data are labeled, the model is retrained, and the process repeats until accuracy gains stabilize or a criterion is reached. We discuss how this is incorporated into the labeling of data and training of a model, and showcase the approach in the story recall task (with in-depth detail in Appendix A, supplementary material and implications of applying this framework—as well as additional techniques—within development and deployment in Appendix B, supplementary material).

Uncertainty Sampling: the Known Unknowns

Uncertainty sampling is used to understand where the model is least confident; finding items *the model knows it does not know.* The goal is to discover unlabeled items closest to decision boundaries in the trained model, obtain their labels, and increase certainty. In a clinical application, uncertainty sampling allows the model to become more confident in cases that exist near decision boundaries or in the extremes of the feature space, whether these are in diagnosis of mild symptoms, early stages of illness, or where patients are transitioning into decline. In a simple binary classification task with two labels, the most uncertain examples are those that the model assigns ~50% probability of the example belonging to either class. These items are the most likely to be incorrectly classified—which can be life-threatening, or life-altering—so the goal is to strengthen confidence around a decision boundary to compensate for this.

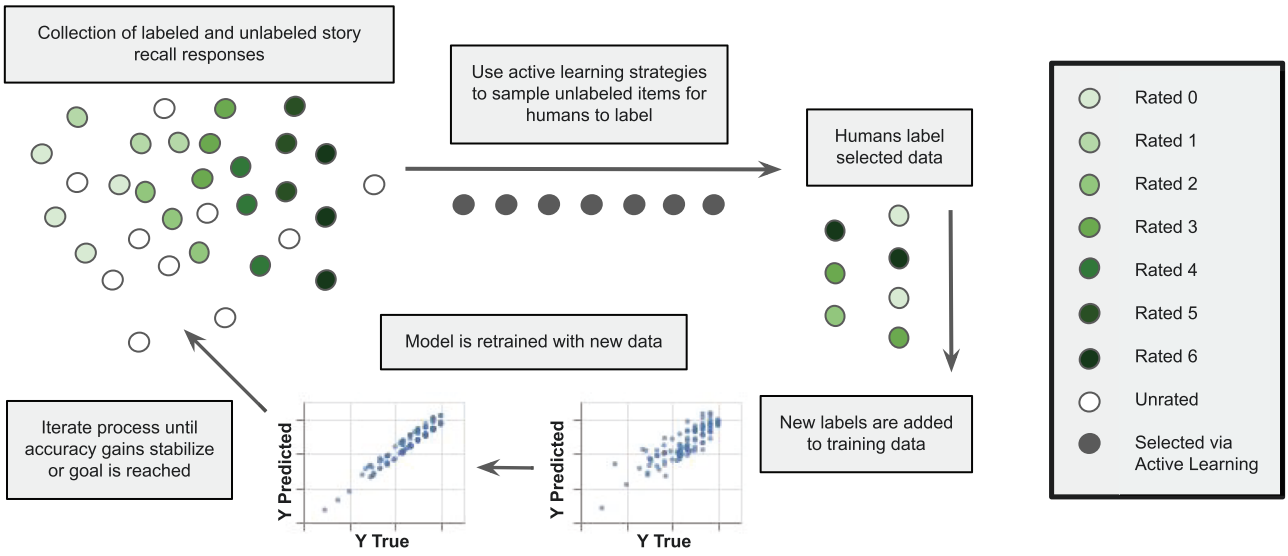

Fig. 2. Active learning process for story recall scoring. Adapted from Human-in-the-Loop Machine Learning.[31]

In uncertainty sampling, items with the lowest confidence in the model are sampled. For instance, if a model predicts whether a speech segment is more y_1: "schizophrenic-like" (loose associations), y_2: "manic-like" (flight of ideas) (The terms are used as in the classic distinction differentiating speech displaying 'loose associations' versus a 'flight of ideas'. These distinctions were formalized in models of language processing to characterize the manner in which incoherence differs in manic versus schizophrenic speech, with the former shifting 'from one coherent discourse structure to another' and the latter being deficient in 'any discourse structure'[32] [p. 831]) or y_3: nondisordered, the output is a vector of size 3 where each number corresponds to the predicted confidence of each label. In the case that the output is $[0.6,\ 0.3,\ 0.1]$ the model assigned 60% probability the speech is "schizophrenic-like", 30% probability it is "manic-like", and 10% probability it is nondisordered. Each unlabeled item in the dataset can be rank-ordered by predicted confidence and lowest ranked items are prioritized for labeling. Several criteria can determine this ranking: (1) *least confidence* criteria determines, for each unlabeled item, 1 minus the highest confidence label (in this example, the maximum confidence is 0.6, so $1 - 0.6 = 0.4$). (2) *Margin of confidence* ranks the unlabeled items by the difference between the two most confident labels (in this example, the margin of confidence would be $0.6 - 0.3 = 0.3$). (3) *Ratio of confidence* ranks the unlabeled items by the ratio between the two most confident labels (in this example, the ratio of confidence would be $0.6/0.3 = 0.2$). (4) *Entropy-based sampling* measures the difference between all predictions to derive a measure of how much every confidence differs from another. Each of these approaches ensures that borderline cases are included in the training data, thus strengthening decision boundaries.

