Developing a single-session outcome measure using natural language processing on digital mental health transcripts

MILLIGAN, Gregor <http://orcid.org/0000-0003-3357-9330>, BERNARD, Aynsley, DOWTHWAITE, Liz, VALLEJOS, Elvira Perez, DAVIS, Jamie, SALHI, Louisa <http://orcid.org/0000-0001-6458-1391> and GOULDING, James

Received: 26 January 2024 | Revised: 16 April 2024 | Accepted: 18 April 2024

DOI: 10.1002/capr.12766

ORIGINAL ARTICLE

WILEY

Developing a single-session outcome measure using natural language processing on digital mental health transcripts

Gregor Milligan[1,2] | Aynsley Bernard[3] | Liz Dowthwaite[4] | Elvira Perez Vallejos[4,5] | Jamie Davis[3] | Louisa Salhi[3] | James Goulding[1]

[1]N/LAB, Nottingham University Business School, University of Nottingham, Nottingham, UK

[2]Horizon CDT, University of Nottingham, Nottingham, UK

[3]Kooth PLC, London, UK

[4]Horizon Digital Economy Research, University of Nottingham, Nottingham, UK

[5]School of Medicine, University of Nottingham, Nottingham, UK

Correspondence
Aynsley Bernard, Kooth PLC, London, W2 1AY, UK.
Email: abernard@kooth.com

Funding information
Engineering and Physical Sciences Research Council; Kooth PLC

Abstract

Background: Current outcome measures in digital mental health lack granularity, especially for single-session interventions. This study aimed to address this by utilising natural language processing (NLP) methods to create a clear and relevant outcome measure. This paper describes the development of the Adult Session Wants and Needs Outcome Measure (Adult SWAN-OM), a novel outcome measure for the Qwell digital mental healthcare platform to understand service user (SU) needs engaging in single-session therapy (SST).

Methods: The research employs a multi-phased approach combining NLP methods with the typical stages of outcome measures development as follows: (1) assumption definition and validation with SUs and clinicians; (2) transcript theme extraction using the RoBERTa large language model (LLM) in conjunction with topic modelling to extract themes from 254 single-session transcripts from 192 SUs; (3) clinical item refinement focus group; (4) content validity with clinicians and SUs to improve the relevance and clarity of the items; and (5) outcome measure finalisation in a workshop held with clinicians to consolidate the final wording.

Results: Ninety-six potential wants and needs were generated and distilled into 12 measure items. The outcome measure was shown to be relevant and clear to both SUs and clinicians when used in the context of SST.

Conclusion: This study highlights the potential of combining NLP approaches with co-creation methods in single-session outcome measure development. We argue that the incorporation of clinical expertise and SU experience ensures the clarity and applicability of such measures and that this approach to capturing single-session wants and needs promises novel insights for supporting digital mental health interventions.

KEYWORDS
digital mental health, large language models, natural language processing, outcome measure, topic modelling

1 | INTRODUCTION

Qwell is a digital mental healthcare platform (DMHP) commissioned by the British National Health Service, local authorities, charities and businesses. Through this platform, a service user (SU) can access an online peer community and a team of emotional well-being practitioners. Qwell is anonymous at the point of use and free to access. Due to the wide-reaching and person-centred service model, there is a varied range of SU needs. This project builds upon the Kooth Session Wants and Needs Outcome Measure (SWAN-OM), designed by De Ossorno Garcia et al. (2021), to develop a novel single-session outcome measure that aids in understanding the idiographic wants and needs of SUs by extracting reoccurring themes from single sessions on the Qwell platform.

Outcome measures can be used within DMHPs to provide insight into the wants and needs of SUs. Single sessions offer SUs the opportunity to talk about problems, receive helpful advice, be referred to other resources and have direct access to intervention (Hymmen et al., 2013). However, De Ossorno Garcia et al. (2021) suggest that there is not a sufficient measurement for this type of intervention, which translates patient needs into achievable outcomes. Two types of sessions are evaluated in this work: 'single sessions' and 'one-at-a-time sessions' of between two and five sessions. This research shows how contemporary machine learning methods can be applied to the ubiquitous and often unused text data generated within DMHPs, and its uses in the development of an outcome measure through the analysis of transcript data.

