HHS Public Access

Author manuscript

DIS (Des Interact Syst Conf). Author manuscript; available in PMC 2018 July 17.

Published in final edited form as:

DIS (Des Interact Syst Conf). 2018 June ; 2018: 559–571. doi:10.1145/3196709.3196776.

"It's hard to argue with a computer:" Investigating Psychotherapists' Attitudes towards Automated Evaluation

Tad Hirsch,
Department of Art + Design, Northeastern University, Boston, MA, tad.hirsch@northeastern.edu

Christina Soma,
Department of Educational Psychology, University of Utah, Salt Lake City, UT, christina.soma@utah.edu

Kritzia Merced,
Department of Educational Psychology, University of Utah, Salt Lake City, UT, k.mercedmorales@utah.edu

Patty Kuo,
Department of Educational Psychology, University of Utah, Salt Lake City, UT, patty.kuo@utah.edu

Aaron Dembe,
Department of Educational Psychology, University of Utah, Salt Lake City, UT, aaron.dembe@utah.edu

Derek D. Caperton,
Department of Educational Psychology, University of Utah, Salt Lake City, UT, derek.caperton@utah.edu

David C. Atkins, and
Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, datkins@uw.edu

Zac E. Imel
Department of Educational Psychology, University of Utah, Salt Lake City, UT, zac.imel@utah.edu

Abstract

We present CORE-MI, an automated evaluation and assessment system that provides feedback to mental health counselors on the quality of their care. CORE-MI is the first system of its kind for psychotherapy, and an early example of applied machine-learning in a human service context. In this paper, we describe the CORE-MI system and report on a qualitative evaluation with 21 counselors and trainees. We discuss the applicability of CORE-MI to clinical practice and explore user perceptions of surveillance, workplace misuse, and notions of objectivity, and system reliability that may apply to automated evaluation systems generally.

Keywords

Mental health; machine learning; psychotherapy; design

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous; I.2.1. Artificial Intelligence: Applications and Expert Systems; J.4 Applications: Social and Behavioral Sciences: Psychology; Design

INTRODUCTION

In 2015 over 5 million Americans accessed mental health services [47], and mental health and addiction problems are among the most common causes of disability in the U.S. [49].

Psychotherapy – a conversation between a client (or a group of clients) and counselor – is an effective treatment across a broad range of mental health and addiction problems [28]. Among approaches to psychotherapy, Motivational Interviewing (MI) is effective in promoting behavior change, particularly for addiction and behavioral health problems [35]. MI is explicit in its reliance on the therapeutic relationship (e.g., empathy, collaboration) and promotes specific counselor relational strategies such as the use of open-ended questions and making high-quality reflections of what the client has said.

Although MI has demonstrated efficacy in many research- based clinical trials [32], the quality of MI delivered in real- world settings can be highly variable, due to heterogeneity in the quality of training and supervision [40]. Gold-standard training and supervision involves recording sessions, deriving performance-based feedback from session recordings, and providing expert consultation and coaching. Unfortunately, this approach is time consuming, expensive, and does not scale up to real-world settings [1, 23]. Due to time, resources, and confidentiality issues, in many clinics supervision either does not occur or is often based upon the self-report of the counselor rather than independent observation of clinical sessions [17]. Independent, performance-based feedback is key to promoting and maintaining counselor skills [46], and so relying on retrospective self-report as the sole source of supervision data limits supervisors' ability to provide constructive, actionable feedback.

Advances in machine learning (ML) and natural language processing (NLP) offer computational methods to automatically map spoken language in psychotherapy sessions to quality indicators and performance based feedback [38]. Recent work has demonstrated that MI sessions can be evaluated using ML and NLP methods, and that machine-coded sessions can be comparable with humancoded sessions [1] for both specific, granular techniques (e.g., the quality of individual counselor reflections [8]) and for broad evaluations of an entire therapy session (e.g., how empathic the counselor is toward their patients [51]). These results suggest that NLP and ML may be used to offer counselors fast, reliable, objective feedback to support highquality therapy.

In this paper, we describe Counselor Observer Ratings Expert for Motivational Interviewing (CORE-MI), an automated performance-based feedback system for MI, and report results from a qualitative evaluation with 21 counselors and trainees. We offer two contributions to the field of interactive system design. We present the first study of perceived applicability of automated evaluation software for psychotherapy, an under-studied but potentially important

domain for interaction design research. We also identify challenges for user adoption of, trust in, and satisfaction with automated evaluation systems that may apply to machine learning systems generally.

RELATED WORK

There is growing interest among interaction design researchers in bringing computation to bear on problems associated with mental health. For instance, there have been recent efforts to automatically identify and diagnose suicidal ideation [15], insomnia [25], and other mental health disclosures [33] by analyzing social media [3] and mediated speech [10].

In a more interventionist vein, researchers have explored the potential for interactive games to teach users about anxiety disorders [48] and treat geriatric depression [30], while others have investigated the potential for mobile technologies to support users to monitor and reflect on symptoms of bipolar disorder [4], anxiety, and autism [41]. There have also been experiments in using augmented and virtual reality environments to offer exposure therapy to sufferers of PTSD [42] and phobia [50].

In the psychotherapy domain, there are several commercially-available systems that provide automated therapy, particularly cognitive behavioral therapy (CBT), a widely-used evidence-based therapeutic practice. Interestingly, a recent evaluation of automated CBT systems found that users fared poorly if they didn't also receive human support [39], a need that has been well-documented in the eMentalHealth literature (e.g., [26]). There are also established services that provide mediated counseling via the Internet and telephony networks (e.g., [29]), although prior work has found high attrition rates among users of these services [12].

While prior work has often emphasized supporting clients and promoting wellbeing, our work focuses on enhancing the work of professional counselors. Although less developed, there have been recent efforts in this area, including tools for counselors to construct therapeutic game environments for clients [13], and mobile phone applications that enable adolescents who experience mental health disorders to monitor their symptoms for later review with healthcare providers [34].

Despite this flurry of activity, there remains a dearth of research into computer support for training and evaluation of mental health practitioners. One notable exception has been the development of virtual standardized clients, which are conversational agents that simulate symptoms of various mental health conditions [27]. We believe that the relative lack of attention given to training and assessment is an unfortunate oversight, because most psychotherapy will continue to be delivered by human counselors for the foreseeable future, and at present, the quality of psychotherapy in the community is highly variable [22].

CORE-MI OVERVIEW

The Counselor Observer Ratings Expert for Motivational Interviewing (CORE-MI) is an automated coding and visualization system that provides report-card like feedback on psychotherapists' adherence with MI [24]. CORE-MI combines speech and language

processing with automated coding and interactive visualization to help counselors and trainers identify strengths and areas for improvement (Figure 1).