The aforementioned sampling techniques are not appropriate for continuous-valued regression tasks. (5) *Query by committee* can be applied in both classification and regression tasks.[33] It is an ensemble-based approach where variations of ML models are trained with differing data subsets or hyperparameters to evaluate the extent to which varied models may disagree. The more disagreement, the more uncertainty around the input. This approach is common in neural networks, where a single unlabeled example is passed through the network multiple times, each iteration with a section of the network dropped out[34] to reveal whether more prediction disagreement occurs in certain problem space areas, features, or demographics. Harnessing model variations to quantify uncertainty is analogous to comparing human opinions: cases where clinicians agree on some label may imply the symptoms in question align more with a "textbook example" and are thus easier to label. In contrast, cases where clinicians make different judgments would indicate that this particular example is abnormal, and including such examples with definitive labels in a training set would strengthen future predictions for such rarities.

Diversity Sampling: the Unknown Unknowns

Diversity sampling is used to understand areas in the dataset that are underrepresented in the model. Sampling items for diversity provides models a more complete picture of the data and feature space. Knowledge gaps are usually areas where feature values are rare, or constitute underrepresented real-world demographics. This is useful in clinical settings where understanding demographic intricacies is crucial for generating valid group predictions. Differences in language and behavior inherently exist between demographic groups, and if not

well-represented in the model, it *will not be able to make accurate predictions in these cases.* This is furthermore a reason that each model must be transparent with respect to assumptions used in their creation—ie, if a model was trained only on data from those who speak English as their first language, it must be made explicit that the model cannot be expected to accurately predict variables of interest for those who learn English as a second language.

Importantly, this sampling can be done with and without a trained model. The first approach that harnesses a trained model is (1) *model-based outlier sampling* which seeks to understand examples that are currently unknown to the model by finding individual feature values or combinations that have not yet been encountered in training. In neural networks, this can be done by investigating neuron activations as each unlabeled example is passed through the network since neurons with the most activation tend to be those that the network has more information about. Activations are ranked and examples that produced the smallest activations are chosen as outliers. For clinical data, this entails sampling data from patients whose symptoms do not align with what is represented in the training data. Model-based outlier sampling can overemphasize a model's statistical biases and thus choose similar outliers each time; hence, the importance of investigating diversity sampling approaches that are not model based.

The next approaches operate distinctly from a trained model. (2) *Cluster-based sampling* divides the data into unsupervised clusters and samples an equal number of examples from each, allowing an even spread of data types and ensuring the training set does not over-represent single areas. Here, for each cluster (created by a standard algorithm such as k-means), sampling is both random and from the centroids and outliers. If clusters match underlying symptoms in a clinical group, the model will be evenly trained on each group. (3) *Representative sampling* finds data most similar to the data in real-world applications,[35] and is achieved in an adaptive manner where one example is chosen per iteration so as to not choose a full set of similarly representative examples. This allows the model to more fully address disparities between the training data and real-world application. If the model is already fully trained on representative data, no further examples are needed for this criterion to be reached. Finally, (4) *sampling for real-world diversity* can take many forms. The model must take into account any meaningful data characteristic that affects model performance in certain contexts. These characteristics could be variables such as race, language, location, gender, socioeconomic status, and education, with the potential that these biases can inflate in combination. For instance, race in certain locations skews data more heavily than in other locations (for an overview on bias in NLP see Refs. [23,36]). Measuring and reducing real-world bias is complex, but various approaches can minimize the impact, such as applying all active learning approaches *stratified over each demographic,* thus sampling a wide range of examples for each group.