Implications for practice and policy

1. This work will enable practitioners to quickly understand the clear and relevant wants or needs of a service user participating in single-session therapy on digital mental health platforms.
2. The work also addresses a key challenge in single-session therapy delivery in which there are no applicable, accessible outcome measures that are actionable in a single-session therapy delivery, and this outcome measure solves the challenge by presenting 12 potential outcomes that a service user could want to achieve in a single session.
3. This study shows the process of developing a single-session outcome measure using contemporary natural language processing techniques and how to combine these methods with well-established qualitative methods, such as workshops. We are providing insight into the application of large language models in evaluating transcripts in a single-session therapeutic context.
4. Implication for Policy: This study may influence future policy changes related to the development and provision of therapeutic outcome measures.

1.1 | Outcome measures: Development and application

For an outcome measure to be useful in a clinical environment, clients must be able to assign meaning to items in the measure and identify goals they find useful (Kwan & Rickwood, 2015). The process for developing outcome measures is typically carried out via focus groups and expert panels with a combination of SUs, clinical experts or practitioners (Blais et al., 1999; Rose et al., 2011); this study builds upon this existing approach of developing outcome measures by combining focus groups and expert panel insight with the evaluation of transcripts from therapeutic sessions between practitioners and clients. Outcome measures often follow two approaches: nomothetic approaches consist of validated outcome items based on population norms, and idiographic approaches are based on personalised items for individual patients rather than broader populations (Ashworth et al., 2019). Outcome measures can be used to understand specific problems or concerns, such as depression (Patient Health Questionnaire-9, PHQ-9; Kroenke et al., 2001) or anxiety (Generalised Anxiety Disorder-7, GAD-7; Spitzer et al., 2006). Nomothetic measures are used for determining more general therapeutic outcomes, including the Young Person's CORE (YP-CORE; Twigg et al., 2009), the Clinical Outcomes in Routine Evaluation Outcome Measure (CORE-OM; Evans et al., 2002) and the Kessler Psychological Distress Scale (K-10; Kessler et al., 2003).

Single-session therapy (SST) is a therapy that lasts for one session; there are several reasons why a client may partake in just one therapy session. Dryden (2020b) defines two primary types of SST; 'single-session therapy by default' is where a client books a series of sessions but only attends the first session and does not return to complete the subsequent sessions. In contrast, 'single-session therapy by design' is when a client arranges and completes a single therapy session with a therapist. The delivery of SSTs differs from other long-form interventions as the help is provided during one session rather than multiple sessions over a period of time. However, it should be noted that having a single session does not rule out the option of future sessions (Dryden, 2020a). 'One-at-a-time' (OAAT) interventions can also be part of the SST approach (Hoyt et al., 2018); although a single session is not initially part of a wider treatment plan, SUs may take part in several OAAT sessions depending on the therapeutic needs of the client. Young and Dryden (2019) suggest the increase in the uptake of SSTs/OAATs can be seen as a response to the increasing need for accessible and responsive mental health service delivery.

1.2 | Developing outcome measures for digital environments

Mindel et al. (2021) evaluated the suitability of three measures to understand the needs of SUs in the context of the DMHP Kooth: the

Short Warwick-Edinburgh Mental Wellbeing Scale (SWEMWBS), the Strengths and Difficulties Questionnaire (SDQ) and YP-CORE (Clarke et al., 2011; Goodman, 1997; Twigg et al., 2009). Based on the judgement of clinical practitioners within Kooth, all three measures demonstrated validity when used as indicators of user-rated needs upon entry to the platform; YP-CORE was the more appropriate measure in this context as YP-CORE can be used to measure both user needs and user outcomes. Although these measures can be used to understand SUs' needs, Hymmen et al. (2013) and Mindel et al. (2021) suggest that a better understanding of SUs' needs and greater impact of outcomes would be achieved by combining standardised measures with a more personalised assessment of the individual.