Speech and Language Processing

CORE-MI utilizes foundational speech signal processing methods to translate an audio-recording of an MI session into a numeric representation of semantic and vocal acoustic data, which are in turn used as features in machine learning predictive models. Key processing steps include: voice activity detection (i.e., is someone speaking or not?), speaker segmentation (also called "diarization"; i.e., which person is speaking?), role identification (i.e., is the speaker the counselor the client?), and automated speech recognition (i.e. what is being said?) [51]. Additionally, paralinguistic information such as prosody, pitch, speech rate, and intensity are all estimated. A variety of machine learning approaches have been used to take these speech and language derived inputs and predict MI-relevant features of the session.

CORE-MI makes use of the Barista open-source speech processing framework [14].

Automated Coding

CORE-MI provides feedback on standard MI quality measures described in the Motivational Interviewing Treatment Integrity Scale [37].

Summary Measures—The report presents users with an overall MI fidelity score: a composite metric of the six standard summary measures of MI quality: empathy, MI spirit, reflection-to-question ratio, percent open questions, percent complex reflections, and percent MI adherence. Overall MI fidelity is rated on a 12-point scale; in each of the six domains, counselors can receive one point for passing basic proficiency benchmarks and two points for passing advanced competency benchmarks, where these benchmarks have been defined by the treatment developers.

The second measure is MI adherence, a percentage-based metric, that divides the total number of MI-adherent utterances (e.g., asking open questions, making complex reflections, supporting and affirming patients, and emphasizing client autonomy) divided by the sum of MI adherent and MI non-adherent counselor behaviors.

Therefore, any score less than 100% indicates a session where the counselor used at least some non-adherent behaviors (e.g., being confrontational or giving unsolicited advice).

MI spirit and empathy are additional global ratings that capture the "gestalt" of the session. MI spirit assesses the overall competence of the counselor along dimensions of collaboration, evocation, and autonomy. Empathy measures the extent to which the counselor makes an effort to understand the client's perspective. Both are rated on a 1–5 Likert scale.

Behavior Counts—CORE-MI also tracks three "behavior count" measures that characterize the quantity and quality of questions and reflections in the discourse. Questions and reflections are key discursive elements in MI. Behavior count quality metrics include:

1. Reflection to question ratio: the total number of counselor reflections divided by the total number of questions. MI developers have set a 1:1 ratio to indicate basic proficiency and a 2:1 ratio to indicate advanced competence.
2. Percent open questions – open questions (as opposed to closed questions) are preferred as they allow a wide range of possible answers. The percent open questions metric is the number of counselor open-ended questions divided by the sum of the counselor's open and closed questions. Fifty percent open questions indicates beginning proficiency and advanced competence is earned at 70% open questions.
3. Percent complex reflections: the percentage of all counselor reflective utterances that are "complex" rather than "simple". Complex reflections add meaning to client utterances or summarize what they client has said across previous statements, whereas simple reflections may be mere restatements. Basic and advanced proficiency levels are achieved at the 40% and 50% levels, respectively.

Interactive Visualization

CORE-MI provides a report card-like, visual summary of counseling sessions (figure 2). The visualization is implemented in HTML and JavaScript, and makes use of the D3 visualization library. It is accessible through standard web browsers on a variety of devices and platfonns, but the design is optimized for tablet-like devices

At the top of the CORE-MI summary report page is information identifying the session, the counselor, and the client. The feedback portion is organized with the highest- level, summary data at the top and shows progressively more specific metrics throughout the page. Near the top-right of the page is a Notes section, which offers a free-field text entry box for trainees and supervisors to add comments on a session.

Interactive "information buttons" are available throughout and can be selected by users to view detailed information about particular measures. For example, a user might select the button next to the MI Adherence title to read about the how that measure is calculated, and desired competency thresholds.

The report prominently features an Overall MI fidelity score that aggregates all measures to give an impressionistic view of the participant's general level of adherence to MI principles. MI adherence, another summary rating, is depicted as a pie chart.

Bar charts indicate global measures (MI spirit, empathy) on a 1–5 scale where 3.5 – 4 indicates basic proficiency, and advanced competency is indicated by scores of 4 or greater. Changes in color highlight the counselor's level of proficiency. MI spirit characterizes the general therapeutic style of a session, captured by how collaborative, accommodating, and supportive the counselor is. Empathy is the degree to which the counselor makes efforts to see and understand their client's worldview.

The behavior counts section of CORE-MI provides finer grained feedback, focusing on summaries of specific types of counselor statements. Like the global ratings, each behavior count is visually depicted as a horizontal bar graph with colors changing as different levels of proficiency are reached.

The session view displays percentage of counselor and client talk time as a bar chart, and includes a session timeline that illustrates turn taking, with counselor and client speech indicated in different colors. Color intensity is mapped to vocally-encoded arousal, such that dark hues indicate a speaker speaking intensely – perhaps in anger, frustration, surprise, or joy. The top section of the session timeline represents the entire session and contains a moveable window that can be stretched and dragged over the bar to focus in on specific sections, shown in the bottom section at a larger scale. Hovering over specific talk turns brings up a transcription of the utterance and how the system coded it (e.g., as a counselor's open question). In this way, the session view provides an overview of key session dynamics, directs the user's attention to areas of greatest interest, and enables her to review key exchanges.

System Reliability

CORE-MI evaluates counselor fidelity to MI via the motivational interviewing skill code (MISC) [36], which is a gold-standard, human-based rating system. The ML / NLP system in CORE-MI was trained on a dataset of 300K utterances from 356 MI session recordings, which were hand labeled by an 8-person coding team using established MISC coding protocols [31]. Various machine learning models have been evaluated for predicting MISC codes (i.e., MI tone) features [9][45][51]. Averaging over all counselor codes, the correlation of model-based predictions with human-generated codes was found to be 93% of human reliability (SD = 16%) [7]. It is worth noting that, because human ratings are measured with error [i.e., inter-rater reliability is < 1.0], machine learning based predictions can never be higher than human reliability).

Overall, common codes and session summary codes (e.g., different types of counselor questions and reflections; empathy) are strongly predicted by ML/NLP models. Infrequent codes (e.g., confront) are more challenging for human and machine coding.

STUDY DESIGN

Counselors in community clinical practice are rarely evaluated, and never by an automated system. As such, we anticipate that CORE-MI may be disruptive to typical clinical workflows and may not be acceptable to all users. To better understand these concerns, we conducted a study of likely users' attitudes towards automated evaluation. We were particularly interested in three questions:

1. Receptivity: How open are counselors to the concept of automated evaluation?
2. Workflow: What role, if any, can counselors imagine automated evaluation playing in their clinical practice?