It is well-known that ML models amplify biases inherent in data.[37] Clinical data in particular is heavily skewed toward the most highly represented demographics: participants tend to be from western, highly educated, industrialized, rich, and democratic areas.[38] If random sampling is used to gather such data, the model will be heavily biased as more examples will naturally be drawn from dominant classes. Thus, increasing diversity in datasets will also increase access for more target populations once implemented. For verifying robustness to various demographics during training, researchers should calculate macro accuracy statistics per demographic. Specifically, the evaluation dataset should include the full range of demographics of interest. Statistics such as minimum score by demographic or harmonic mean over all demographics can verify consistency in predictions. If accuracy is disproportionately lower in certain populations, effort must be made to add more data from those populations, or explore approaches such as synthetic data generation (Appendix B, supplementary material).

METHODS

We demonstrate an implementation of a human-in-the-loop ML framework applied to an automated story recall assessment model. Evaluation of human verbal memory is a critical component of establishing neurocognitive function in psychiatry, and arguably has some similarities to the anamnesis process of medical history taking, where the clinician asks questions to probe the patient to recall information to facilitate the diagnosis process. Given its importance, it is a core component of the Wechsler Memory Scale,[39] where the Logical Memory subtest requires the patient to repeat short stories immediately after they have been spoken by the examiner, and after a delay. Automation of such an assessment may enable more regular and/or remote assessments, which may be beneficial for longitudinal monitoring of patient health.

For this experiment, we harness a dataset of 846 labeled responses from 79 healthy participants and 23 participants with affective disorders. Responses were scored on a scale of 1–6. Active learning is simulated by ignoring labels until responses are chosen to demonstrate how a superior model can be obtained more efficiently than can be done with random sampling or by labeling all of the data. This simulates the human-in-the-loop process as the algorithm picks which data is to be labeled by a human at each iteration, rather than requiring it all to be labeled initially. We randomly sampled 100 training responses (*training set*), 100 for tuning parameters (*validation set*), and 100 for obtaining accuracy measures (*evaluation set*;

sampled with stratified sampling so as to achieve an equal spread of ratings and demographics). Creating three subsets of 100 responses is an arbitrary choice, and can be implemented in many other ways.

What remains constant in all approaches is that the evaluation set is put aside until the model is fully trained and tuned, such that evaluation is not biased or overfitted. It is good practice to generate multiple evaluation datasets. One strategy is to create two: one drawn from the same dataset as the training and validation and one drawn from a different source. The out-of-domain dataset enables evaluation of model generalization to the problem, rather than to particular idiosyncrasies of the training set. In addition to the evaluation set sampled from the original 846 responses, we also tested the active learning approach on a separate dataset of 1177 responses collected from 120 healthy participants and 105 participants with serious mental illness. With this testing approach, it becomes clear whether the model has suffered from overfitting. (A popular example from computer vision is that of a model thought to distinguish huskies from wolves.[40] The model learned to identify wolves by learning that photos taken with snow were usually wolves. It never learned to distinguish between the animals, but rather learned an idiosyncratic correlation in the data. From a clinical perspective this is analogous to a model trained to predict whether a patient has schizophrenia. Due to the dataset comparing highly educated healthy controls to less educated schizophrenia patients, the model learns to predict by education artifacts rather than disease symptoms.)

Active learning starts with a minimum viable product: here it is a ridge regression model trained on the initial small training set. This model achieved a 0.66 Pearson correlation to human ratings when tested on the responses in the evaluation set. Each active learning iteration sampled 100 additional items: 45% via uncertainty sampling, 45% via diversity sampling, and 10% via random sampling. These percentages must be tuned based on the application. For example, if the problem is unbalanced (ie, majority class examples outnumber minority class examples) and more samples from minority classes are needed, it is better to oversample from diversity sampling as majority class outliers are less likely to be of interest.

For each sampling approach, we implemented the best-suited techniques for a regression model (see Appendix A, supplementary material for a detailed overview of the first iteration of active learning). Figure 3 shows data selected in the first iteration, where uncertainty sampling (*query by committee*), diversity sampling (*clustering*: centroids, outliers, random, and *real world*), and *random sampling* were applied. These 100 examples were "labeled", added to the training set, and the regression model was retrained. This process was repeated five times until all remaining responses (N = 646) were included in the training set.