An initial attempt at a data-informed outcome measure is the SWAN-OM (De Ossorno Garcia et al., 2021); the SWAN-OM consists of a total of 21 outcome items split across six themes that aim to capture in-session goals and focus on the elements critical to the success of single-session and brief interventions. When administered across a 6-month period to 1401 SUs, the most frequently selected responses were '*Feel better*' and '*Find ways I can help myself*', while less commonly selected responses included '*Feel safe in my relationships*' and '*Learn the steps to achieve something I want*' (De Ossorno Garcia et al., 2022). Although the SWAN-OM provided insight into the development of a more robust outcome measure tailored to the digital mental health space, there were limitations to this work, primarily the small sample of transcripts analysed. This research builds upon previous work developing outcome measures (De Ossorno Garcia et al., 2021; Denner & Reeves, 1997; Honary et al., 2018) by evaluating 38,420 transcript messages across 254 conversations using contemporary natural language processing (NLP) methods to extract a more representative configuration of SU wants and needs.

1.3 | Natural language processing in digital mental health

NLP is a collection of computational techniques for learning, understanding and producing human language content (Hirschberg & Manning, 2015). Topic modelling is an NLP technique that can be used to represent large amounts of data in low dimensions and present hidden concepts, latent variables and prominent features of a corpus (Kherwa & Bansal, 2018). Dynamic topic models can provide a more nuanced understanding of the topics in a corpus and how the topics change over time (Blei & Lafferty, 2006). Transformer models have enabled considerable advancements in the field of NLP, with a widely cited transformer architecture being the Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019). BERT and other large language models (LLMs) can be modified to focus on a specific discipline by fine-tuning them on a corpus from that discipline, for example, Med-BERT (Rasmy et al., 2021) and BioBERT (Lee et al., 2020). These methods have wide-ranging applications within mental health fields, such as detecting specific mental health concerns or improving practitioner workflows.

In the context of mental health informatics, NLP has been employed for a wide range of applications, such as understanding general well-being, for example, predicting satisfaction with life by applying topic modelling techniques to Facebook message transcripts (Schwartz et al., 2016) or estimating the well-being of populations using 1.53 million geo-tagged tweets (Jaidka et al., 2020). Examples of NLP use cases for specific mental health concerns include recognising schizophrenia in corpora, detecting suicide ideation from counselling transcripts, and analysing social media data to detect depressive symptoms (Althoff et al., 2016; Coppersmith et al., 2018; Oseguera et al., 2017; Strous et al., 2009). Cook et al. (2016) showed that NLP models can generate relatively accurate predictions in identifying individuals at risk of psychological distress or suicide by using answers from a simple questionnaire about the patient's mood. These approaches can assist clinicians by quickly extracting and synthesising valuable information from text written by SUs.

Despite the versatile applications of NLP in mental health support, there are often limitations to the studies that prevent actionable insight for service delivery based on the outputs of said NLP techniques. This is primarily due to limited access to real-world data (Liu et al., 2021). Publicly available data, often from social media platforms, such as Facebook, X (formerly Twitter) or Reddit, are often used in place of sensitive and difficult to acquire therapeutic transcript data (Zeberga et al., 2022). This leads to insights that are not completely transferable to therapeutic practice owing to the different contexts of a social media platform compared with a DMHP. Although these research insights are useful, the way in which users engage with a public-facing social media platform is significantly different from the approach an SU would take in engaging with a digital mental health intervention. Therefore, the textual insights gleaned from an NLP model would also be different. This work addresses this contextual disparity by using real-world transcript data from DMHPs and applies contemporary NLP methods to understand the needs of SUs in the context of SSTs.