3. Concerns: What concerns, if any, do counselors have about introducing automated evaluation into their practice?

Participants

To evaluate potential users' perceptions of CORE-MI, we conducted a study with 21 counselors who were recruited at or nearby a large, North American university. 18 clinicians identified as White, two as Hispanic/Latinx (n = 2), and one as Asian American/Pacific Islander. The mean age of participants was 42.1 (*SD* = 11.7).

Counselors were sampled from two training categories. Clinicians-in-training (n = 10) were recruited from master's (n = 8) and doctoral (n = 2) level clinical training programs, and all clinicians-in-training were enrolled in introductory counseling training courses. The trainees reported having either no training or experience with MI (n = 4), or some familiarity, but no training, with MI (n = 5). One participant did not provide information on their experience in MI.

Experienced counselors (n = 11) were recruited from local community mental health sites. Contact information for experienced counselors was obtained via a snowballing sample where individuals within the mental health community spread information about the study via word of mouth. Seven of the experienced counselors were doctoral level practitioners who have been licensed for more than two years (n = 7), and four of the experienced counselors were licensed master's level practitioners. Regarding training experience, participants reported that they received training MI training and integrate MI in their practice (n = 2), are members of an MI organization (n = 1), have familiarity with and formal training in MI (n = 7), or have familiarity with, but no formal training, MI (n = 1).

Methods

This was a qualitative, mixed-methods study. Participants conducted brief counseling sessions with standardized clients, which were analyzed by CORE-MI. Participants were shown their results and interviewed by the research team. Interviews were recorded and transcribed, and thematic analysis was employed to analyze transcripts.

Standardized Patients

Each counselor participated in a 10-minute mock counseling session with a single graduate student, who was trained to portray a client seeking substance abuse treatment. This approach is often called a standardized patient (SP), and is a well-established research technique among psychotherapy researchers (for an example, see [2]). The SP profiles included an individual who was addicted to methamphetamines, and two college aged individuals seeking treatment for their alcohol usage. The SP profiles were randomly assigned to participants and enacted by four different graduate students.

Semi-structured interviews

SP sessions were recorded and analyzed by the CORE-MI system. Counselors received an email containing a URL to a password-protected CORE-MI summary report of their session. After the counselor had an opportunity to independently review his or her results, a semi-

structured interview was conducted by a member of the research team. During the interview, the researcher presented the report on a laptop and used a think-aloud protocol as the counselor navigated the report, asking clarifying questions when necessary. Participants were also asked about their perceptions of the report, and to provide feedback on each report section. Researchers probed on counselors' responses to receiving automated feedback on their clinical work, thoughts on incorporating CORE-MI into their clinical or training practice, and about the counselor's comfort and experience with MI.

Thematic analysis

Interview sessions were video recorded and transcribed verbatim. Following transcription, a team of eight researchers used thematic analysis [5] to identify common themes in participant interviews. First, each researcher read four transcripts and created their own code-books of salient themes. After independent coding, researchers convened to compare themes. The researchers reached consensus on common themes, and developed domain categories: higherlevel codes that describe similar themes [11]. Finally,researchers created a master codebook based on the generated domains and themes. All of the transcripts were then coded using the master codebook. To increase the trustworthiness and dependability of the findings, the researchers engaged in consensus coding throughout the study [18]. Each feedback session was coded four times by different researchers, and researchers met weekly to resolve disagreements and reach agreement on codes until all transcripts were analyzed. The researchers shared their rationale for specific codes and discussed how their codes may have been influenced by their views and biases.

RESULTS

Participants generally provided two categories of responses. The bulk of each interview was dedicated to providing specific responses to the CORE-MI summary report. However, participants also speculated on CORE-MI's applicability to clinical practice. For clarity, we report these responses separately.

Reactions to the CORE-MI summary report

Three major themes emerged in participants' discussions of the CORE-MI summary report: user interface, quantitative measures, and personal reflections. The user interface codes encapsulated participants' comments about layout, usability, accessibility of information, and aesthetics. The quantitative measures theme referred to the different MI quality metrics and language summaries presented by the tool. Personal reflections pertained to participants thinking about their own clinical work with patients, prompted by the CORE-MI report.

User Interface—Participants provided substantial feedback on the user interface. Much of this is specific to the CORE-MI report and may not be easily generalizable to other systems. We report these results in the interest of completeness, and to aid the reader in determining what if any influence user interface may have had on participants' overall impressions of CORE- MI.

Upon opening the feedback tool, many (10) participants said the layout was clean, uncluttered, and simple, though three participants attributed overlooking some information due to the simplicity of the layout. One participant found the amount of information initially overwhelming and difficult to navigate. Twelve participants described initially looking at the fidelity and adherence scores or the global ratings graphs and moving across the interface from left to right and then up to down. If the graph was viewed first, two participants noted that it was because it was in the middle of the page. Eight people found the bar graph benchmarks to be visible, eye-catching, and a familiar and easy way to assess proficiency, although some reactions were mixed regarding the ultimate meaning of these benchmarks and the information they conveyed. Participant 2E noted that the graphs "present nice and look really good."

16 participants used information buttons to gather information from the written statements, when they were feeling confused, and to retroactively confirm their assumptions about definitions and interpreting scores. As noted by participant 8E, "the info boxes were helpful," and the 16 participants indicated that they were intuitive to find.

CORE-MI presents detailed information on the Session Timeline, including talk-turns, vocally-encoded arousal, transcription, and behavioral codes linked to transcript. Participants described scrolling through the timeline in conjunction with reflecting on their session, examining the transcript as an indicator of their performance, and using the timeline transcript to confirm the accuracy of other measures. The participants also scrolled over the color blocks and saw the transcription and behavioral codes. There were concerns with the accuracy of the computer-generated transcript, but participant 14N noted that "eighty percent of it made sense." Three participants did not interact with theTimeline.

Different colors throughout the layout represented different concepts, and participants described a variety of thoughts surrounding the colors. Participant 5N noted the colors were, "calm…not passive." Nine participants used the colors to distinguish between counselor and client utterances on the session timeline as well as to distinguish between basic, intermediate, and advanced competency levels. Five participants noted that they recognized the change in color as a metric for proficiency (basic to advanced), though there were mixed feelings about how to interpret these findings. These participants described their interpretations of the competency metrics in conjunction with their personal feelings of their performance during the session.

Five participants did not notice varying color intensity on the session timeline. This may be due to relatively consistent levels of arousal during the SP sessions, and the corresponding consistency in color intensity across the session timeline.

Overall, participants found the interface usable and demonstrated basic competency with the system, including the ability to access information, interpret their scores, and generally understand how automated evaluation related to their sessions.