RESULTS

As a result of the first round of active learning, the Pearson's correlation on the evaluation set increased from an initial 0.66 correlation to 0.78. For comparison, this was replicated with 100 random samples and resulted in a 0.69 Pearson correlation. Figure 4 shows the relative increase in correlation with the held out evaluation set as active learning progressed (blue lines with circles overlaid; solid for active learning and dashed for random sampling). The model required 200 training examples to reach a correlation of 0.78 with human labels, while an approach with random sampling reached this correlation only after the incorporation of 500 responses. Results were similar when testing both approaches on an external dataset, labeled "Transfer" and depicted with orange lines without overlaid circles in figure 4. As the external dataset was collected from a different population, had a different distribution of feature values and ratings, and had a lower inter-annotator agreement, it required two rounds of active learning to reach its maximum accuracy. Nevertheless, when more critical data are chosen to train the model earlier, significantly less data are needed to reach a given accuracy.

Human-in-the-loop strategies may more generally apply as safeguards to investigate model stability and robustness when deployed in new situations (Appendix B, supplementary material). We simulated a deployment of the story recall model to the Transfer set. Implementing the *query by committee* technique, six regression models were trained with varied parameters and subsets of training data. Ranges in the six predicted values (maximum − minimum predicted rating) were computed and responses were ranked accordingly. The two most uncertain responses were from the same participant who had profanity and violence in their response. The responses, however, were verbose and fairly on topic to the prompt, making the feature distributions highly irregular. Both responses had an expert score of 1 (the lowest score); however, the fully trained model scored both between 3 and 4 points. After the two most uncertain responses, uncertainty levels severely dropped and no other responses from the Transfer set were deemed *highly* unusual from the data the model was trained on (as the task is fairly constrained). As a means to understand how the model performs in different levels of certainty, the top and bottom 10% of responses with respect to their prediction ranges were extracted and tested with the fully trained model. The 10% with the lowest range (most certain) resulted in a correlation with the human ratings of r = 0.90 and the 10% with the highest range (most uncertain) correlated r = 62. Thus, this method is able to be harnessed as a means to alert clinicians when not to rely on model predictions. *Real world* sampling was also implemented at this stage to flag responses from humans whose demographics were not fully covered in the training data.