1.4 | Rationale and research aims

The rationale of this study was to explore how NLP techniques can be harnessed to aid the understanding of the wants and needs of SUs concerning single-session therapies on DMHPs. This research aimed to answer the question of how NLP methods can be applied to a corpus of DMHP transcripts to generate a clear and relevant outcome measure for adult users participating in SSTs on DMHPs.

This study aimed to make the following contributions:

1. A novel outcome measure for adults participating in SSTs.
2. Utilising NLP methods in the development of a single-session outcome measure for a DMHP.
3. Providing insight into the development of an outcome measure through the analysis of conversation-level transcripts.
4. Incorporating the perspective of individuals with experience of engaging with SSTs in the design of a single-session outcome measure.

2 | METHODS

The development of this outcome measure followed a multi-phased design process, informed by outcome measure development literature, particularly on participatory research, involving focus groups comprised of clinicians and individuals with lived experience of engaging with DMHPs. The outputs from the NLP analysis of conversation transcripts between SUs and practitioners underwent evaluation by both clinicians and individuals engaging with DMHPs. This evaluation guided the creation of a set of outcome items, which were further refined through content validity sessions involving Qwell clinicians, practitioners and experts in SSTs. The data were provided and anonymised by Qwell. These secondary data were processed by Qwell to ensure the transcripts did not contain any personally identifiable information, and this study received a favourable ethical opinion from the University of Nottingham's Business School ethics board. The transcripts analysed were from Qwell SUs that had previously provided research consent regarding the evaluation of their therapeutic transcripts. Informed consent was also gathered from participants of the workshops undertaken throughout this study.

2.1 | Dataset and demographics

The data used in this study consisted of transcripts between practitioners and SUs ($n = 874$) at the conversation level ($n = 2323$). A filter was applied to the dataset to ensure that various inclusion criteria were met. The inclusion criteria were the following: the SU must have previously given consent for their data to be included in Qwell research studies. Each session must be longer than 8 min, a timeframe defined by Qwell to ensure the SU actively participated in the session. The session should have no 'named worker' associated with it and must be a drop-in single session. Finally, the transcript must have an associated End of Service Questionnaire (ESQ). The process for these filtering criteria is shown in Figure 1.

Individuals in the final selected cohort ($n = 192$) are not significantly different from the wider Qwell SU population during the study period ($n = 874$) in terms of age, gender or ethnicity, suggesting that the cohort is representative of the wider target Qwell SU population. Each conversation was assigned a label based on the SU's rating of the session via an ESQ presented to the SUs at the conclusion of the session.

2.2 | Phase 1: Assumption definition and validation workshop

An initial workshop included people with experience engaging with DMHPs and Qwell clinicians to walk through the assumptions that were to be made when collecting and evaluating the transcript data used in this study. Assumption 1: The ESQ is an accurate representation of the extent to which an SU's wants and needs are met during a single session. Assumption 2: SUs need to identify their wants and needs in the session for them to be met. These assumptions facilitated the identification of transcripts and conversational components that align with wants and needs being met within a single session. The text data from successful or 'useful' single sessions (determined by Assumption 1) could be used to find topics to determine the wants and needs of SUs when entering single sessions. This workshop session enabled the researchers to gain insight into the thoughts of people with experience of engaging with DMHPs and contextualise the planned methodological approach with the participation of SUs and Qwell clinicians.