Quantitative Measures—Overall MI fidelity was described as a salient measure. Six participants indicated they looked first at fidelity and considered this metric as important

when interpreting the feedback page overall. Participant 10E, a novice counselor, noted that they looked to see whether they got a “passing grade.” Novice counselors seemed to interpret these scores as academic ratings, possibly due to the majority of their professional development occurring through classes. Alternatively, participant 3E, an experienced counselor, reflected on what this number meant for their performance and professional development. Participants, both novice and experienced, expressed a desire to see how their scores changed over time, broadly and with particular clients.

Several participants indicated they struggled to understand some of the measures. This was particularly true for global ratings (empathy and MI spirit) as well as overall evaluations of fidelity and adherence. Seven participants, a majority of novice counselors, reported not fully comprehending the meaning of MI Fidelity score; perhaps due to a lack of experience in the field and not having a more practical understanding of fidelity that might be held by experienced counselors. Nonetheless, these participants expressed confidence in the accuracy of the scores.

Four participants who said they didn’t understand some of the measures wanted to know more about their definitions or calculations. This desire seemed primarily motivated by participants seeking to improve their scores.

Several counselors commented on the apparent objectivity of quantitative data and automated systems. Five participants preferred the simplicity of behavioral count data to interpretive measures because it was “raw” and “objective.” The same general attitude held for percent talk time.

Several novice counselors cited objectivity as contributing to their willingness to accept automated evaluation. For example, participant 16N said it was “hard to argue with a computer,” and that the feedback data was objective because it was computationally derived instead of human evaluated. Participant 9E said it was in some ways easier to accept automatic feedback because it was “objective” and did not generate defensiveness in the same way as a human supervisor might when delivering critical feedback.

Overall, participants accepted the accuracy of the measures. Sixteen participants, both novice and experienced counselors, expressed a general acceptance of the accuracy of the measures despite not knowing how some constructs were defined, and in some cases, questioning transcription accuracy.

Experienced counselors were more likely to question the accuracy and calculation of feedback. For example, participant 1E expressed skepticism about how the computer program coded specific utterances and questioned the accuracy of his scores.

In summation, counselors were receptive to automatedfeedback, and were generally confident in the accuracy ofCORE-MI measures despite not fully understanding how they were derived. Several explicitly invoked notions of objectivity as underlying CORE-MI’s appeal.

Personal Reflections—The majority of participants (18) reflected upon how they interact with clients and their use of particular techniques, while they examined the CORE-MI report from their session with the standardized patient. Participants were particularly aware of how personal experience influenced their approach therapy. For example, participant 3E shared, "I think we all internally have a sense of space and how dominant we are in the room but that's I think so contingent upon your history and your family…so actually [the percent talk time] was a really helpful number for me to look at."

Nine participants compared scores from measures on the CORE-MI with their perceptions of how they conducted therapy. In particular, participants described how their CORE-MI scores were consistent with their therapeutic style or experience level. For example, participant 1E indicated that, "I tend to ask a lot of questions and that is my form of directiveness in session which is something that in the past I've tried to curb.. .and this just really brought out to me that I'm still being pretty directive." In addition, participant 6N indicated that, "I'm on a good course you know I'm just under basic and given my level of experience I think that's fine."

Seven participants reported that the software sparked reflection on areas of improvement, and that the CORE-MI offered an objective measure of skills to improve upon. Quantitative measures on the CORE-MI helped participants focus upon different domains and explore alternative ways of implementing techniques in therapy. For example, participant 14N shared that, "I felt like it…just clarified my own thoughts about my skills…it was nice to have maybe a professional opinion about what I needed to develop". Participant 3E stated that, "looking at these ratings…my sense is okay…like proficient like consistent enough…but potentially really looking at what are the things that I could be doing that [are] more in line with the overall spirit of MI…and empathy."

Clinical uses

Participants also speculated about CORE-MI's suitability for clinical use. Three subthemes emerged in these discussions: personal data benchmarks, supervision and training, and workplace concerns.

Personal Data Benchmarks—Throughout the interviews, participants seemed particularly contemplative about how to internalize feedback and curious about how scores were generated in order to integrate the feedback into practice. Participants expressed the desire to make personal improvements based upon the feedback provided. As was mentioned above, participants' questions about how measures were calculated appeared to be motivated by the desire to improve their scores (rather than, say, to challenge an assessment, or to simply understand how the system worked).

Many participants expressed interest in using CORE-MI to monitor how their practice evolves over time. Both novice and experienced counselors projected using this tool across multiple sessions with multiple clients to examine trends and patterns and study their counseling style with different clients.

Participants were curious about including information related to the counselor's progress in fidelity and adherence scores across multiple sessions. Two suggested including a personal average (based on the average across one's clients), as opposed to only comparing scores to a normative sample. Two others (8E and 9E) mentioned tracking how their practice evolved with particular clients. Participant 9E noted that "is not only how do I measure compared with the general (normative sample), but how do I compare myself with that particular client." The same participant (9E) referred to this concept as "personal data benchmarks"; they explained that this information could provide "longitudinal information of what's going on with a specific client", thus, suggesting the system's use for treatment plan and recommendations.

Three participants suggested additional behavioral count data such as amount of silence, and total time spent on reflections and questions, and imagined using these raw, basic numbers to gauge their own performance over time, or compare themselves to other practitioners. Participants also wondered about a method of continuously interacting with the feedback by adding comments about their clinical work. For example, participant 7E was curious about the tool's possibility to generate specific strengths and growth areas based on the computer-generated feedback. They indicated that an overview of these might better contextualize the information about the scores and the presented information.

One participant (15N) expressed her interest in delivering specific psychotherapy training before and after the implementation of the tool, as a way for counselors to stay current. Participant 12E recommended the use of alternate data to verify the tool's assessment; he noted that "it would be cool if it had physiological data (e.g. heart rate, skin response) to rate the session."

Training and Supervision—All the participants noted the potential use of the feedback tool in their clinical practice and supervision. A novice counselor (15N) stated "I could see it being used with therapy sessions being recorded and then using this tool to provide feedback. I could see a supervisor reviewing it first and then adding notes here." Eleven participants talked about the value of the tool within a training context, particularly noting the potential for supervisors to monitor trainees' progress.

Eight participants described how the CORE-MI could be integrated into the supervision and training of new counselors by generating detailed feedback for trainees and opening discussion of development of specific skills. In particular, they suggested that CORE-MI could be used to give quick, immediate feedback after sessions and could help counselors objectively see how they are interacting with clients. Participant 9E stated that "if I was doing this with someone, like a supervisee…we can visit numbers and I think they would be really good talking points. You know, like data that we can really reflect on." Similarly, participant 3E indicated that, "a lot of times in supervision we're looking at these more global things around content and really drilling down into the phrasing of questions and so on, [CORE-MI] would be really valuable I think." Eight participants also described how the notes section could be helpful by being used to give feedback of strengths and weaknesses to providers. Notes could be linked to specific instances in therapy sessions to explore ways that counselors could improve. For example, participant 10E shared that, "I think it would be

useful for the clinician as well to jot down things that they thought were good that they did, areas for growth." Finally, seven participants indicated that CORE-MI could establish benchmarks for whether providers are meeting adequate standards. Participants discussed how CORE-MI could be particularly useful in evaluating providers after they graduate. For instance, participant 9E stated that "at some level this could be used as a performance evaluation for an experienced counselor. Because once you become licensed you literally don't have to be supervised."