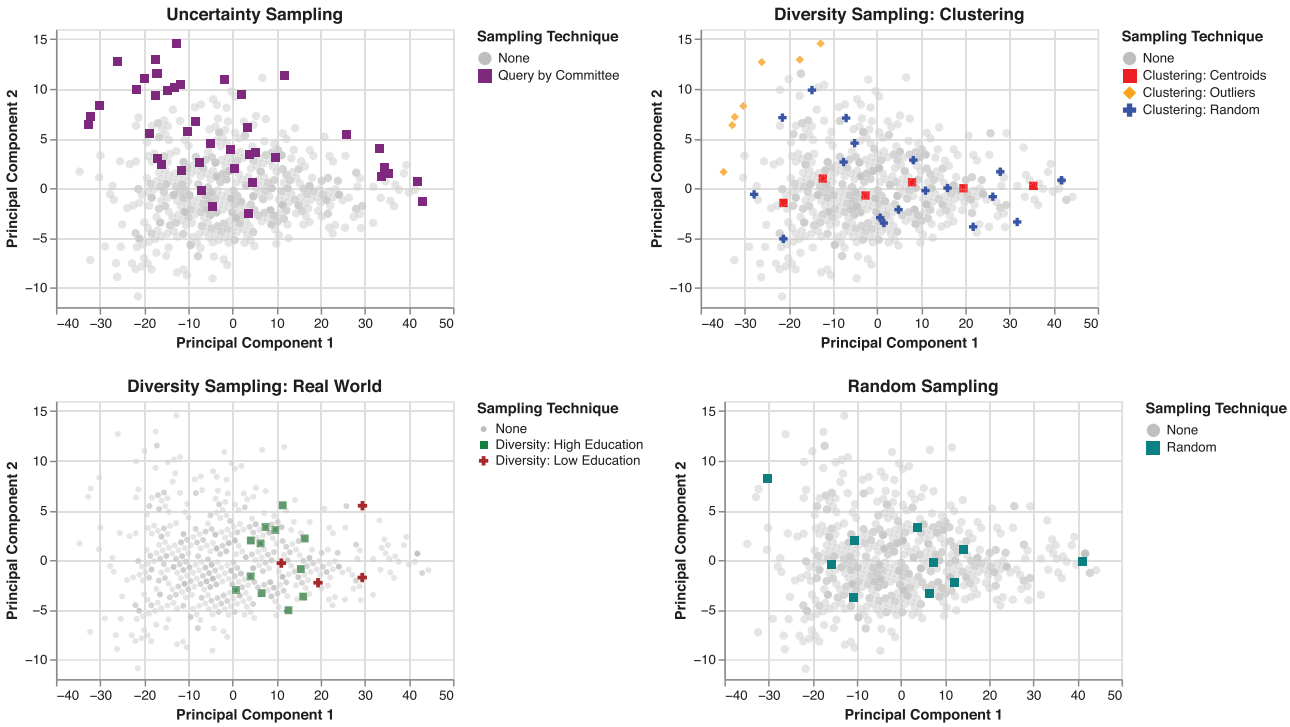

Fig. 3. Scatter plot of unlabeled story recall responses chosen via uncertainty ($N = 45$), diversity ($N = 30$ clustering, $N = 10$ real world), and random ($N = 10$) sampling in the first active learning iteration, visualized with the first two components of PCA.

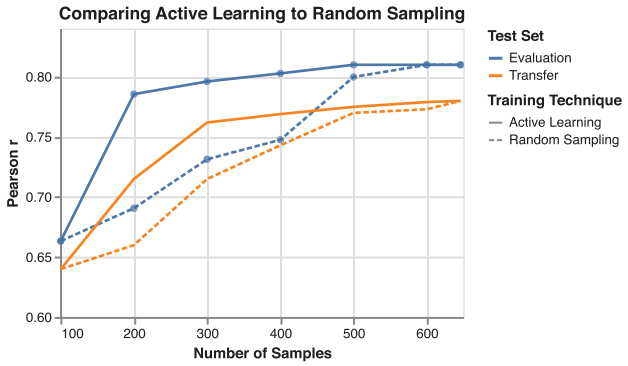

Fig. 4. Line plot of story recall model prediction and human rating correlations when harnessing active learning (solid lines) versus random sampling (dashed lines), testing on the evaluation set (blue lines with overlaid circles) and the transfer set (orange lines without overlaid circles). Both approaches continue until the full dataset is harnessed and correlations converge. For interpretation of the references to color in figures, the reader is referred to the web version of this article.

Discussion

Human-in-the-loop approaches will not solve every ML pitfall. They remain limited in the same manner as traditional approaches by the quality of a dataset. If a problem is ill-formed, or chosen features do not predict variables of interest with high accuracy, human-in-the-loop techniques will not make up for this. While diversity sampling can alert humans to discrepancies in the model's representation of demographic groups, it will never solve the issue of missing data. Furthermore, the implementation of evaluation metrics such as test-retest reliability and divergent validity[41] remain critical for garnering trust in these approaches.

The creation of AI methods for psychiatric applications requires the involvement of *many* humans. Whether that is researchers conceptualizing a computerized solution, engineers developing a model, or clinicians incorporating such a model in the clinic, it should be clear that the absence of the "human" from the "loop" would be catastrophic. Here, we proposed a specific type of human-in-the-loop ML, where best practices are harnessed to create the most robust and reliable model possible. These methodologies—now widespread in many areas of AI—can apply to various types of data and experimental studies. While there has been much success in the analysis of structured speech samples for clinical applications, the human-in-the-loop framework offers equal, if not more, power in applications with less constraint and more freedom for diverse data. This is especially the case in free speech, where a wide range of task types exist (eg, semantic fluency, story recall, process questions, open-ended prompts, and unprompted language including a person's emails, texts, or day-to-day speech). As task constraints are lifted, more diverse data are generated and more safeguards must be incorporated.

This paper is intended to serve as a framework that researchers and clinicians can follow for the creation and deployment of reliable models that incorporate implementation safeguards to protect patients from spurious model predictions. We must move beyond the traditional approaches to implementing ML models as the results tend to be brittle and inappropriate for clinical implementation. The way forward must integrate a robust framework placing control in the hands of the human and providing safeguards for the users. Hence, incorporating human-in-the-loop methodologies in clinical practice will result in humans remaining accountable in the decision making process. The future of AI in psychiatry is not that of computer *or* human, but rather will harness the best of both while maximizing explainability, minimizing bias, and keeping algorithmic accountability in the hands of the human.

Supplementary Material

Supplementary material is available at *Schizophrenia Bulletin*.

Acknowledgment

The authors have declared that there are no conflicts of interest in relation to the subject of this study.