2.3 | Phase 2: Transcript theme extraction process

When SUs complete a text-based therapeutic session with a Qwell practitioner, they are presented with an optional ESQ, a four-item questionnaire that the SUs complete to reflect on the quality of the session. The ESQ contains four items, 'I felt heard, understood and respected', 'What we talked about was important to me', 'The person helping me was a good fit for me' and 'Overall, the session was right for me'; the items are rated on a scale of −1 (Not at all), 0 (A little) and +1 (A lot), for a total score for the ESQ of −4 to +4 for each user. The outcome score was used as the dependent variable in the training of a supervised learning algorithm to extract the textual features that contribute towards positive and negative single-session outcomes. A RoBERTA regression classifier was trained on transcript data from 2323 single sessions to extract textual elements that contributed to a successful session (higher ESQ scores). To understand what elements of each message were most impactful in the outcome of a session, the transformers-interpret model (Janizek et al., 2021) was used to extract word attributions from positively classified messages. Positive word attributions were then passed through a contextualised topic model (CTM; Bianchi et al., 2021) to extract and group positive word attributions into 10 topics; this process flow is shown in Figure 2.

2.4 | Phase 3: Clinical item refinement

The topics generated in Phase 1 were presented to a focus group of expert clinicians within Qwell to establish an initial set of outcome measure items; these items were refined into a smaller subset of items that encapsulate the spectrum of potential SU wants and needs. The initial measure items were structured into a content validity survey, which involved presenting the items to SUs and experts to assess the relevance and clarity of the items in the context of SSTs. The experts were then able to judge the quality of the outcome measure while also providing insight into the clarity and relevance of the items.

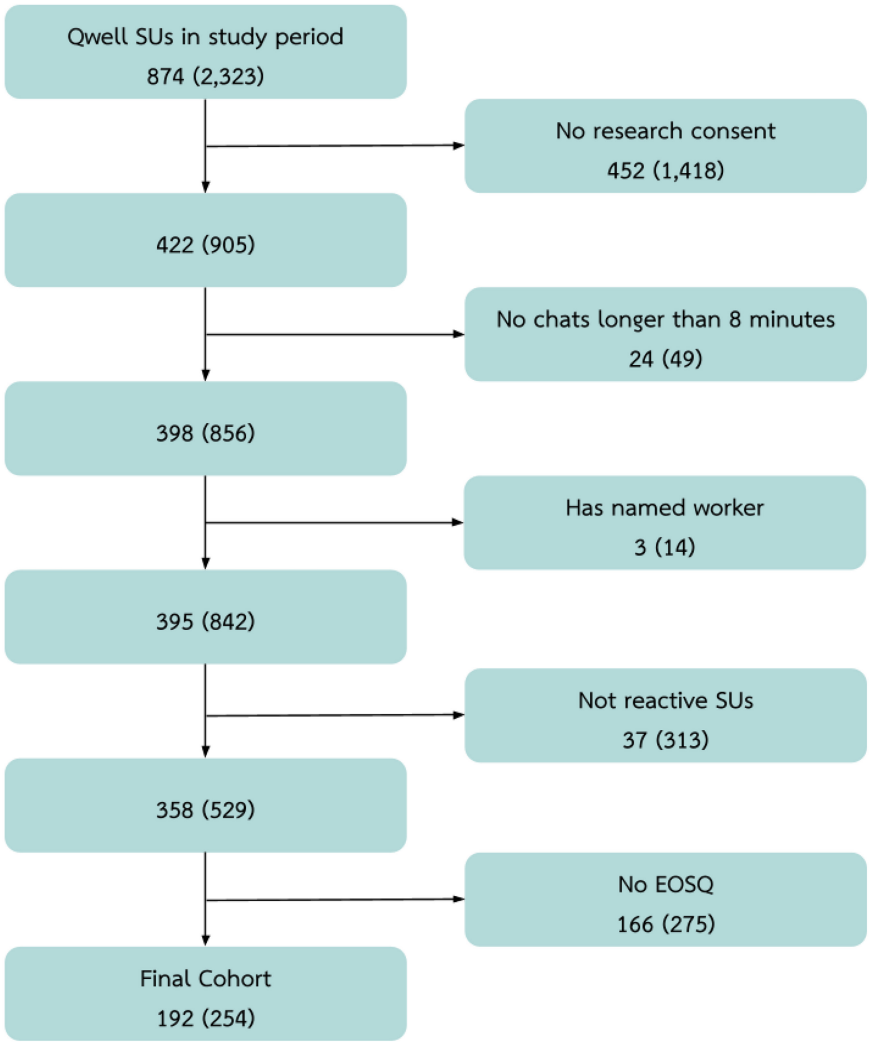

FIGURE 1 Cohort flow diagram (numbers in brackets represent the count of unique conversations between a service user [SU] and a worker).