Workplace concerns—Several experienced counselors (n=3) expressed concern about the unintentional consequences of including this automated tool in clinical practice.

Experienced counselors mentioned worries about workplace surveillance, and the possibility of the CORE-MI serving as a benchmarking tool to penalize providers. Participant 18E said, "I would be open to [CORE-MI] but I'm concerned about, you know, like, is my performance based on the tool and then I would be evaluated by my boss. Is it gonna lead to getting fired... I don't know if that's the idea." Another participant (19N) raised concerns about the possibility of misusing the tool to rate a clinician's performance in contexts outside the individual therapy format, referring to intake assessments (initial interviews), psychological testing, and group therapy.

Three participants highlighted the need for training on how to use and interpret the different measures included in the feedback tool, both to enhance their own understanding and as a bulwark against misuse.

DISCUSSION

Our findings demonstrate that counselors are generally receptive to introducing a computer-generated assessment and evaluative technology into clinical settings. Trainee participants generally perceived the application as usable, and professional participants seemed genuinely excited about its potential to enhance clinical practice, promote skill development, and improve the quality of supervision.

Further, CORE-MI seems to encourage psychotherapists' self-reflection on their practice, which we take as a good indicator that the system can play a role in ongoing skill development. Participants drew links from the CORE-MI report to clinical activities and their approach to therapy, with little prompting from investigators. As described above, several participants described CORE-MI as offering a quantitative perspective on their implicit beliefs about therapeutic style and areas for improvement.

We also note that novice and expert counselors alike expressed interest in using CORE-MI to monitor changes in their practice over time, and to use CORE-MI for training and supervision. This suggests that counselors of varying levels of expertise are generally interested in continuing to improve their practice and see a role for automated evaluation in ongoing education and skill development.

However, we did observe differences in how novice and expert counselors perceived the system. Trainees generally accepted the accuracy of CORE-MI scores (regardless of whether

they understood how those scores were generated, or what they meant), while experienced counselors were more likely to question the validity of scores that did not match their subjective perceptions of their own performance. It may be the case that experienced counselors have a better understanding of the nuances of motivational interviewing and are thus able to probe more deeply into the meaning of particular scores. It might also be the case that novice counselors are more habituated to receiving performance evaluations, and thus are more accepting of this kind of feedback than their more experienced counterparts.

This study also suggests several challenges and opportunities for future development of CORE-MI. The first of these stems from the surprising finding that participants expressed great enthusiasm for and general trust in CORE-MI, even as they acknowledged they did not fully understand how feedback was derived or what precisely it meant. This tension applied to the application as a whole, and also to specific measures: as described above, participants relied on MI Fidelity as a primary performance measure and reflected on its implications for professional development, while simultaneously expressing a lack of understanding of the components of their score.

From one perspective, it may seem encouraging that participants were willing to accept feedback results as valid, as future iterations of similar tools may not struggle overmuch to establish legitimacy. However, this also raises concerns that counselors and supervisors could uncritically accept machine-generated ratings. It may become necessary to provide additional information to ensure users' interpretations of automated ratings matches their intended meaning and actual system capabilities as closely and reliably as possible.

We suspect that these findings may be at least partially attributable to general enthusiasm for the CORE-MI system, and to a broader cultural valorization of quantification and machine learning. We also hypothesize that recruiting participants with limited MI training may explain some of the misunderstandings about standard MI evaluation measures. Nonetheless, users' ability to trust in a system that they don't understand has been previously observed [21], and is endemic to ML systems—particularly when system predictions are based on trained non-linear ML models rather than on theoretical models. This is not to say that users will blindly trust ML systems. But, prior work suggests that evaluations of a system's reliability are based on subjective perceptions of its output and the perceived "soundness" of its reasoning rather than on statistical evidence of its accuracy[43]. The implication for system designers is that users' willingness to trust and ultimately adopt ML systems is dependent on perceived "legibility," meaning, the degree to which system behavior seems to "make sense" to its users, regardless of its mechanism or statistical accuracy [19].

We also uncovered apprehension among potential users, including concerns about surveillance and misuse in the workplace. Several participants expressed concern that CORE-MI assessments might be adopted by supervisors and administrators as performance benchmarks, with at least one participant expressly invoking the possibility of "getting fired" based on a CORE-MI evaluation.

These concerns were particularly pronounced among expert counselors, who presumably have more professional experience than their less experienced counterparts and are thus more sensitive to workplace issues.

We think there is an interesting link between these concerns, and the perception that CORE-MI provides objective assessments of counselor performance. As has been previously shown, ML systems can inherit biases implicit in their design and in the data sets upon which they are trained [6][44]. Nonetheless, they are often perceived as and valued for their apparent lack of prejudice.

Participants in our study suggested that automated evaluation's perceived objectivity might facilitate better, more open reflection and communication, because counselors might be less defensive receiving criticalfeedback from a machine than from a human supervisor.

Most notably, automated assessment was also seen as more definitive than human evaluation. This could be taken to suggest that CORE-MI predictions will lead to greater acceptance of counselor's shortcomings, and thus will promote learning and enhance performance. Viewed through the lenses of workplace justice, though, this observation takes on a darker hue. As system designers, we acknowledge the very real potential for CORE-MI to be adopted as a performance evaluation and selection tool. At the same time, we also recognize the inherent fallibility of machine learning systems. As such, we understand the possibility that decisions about counselor performance may be based on mistaken predictions made by machines, whose very statusas machines simultaneously render them unassailable even as they make mistakes. While it may be "hard to argue with a computer," as designers it is our responsibility to createmechanisms that enable users to do precisely that – bparticularly when the predictions that our systems make havethe potential to adversely impact human welfare [19].

We believe that concerns about system fallibility, perceived objectivity, and workplace justice may be broadly applicable to the design of ML-based automated evaluation systems for human service and other contexts. This suggests a vital new research area for research.