References

1. Bedi G, Carrillo F, Cecchi GA, et al. Automated analysis of free speech predicts psychosis onset in high-risk youths. *NPJ Schizophr.* 2015;1:15030.
2. Corcoran CM, Carrillo F, Fernández-Slezak D, et al. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry* 2018;17(1):67–75.
3. Iter D, Yoon J, Jurafsky D. Automatic detection of incoherent speech for diagnosing schizophrenia. In: *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*; June 5, 2018, New Orleans, LA: Association for Computational Linguistics; 2018:136–146.
4. Corcoran CM, Mittal VA, Bearden CE, et al. Language as a biomarker for psychosis: a natural language processing approach. *Schizophrenia Res.* 2020;226:158–166.
5. Mota NB, Copelli M, Ribeiro S. Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *NPJ Schizophr.* 2017;3(1):1–10.
6. Voppel A, de Boer J, Slegers F, Schnack H, Sommer I. S136. Classifying schizophrenia using phonological, semantic and syntactic features of language: a combinatory machine learning approach. *Schizophr Bull.* 2020;46(Suppl 1):S87–S87. doi:10.1093/schbul/sbaa031.202.
7. Rezaii N, Walker E, Wolff PA. machine learning approach to predicting psychosis using semantic density and latent content analysis. *NPJ Schizophr.* 2019;5(1):9. doi:10.1038/s41537-019-0077-9.
8. Tang SX, Kriz R, Cho S, et al. Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. *NPJ Schizophr.* 2021;7(1):25. doi:10.1038/s41537-021-00154-3.
9. Cohen AS, Fedechko TL, Schwartz EK, et al. Ambulatory vocal acoustics, temporal dynamics, and serious mental illness. *J Abnorm Psychol.* 2019;128(2):97–105. doi: 10.1037/abn0000397.
10. Chandler C, Foltz PW, Cohen AS, et al. Machine learning for ambulatory applications of neuropsychological testing. *Intelligence-Based Med.* 2020;1-2:100006. doi: 10.1016/j.ibmed.2020.100006.
11. Cohen AS, Mitchell KR, Docherty NM, Horan WP. Vocal expression in schizophrenia: less than meets the ear. *J Abnorm Psychol.* 2016;125(2):299–309. doi:10.1037/abn0000136.
12. Cohen AS, Renshaw TL, Mitchell KR, Kim Y. A psychometric investigation of "macroscopic" speech measures for clinical and psychological science. *Behav Res Methods.* 2016;48(2):475–486. doi:10.3758/s13428-015-0584-1.
13. Cheng J, Bernstein J, Rosenfeld E, et al. Modeling Self-Reported and Observed Affect from Speech. Proc. *Interspeech* 2018;365:3–3657. doi:10.21437/Interspeech.2018-2222.
14. Chandler C, Foltz PW, Cheng J, et al. Predicting self-reported affect from speech acoustics and language. In Kokkinakis D, Fors KL, Themistocleous C, Antonsson M, Eckerström M, eds. *Proceedings of the LREC 2020 Workshop on: Resources and Processing of Linguistic, Para-linguistic and Extra-linguistic Data from People with Various Forms of Cognitive/Psychiatric/Developmental Impairments (RaPID-3)*. Paris, France: European Language Resources Association (ELRA). 2020b:9–14. https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/RaPID3book.pdf
15. Elvevåg B, Foltz PW, Weinberger DR, Goldberg TE. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophr Res.* 2007;93(1-3):304–316.
16. Pauselli L, Halpern B, Cleary SD, Ku BS, Covington MA, Compton MT. Computational linguistic analysis applied to a semantic fluency task to measure derailment and tangentiality in schizophrenia. *Psychiatry Res.* 2018;263:74–79. doi:10.1016/j.psychres.2018.02.037.
17. Ku BS, Pauselli L, Covington MA, Compton MT. Computational linguistic analysis applied to a semantic fluency task: A replication among first-episode psychosis patients with and without derailment and tangentiality. *Psychiatry Res.* 2021;304:114105. doi:10.1016/j.psychres.2021.114105.
18. Sarzynska-Wawer J, Wawer A, Pawlak A, et al. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Res.* 2021;304:114135. doi:10.1016/j.psychres.2021.114135.
19. Chandler C, Foltz PW, Elvevåg B. Using machine learning in psychiatry: the need to establish a framework that nurtures trustworthiness. *Schizophr Bull.* 2020;46(1):11–14. Doi:10.1093/schbul/sbz105.
20. Kim NS, Ahn WK. Clinical psychologists' theory-based representations of mental disorders predict their diagnostic reasoning and memory. *J Exp Psychol Gen*. 2002;131(4):451–76. PMID: 12500858.
21. Grzenda A, Kraguljac NV, McDonald WM, et al. Evaluating the machine learning literature: a primer and user's guide for psychiatrists. *Am J Psychiatry*. 2021;178(8):715–729. doi: 10.1176/appi.ajp.2020.20030250. PMID: 34080891.
22. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision

support systems: benefits, risks, and strategies for success. *NPJ Digit. Med.* 2020;3:17. doi: 10.1038/s41746-020-0221-y.

23. Hitczenko K, Cowan HR, Goldrick M, Mittal VA. Racial and ethnic biases in computational approaches to psychopathology. *Schizophr Bull.* 2021;48(2):285–288. doi: 10.1093/schbul/sbab131.
24. Strickland EK. IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectr.* 2019;56:24–31.
25. Guo LN, Lee MS, Kassamali B, Mita C, Nambudiri VE. Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection-A scoping review. *J Am Acad Dermatol.* 2021;10:S0190-9622(21)02086-7. doi: 10.1016/j.jaad.2021.06.884. Epub ahead of print. PMID: 34252465.
26. Oliva J. *Dosing Discrimination: Regulating PDMP Risk Scores*. 110 California Law Review (forthcoming 2022). 2021. Available at SSRN: https://ssrn.com/abstract=3768774 or http://dx.doi.org/10.2139/ssrn.3768774
27. Chandler C, Foltz PW, Cheng J, et al. Overcoming the bottleneck in traditional assessments of verbal memory: Modeling human ratings and classifying clinical group membership. In Niederhoffer, K., Hollingshead, K., Resnik, P., Resnik, R., Loveys, K. (Eds), *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology.* Minnesota, USA: Minneapolis, 2019:137–147. http://dx.doi.org/10.18653/v1/W19-3016
28. Holmlund TB, Chandler C, Foltz PW, et al. Applying speech technologies to assess verbal memory in patients with serious mental illness. *NPJ Digital Med.* 2020;3:33. doi: 10.1038/s41746-020-0241-7.
29. Fisher AJ, Medaglia JD, Jeronimus BF. Lack of group-to-individual generalizability is a threat to human subjects research. *Proc Natl Acad Sci USA.* 2018;115(27):E6106–E6115. doi: 10.1073/pnas.1711978115.
30. Foltz P, Rosenstein M, Elvevåg B. Detecting clinically significant events through automated language analysis: Quo imus? *NPJ Schizophr.* 2016;2:15054. doi: 10.1038/npjschz.2015.54.
31. Monarch, RM. *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI*. Shelter Island, NY: Manning Publications Co. 2021.
32. Hoffman R, Stopek S, Andreassen N. A comparative study of manic vs. schizophrenic speech disorganization. *Arch Gen Psychiatry.* 1986;43:831–838.
33. Dagan I, Engelson SP. Committee-based sampling for training probabilistic classifiers. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning (ICML'95)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995:150–157.
34. Siddhant A, Lipton Z. Deep Bayesian active learning for natural language processing: results of a large-scale empirical study. In. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* 2904;2018:2909. doi:10.18653/v1/D18-1318.
35. McCallum A, Nigam K. Employing EM and Pool-Based Active Learning for Text Classification. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998:350–358.
36. Blodgett S, Barocas S, Daume H, Wallach H. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020:5454–5476. 10.18653/v1/2020.acl-main.485.
37. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* 2021;54(6):1115. 10.1145/3457607 35
38. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? *Behavioral and Brain Sciences.* 2010;33(2-3):61–83; discussion 83-135. doi: 10.1017/S0140525X0999152X. Epub 2010 Jun 15. PMID: 20550733.
39. Wechsler D. Wechsler Memory Scale - Third Edition, WMS-III: Administration and scoring manual. San Antonio, TX: *The Psychological Corporation* 1997.
40. Ribeiro MT, Singh S, & Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. New York, NY, USA: Association for Computing Machinery, 2016:1135–1144. DOI: 10.1145/2939672.2939778
41. Cohen AS, Rodriguez Z, Warren KK, et al. Natural language processing and psychosis: on the need for comprehensive psychometric evaluation. *Schizophr Bull.* 2022;48(5):939–948.