2.5 | Phase 4: User and clinical content validity analysis

Content validity is the degree to which an instrument has an appropriate sample of items for the construct being measured. The Content Validity Index (CVI) is a widely used index for the evaluation of outcome measures; a panel of Qwell clinicians and Qwell SUs were asked to rate each scale item in terms of its relevance and clarity as a want or need within the Qwell single session. SUs were consulted first, being presented with the initial items generated in Phase 2; they rated the items for relevance and clarity following the Item-level CVI (I-CVI) framework (Lynn, 1986). There was also a free text box at the end of the survey which prompted SUs to suggest their own items that they felt were relevant. The I-CVI framework used a 4-point scale for relevance from 1 = 'not relevant', 2 = 'somewhat relevant', 3 = 'quite relevant' to 4 = 'highly relevant', and similarly for clarity, 1 = 'not clear', 2 = 'somewhat clear', 3 = 'quite clear' to 4 = 'highly clear'.

After the items were scored by both the SUs and the clinicians, items that scored equal to or above 0.75 I-CVI for both relevance and clarity were included in the initial outcome measure. Items with a relevance or clarity score of lower than 0.5 were not included in the measure, and if item scores were between 0.5 and 0.75 for either clarity or relevance, the item was reviewed in a workshop with clinicians ($n = 12$) and SUs ($n = 28$). Scale-level CVI (S-CVI), the average of the I-CVIs for all items on the scale, was also calculated for both SU and clinician scoring. The average item quality enables scrutiny of the total relevance of the measure (Polit & Beck, 2006). The acceptable S-CVI scores and the number of experts required to ascertain robust calculations have been debated in the literature, with a recommended number of experts ranging between two and nine, and S-CVI scores between 0.78 and 1 for excellent content validity (Yusoff, 2019). This process enables the finalised set of items to be defined and clinically validated.

2.6 | Phase 5: Outcome measure finalisation

After the SUs and the clinical team completed the content validity process and the scores were calculated, a workshop was held with clinicians to review the wording of the items that achieved an I-CVI score between 0.50 and 0.75. This workshop consisted of expert clinicians discussing each item to determine whether it should be