Prior research on interaction design and machine learning often focused on consumer applications (e.g., recommender systems), and much of it was conducted at a time when ML systems were relatively uncommon. In recent years, these systems have increasingly "become real" [20], finding widespread adoption and media attention. At the same time, they are finding their way into human service domains, including healthcare [52], public safety [16], and transportation planning. With these applications, ML is being used not simply to predict user preferences, but to anticipate and evaluate human performance in contexts involving human life, livelihood, and well-being. These are high-stakes applications where trust and perceptions of accuracy have significant consequences for end-users, and for the lives that depend on the decisions they make. As researchers, we have an important role to play in understanding these implications, and hopefully, helping to mitigate these risks.

LIMITATIONS AND FUTURE WORK

We acknowledge the limitations of our recruiting methods. Half of our study sample consisted of psychotherapy students, who are accustomed to having their counseling sessions recorded and evaluated. As such, they may be unusually willing to adopt an automated assessment system. Similarly, using snowball sampling may have biased our study sample of experienced counselors towards those who are generally accepting of automated evaluation. We look forward to evaluating CORE-MI with populations who are more representative of professional clinicians.

We also recognize inherent limitations of our study design. While this study has provided useful insights into users' initial impressions of CORE-MI's potential utility, usability, and desirability, we expect that these perceptions may change with ongoing use of the system. We look forward to evaluating users' actual experience with the system in clinical settings, over prolonged periods of time.

ACKNOWLEDGMENTS

Funding for the preparation of this article was provided by the National Institutes of Health/National Institute on Alcohol Abuse and Alcoholism (NIAAA) under award R01/AA018673 and National Institute on Drug Abuse under award R34/DA034860. In addition, David C. Atkins time was supported in part by K02 AA023814. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Drs. Atkins, Hirsch, and Imel are co-founders with equity stake in a technology company, Lyssn.io, focused on tools to support training, supervision, and quality assurance of psychotherapy and counseling.

REFERENCES

[1]. Atkins DC, Steyvers M, Imel ZE, Smyth P. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification Implementation Science. 2014; 9:1. Dec. 2014. . doi: 10.1186/1748-5908-9-49 [PubMed: 24398253]

[2]. Baer JS, Wells EA, Rosengren DB, Hartzler B, Beadnell B, Dunn C. Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors Journal of Substance Abuse Treatment. 2009; 37(2):191–202. Sep. 2009. DOI: 10.1016/j.jsat.2009.01.003 [PubMed: 19339139]

[3]. BagroyS, KumaraguruP and De ChoudhuryM 2017 *A Social Media Based Index of Mental WellBeing in College Campuses*. Proceedings of the 2017CHI Conference on Human Factors in Computing Systems(2017), 1634–1646.

[4]. Bardram JE, Frost M, Szántó K, Faurholt-Jepsen M, Vinberg M, Kessing LV. Designing mobile health technology for bipolar disorder: a field trial of the monarca system CHI '13 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2013; (2013):2627.

[5]. Boyatzis RE, Transforming qualitative information: thematic analysis and code developmentSage Publications; 1998

[6]. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases Science. 2017; 356(6334):183–186. Apr. 2017. . DOI: 10.1126/science.aal4230 [PubMed: 28408601]

[7]. Can D, Atkins DC, Narayanan SS. A dialog act tagging approach to behavioral coding: a case study of addiction counseling conversations *INTERSPEECH* (2015). 2015

[8]. Can D, Georgiou PG, Atkins DC, Narayanan SS. A case study Detecting counselor reflections in psychotherapy for addictions using linguistic features. 2012:2251–2254.

[9]. Can D, Marín RA, Georgiou PG, Imel ZE, Atkins DC, Narayanan SS. 2016. "It sounds like...": A natural language processing approach to detecting counselor reflections in motivational

interviewing Journal of Counseling Psychology. 2016; 63(3):343–350. . DOI: 10.1037/cou0000111 [PubMed: 26784286]

[10]. Chang K, Chan MK, Canny J. AnalyzeThis: unobtrusive mental health monitoring by voice CHI EA '11 CHI '11 Extended Abstracts on Human Factors in Computing Systems. 2011; (2011): 1951.

[11]. Charmaz K, Constructing grounded theorySage Publications; 2006

[12]. Christensen H, Griffiths K, Groves C, Korten A. Free range users and one hit wonders: community users of an Internet-based cognitive behaviour therapy program Australian and New Zealand Journal of Psychiatry. 2006; 40(1):59–62. Jan. 2006. . DOI: 10.1111/j.1440-1614.2006.01743.x [PubMed: 16403040]

[13]. Coyle D, Doherty G, Sharry J. PlayWrite: end-user adaptable games to support adolescent mental health *CHI EA '10 CHI '10 Extended* Abstracts on Human Factors in Computing Systems. 2010; (2010):3889.

[14]. CanD, GibsonJ, VazC, GeorgiouPG and NarayananSS 2014 *Barista: A framework for concurrent speech processing by USC-SAIL*. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014): Florence, Italy, 4 - 9 May 2014 (Piscataway, NJ, May 2014), 3306–3310.

[15]. De ChoudhuryM, KicimanE, DredzeM, CoppersmithG and KumarM 2016 *Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. CHI '16 Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), 2098–2110.

[16]. Emerging Technology from the arXiv. Neural Network Learns to Identify Criminals by Their Faces MIT Technology Review. 2016

[17]. Falender CA, Cornish JAE, Goodyear R, Hatcher R, Kaslow NJ, Leventhal G, Shafranske E, Sigmon ST, Stoltenberg C, Grus C. 2004. Defining competencies in psychology supervision: A consensus statement Journal of Clinical Psychology. 60(7):771–785. Jul. 2004. . DOI: 10.1002/jclp.20013 [PubMed: 15195339]

[18]. Hill CE, ed. 2012 Consensual qualitative research: a practical resource for investigating social science phenomena. American Psychological Association .

[19]. Hirsch T, Merced K, Narayanan S, Imel ZE, Atkins DC. Designing Contestability: Interaction Design, Machine Learning, and Mental Health ACM Conference on Designing Interactive Systems (DIS'17). 2017; (2017):95–99.

[20]. Höök K. Steps to take before intelligent user interfaces become real Interacting with Computers. 2000; 12(4):409–426. Feb. 2000. . DOI: 10.1016/S0953-5438(99)00006-5

[21]. Höök K, Karlgren J, Wærn A, Dahlbäck N, Gustaf Jansson C, Karlgren K, Lemaire B. A glass box approach to adaptive hypermedia User Modeling and User-Adapted Interaction. 1996; 6(2–3):157–184. Jul. 1996. . DOI: 10.1007/BF00143966

[22]. Imel ZE, Baldwin SA, Baer JS, Hartzler B, Dunn C, Rosengren DB, Atkins DC. Evaluating therapist adherence in motivational interviewing by comparing performance with standardized and real patients Journal of Consulting and Clinical Psychology. 2014; 82(3):472–481. Jun. 2014. . DOI: 10.1037/a0036158 [PubMed: 24588405]

[23]. Imel ZE, Steyvers M, Atkins DC. Computational psychotherapy research: Scaling up the evaluation of patient–provider interactions Psychotherapy. 2015; 52(1):19–30. 2015. . DOI: 10.1037/a0036841 [PubMed: 24866972]

[24]. Gibson James, Gray Geoff, Hirsch Tad, Imel Zac E., Narayanan Shrikanth; and Atkins David C. 2016 Developing an Automated Report Card for Addiction Counseling: The Counselor Observer Ratings Expert for MI (CORE-MI) . (San Jose, CA , 2016).