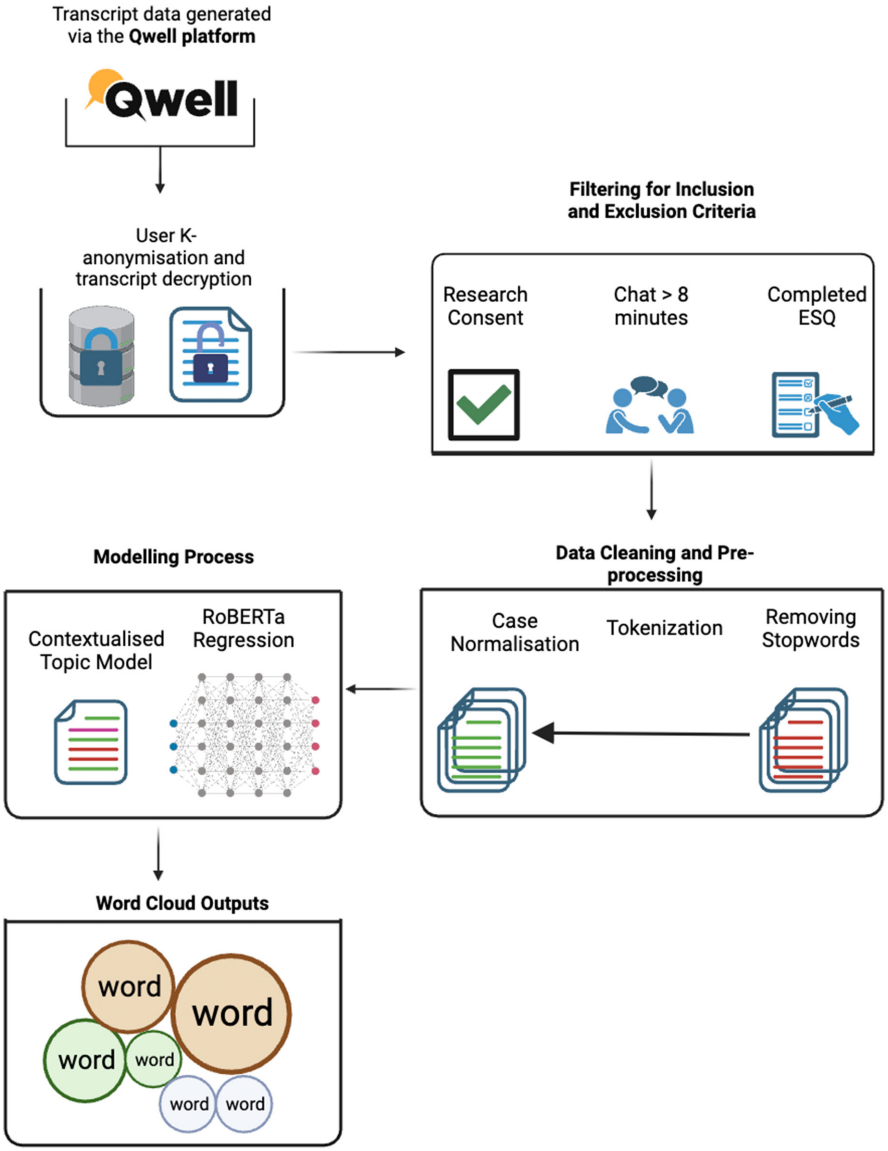

FIGURE 2 Topic modelling process flow diagram.

kept in the measure or removed and what alternative wording would be appropriate for the item to make it clearer or more relevant for an SST. Once the items were reviewed, a final content validity survey was sent to the original set of clinicians with the finalised wordings to gain the I-CVI and S-CVI scores for the outcome measure.

3 | RESULTS

3.1 | Machine learning modelling results

To determine a positive session, a RoBERTa model was applied through the Hugging Face transformer library. This model was used to model patterns between the text and ESQ scores. The outputs were evaluated through a BERT interpreter to extract the parts of the text that were most strongly associated with positive ESQ scores. To identify themes, a CTM was employed to group the parts of the text into 10 topics, represented visually as word clouds, each accompanied by a set of phrases to provide more contexts to each word cloud. These were assessed to formulate an initial collection of 96 potential wants and needs; the topic theme and a sample of the initial items are shown in Table 1.

3.2 | Content validity index results

The 96 initial item statements that were generated from the 10 word clouds were presented to clinicians during a workshop and refined into 11 items that represent a range of potential wants and needs that SUs could have when engaging in a SST on a DMHP. The initial 11 items were validated by the SUs who also suggested a range of additional outcome items that would be useful to include within the measure based on their experiences engaging with DMHPs. These suggestions were reviewed and grouped into four further items for a total of 15 outcome items (Table 2) before content validity was undertaken by the clinical team. Suggested items

TABLE 1 Word cloud topics and initial measure item examples.