[25]. Jamison-Powell S, Linehan C, Daley L, Garbett A, Lawson S. "I can't get no sleep": discussing #insomnia on twitter CHI '12 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2012; (2012):1501.

[26]. Johansson R, Andersson G. Internet-based psychological treatments for depression Expert. 2012; 2006; 1614:01743.x.

[27]. Kenny P, Parsons TD, Gratch J, and Rizzo AA, 2008 Evaluation of Justina: A Virtual Patient with PTSD Intelligent Virtual Agents. Prendinger H, Lester J, , and Ishizuka M. , eds. Springer Berlin Heidelberg 394–408 .

[28]. Lambert MJ, ed. 2013 Bergin and Garfield's handbook of psychotherapy and behavior change. John Wiley & Sons .

[29]. Lederman R, Wadley G, Gleeson J, Bendall S, Álvarez-Jiménez M. Moderated online social therapy: Designing and evaluating technology for mental health ACM Transactions on Computer-Human Interaction. 2014; 21(1):1–26. Feb. 2014. . DOI: 10.1145/2513179

[30]. Li J. Examining the impact of game interventions on depression among older adults CHI PLAY '14 Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play. 2014; (2014):291–294.

[31]. Lord SP, Can D, Yi M, Marin R, Dunn CW, Imel ZE, Georgiou P, Narayanan S, Steyvers M, Atkins DC. Advancing methods for reliably assessing motivational interviewing fidelity using the motivational interviewing skills code Journal of Substance Abuse Treatment. 2015; 49:50–57. Feb. 2015. . DOI: 10.1016/j.jsat.2014.08.005 [PubMed: 25242192]

[32]. Lundahl B, Burke BL. The effectiveness and applicability of motivational interviewing: a practice-friendly review of four meta-analyses Journal of Clinical Psychology. 2009; 65(11): 1232–1245. Nov. 2009. DOI: 10.1002/jclp.20638. [PubMed: 19739205]

[33]. ManikondaL and De ChoudhuryM 2017 *Modeling and Understanding Visual Attributes of Mental Health Disclosures in Social Media. CHI '17 Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), 170–181.

[34]. MatthewsM and DohertyG 2011 *In the mood: engaging teenagers in psychotherapy using mobile phones. CHI '11 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2011), 2947.

[35]. Miller W, Rollnick S. *Motivational* Interviewing Guilford. 2012

[36]. Miller WR, Moyers TB, Ernst D, Amrhein P. Manual for the Motivational Interviewing Skill Code (MISC), Version 2.1 University of New Mexico Center on Alcoholism, Substance Use, and Addictions. 2008

[37]. Moyers TB, Martin T, Manuel JK, Miller WR, Ernst D. Revised Global Scales: Motivational Interviewing Treatment Integrity 3.1.1 (MITI 3.1.1) University of New Mexico Center on Alcoholism, Substance Abuse and Addictions (CASAA). 2010

[38]. Pace B, Tanana M, Xiao B, Dembe A, S Ba C, Soma C, Steyvers M, Narayanan S, Atkins D, Imel Z. What About the Words? Natural Language Processing in Psychotherapy. 2016

[39]. Rennick-EgglestoneS, KnowlesS, TomsG, BeeP, LovellK and BowerP 2016 *Health Technologies "In the Wild": Experiences of Engagement with Computerised CBT. CHI '16 Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), 2124–2135.

[40]. Schwalbe CS, Oh HY, Zweben A. Sustaining motivational interviewing: a meta-analysis of training studies Addiction (Abingdon, England). 2014; 109(8):1287–1294. Aug. 2014. . DOI: 10.1111/add.12558

[41]. SimmW, FerrarioMA, GradinarA, Tavares SmithM, ForshawS, SmithI and WhittleJ 2016 *Anxiety and Autism: Towards Personalized Digital Health. CHI '16 Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), 1270–1281.

[42]. Simms DC, O'Donnell S, and Molyneaux H, 2009 The Use of Virtual Reality in the Treatment of Posttraumatic Stress Disorder (PTSD) Virtual and Mixed Reality. Shumaker R. , ed. Springer Berlin Heidelberg 615–624 .

[43]. Stumpf S, Rajaram V, Li L, Burnett M, Dietterich T, Sullivan E, Drummond R, Herlocker J. Toward harnessing user feedback for machine learning. 2007; (2007):82.

[44]. Sweeney L. Discrimination in Online Ad Delivery Queue. 2013; 11(3):10. Mar. 2013. . doi: 10.1145/2460276.2460278

[45]. Tanana M, Hallgren KA, Imel ZE, Atkins DC, Srikumar V. A Comparison of Natural Language Processing Methods for Automated Coding of Motivational Interviewing Journal of Substance Abuse Treatment. 2016; 65(2016):43–50. . DOI: 10.1016/j.jsat.2016.01.006 [PubMed: 26944234]

[46]. Tracey TJG, Wampold BE, Lichtenberg JW, Goodyear RK. Expertise in psychotherapy: an elusive goal? The American Psychologist. 2014; 69:3. Apr. 2014. . doi: 10.1037/a0035099

[47]. *Treatment Episode Data Set (TEDS): Substance Abuse Treatment Admissions by Primary Substance of Abuse, According to Sex, Age Group, Race, and Ethnicity among admissions aged*

12 and older Year = 2015, UNITED STATES: 2015 https://wwwdasis.samhsa.gov/webt/quicklink/US15.ht m.

[48]. WehbeRR, WatsonDK, TondelloGF, GanabaM, StoccoM, LeeA and NackeLE 2016 *ABOVE WATER: An Educational Game for Anxiety*. CHI PLAY Companion '16 Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts (2016), 79–84.

[49]. Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, Charlson FJ, Norman RE, Flaxman AD, Johns N, Burstein R, Murray CJ, Vos T. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010 The Lancet. 2013; 382(9904):1575–1586. Nov. 2 013. . DOI: 10.1016/S0140-6736(13)61611-6

[50]. Wrzesien M, Burkhardt J-M, Alcañiz Raya M, Botella C. Mixing psychology and HCI in evaluation of augmented reality mental health technology CHI EA '11 CHI '11 Extended Abstracts on Human Factors in Computing Systems. 2011; (2011):2119.

[51]. Xiao B, Imel ZE, Georgiou PG, Atkins DC, Narayanan SS. "Rate My Therapist": Automated Detection of Empathy in Drug and Alcohol Counseling via Speech and Language Processing PLOS ONE. 2015; 10(12):e0143055. Dec. 2015. . doi: 10.1371/journal.pone.0143055 [PubMed: 26630392]

[52]. Yang Q, Zimmerman J, Steinfeld A, Carey L, Antaki JF. Investigating the Heart Pump Implant Decision Process: Opportunities for Decision Support Tools to Help. 2016; (2016):4477–4488.

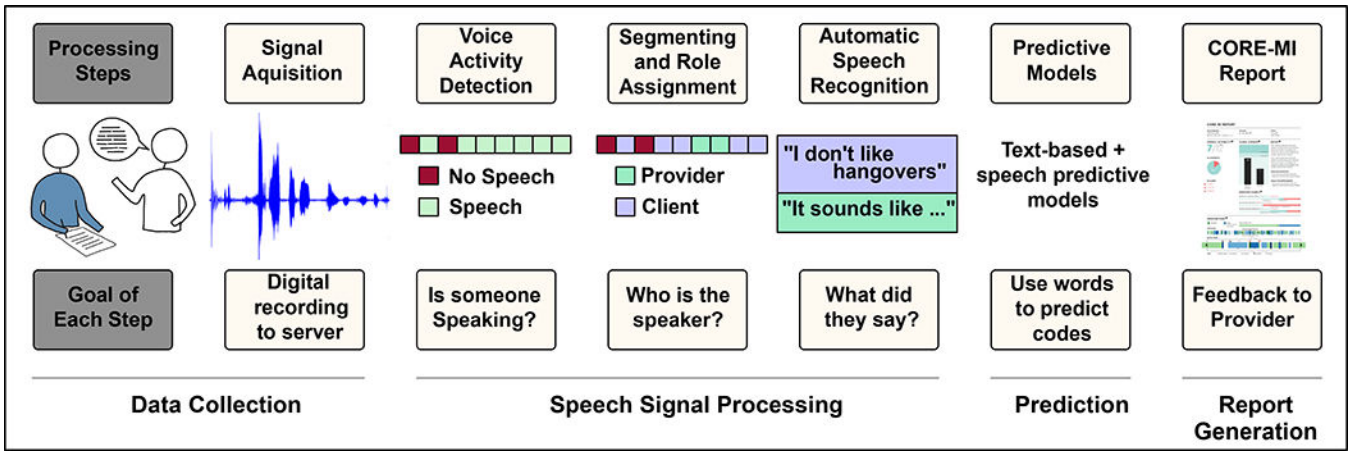

Figure 1.

: CORE-MI system architecture

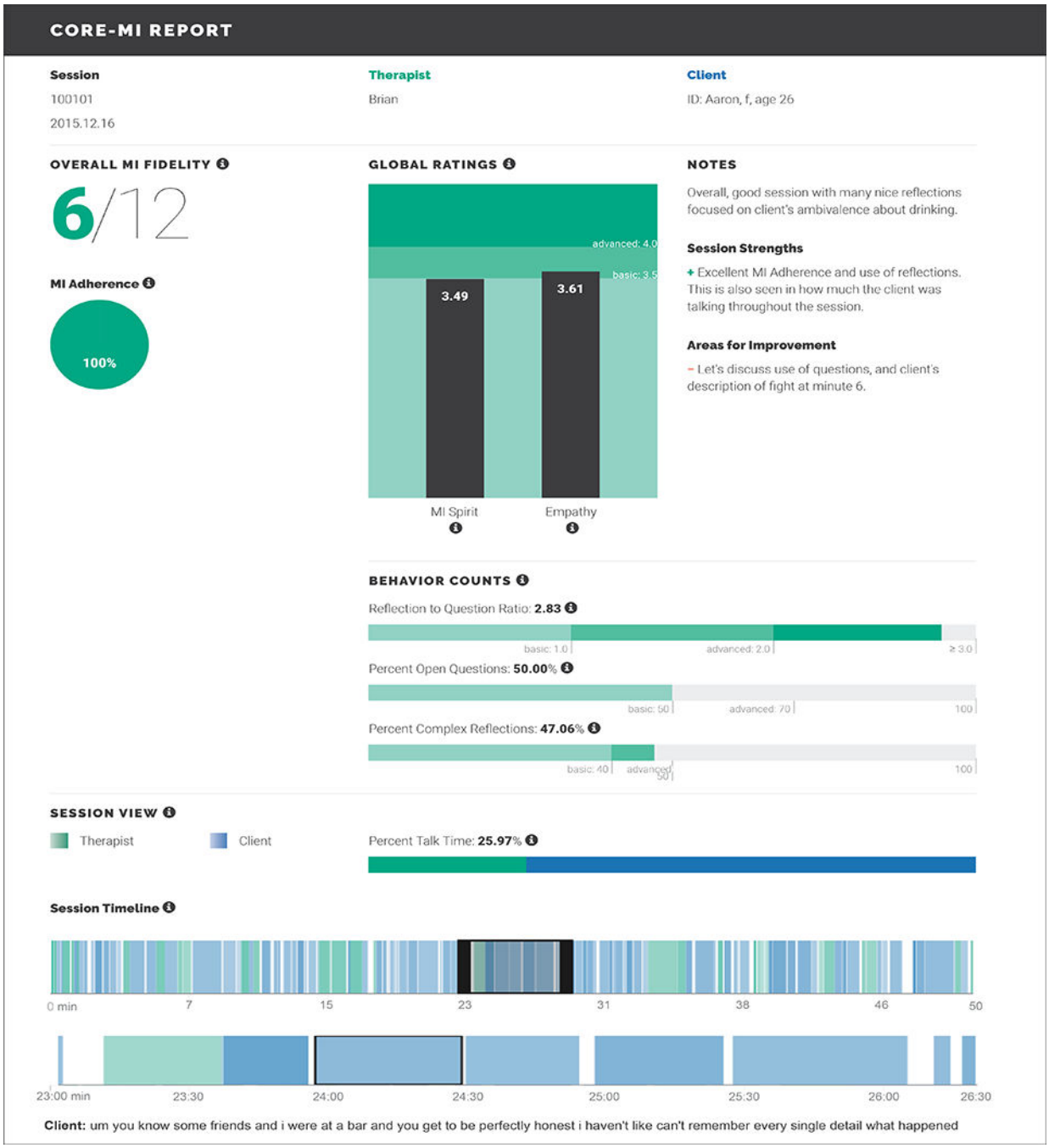

Figure 2.
: CORE-MI summary report