Word cloud topic	Initial measure item examples (I want to…)
Improve and build my relationships/ Understand more about relationships	Build better relationships with my family Nurture my relationships Improve my relationships
Coping with life/Setbacks in life	Find ways of coping with setbacks Learn how to better manage my distress Learn some new coping tools
To be heard/Understood by others/ Acknowledged	Be heard and understood My efforts to be recognised and appreciated Have more fulfilling conversations
Access Support	Understand what professional help will work for me Build a support network Try things out to see what works for me
Understand/Feel better about/Express myself	Feel better about myself Be able to concentrate on what makes me feel better Have a better work-life balance
Understand why I feel a certain way (shame, guilt, fear)	Understand why I feel the way I do Have a space to grieve a loss Feel hopeful for the future
Understand my negative behaviour	Manage my negative behaviour Reduce the harm I am causing myself/others Understand why I hurt myself
Improve my physical health	Have a healthier lifestyle Do activities that improve my physical health Understand how to better manage my physical health condition
Take control	Take control of… Stop struggling with… Make the right decisions
Mindfulness	Work through areas of confusion Learn to relax and be calmer Remember the things I feel grateful for

included '*Learn more about mental health*' and '*Build a trusted relationship with the practitioner*'.

The I-CVI scores were calculated for both the clinical experts and SUs' surveys;[i] from the clinical I-CVI, survey six items (46.15%) were marked as relevant and clear, with I-CVI scores between 0.75 and 1 for both scales. Two items scored below the threshold in the relevance I-CVI but scored highly on clarity (I-CVI = 0.50–0.66). One item scored below the threshold for clarity (I-CVI = 0.50) but highly on relevance. The S-CVI average scores for the clinical responses were 0.66 and 0.70 for relevance and clarity, respectively, and for the SU responses, the S-CVI average scores were 0.70 and 0.63 for relevance and clarity, respectively.

3.3 | CVI expert reference group workshop

A final workshop was held with an expert panel ($n = 5$) to evaluate the items which attained an I-CVI score of between 0.5 and 0.75 for either relevance or clarity. Each item that did not meet the relevance or clarity threshold was reassessed, and further wording suggestions were defined. The workshop presented each item alongside the alternative item suggestions; experts were given the option to select an alternative wording, put forward new wording suggestions or remove an item; for example, the item 'Build a trusted relationship with the practitioner' was removed from the final list because, according to the experts, this did not seem achievable in a single session. Based on the insight from this final workshop, the items were consolidated into a final list of 13 items which were sent to the original set of clinicians to get the final I-CVI scores for clarity and relevance in preparation for face validity of the measure.

3.4 | Finalising outcome measure items

The I-CVI scores were calculated for the final set of items (Table 3). Nine items (84.62%) were marked as relevant and clear, with I-CVI scores between 0.75 and 1 across both scales. Item 1 scored below the threshold in relevance (I-CVI = 0.71) but scored highly in clarity (I-CVI = 0.83). In addition, two items scored below the threshold for clarity (I-CVI = 0.66) but highly on relevance (I-CVI = 0.83); these items were still included in the final measure after the expert panel feedback and will be trialled during face validity testing. The item

TABLE 2 Initial 15 outcome items.

Item #	Statement pre-chat	Statement post-chat
Item 1	Understand more about my relationships	I now understand more about my relationships
Item 2	Learn about coping strategies	I learned about coping strategies
Item 3	Feel heard or understood	I felt heard and understood
Item 4	Build my support system	I have found ways to build my support system
Item 5	Be more open to new experiences	I feel more open to new experiences
Item 6	Understand, express or improve my relationship with myself	I feel able to understand, express or improve my relationship with myself
Item 7	Help with grieving a loss	I had help with grieving a loss
Item 8	Understand how to improve my physical health	I understand how to improve my physical health
Item 9	Feel more aligned and progress with my values and intentions	I feel more aligned and am progressing with my values and intentions
Item 10	Work through a specific problem	I worked through a specific problem
Item 11	Feel calmer	I feel calmer
Item 12	Learn more about mental health	I learnt more about mental health
Item 13	Build a trusted relationship with the practitioner	I built a trusted relationship with the practitioner
Item 14	Find a safe, non-judgemental space	I found a safe, non-judgemental space
Item 15	Help overcoming set patterns	I felt helped overcoming set patterns

TABLE 3 Finalised items selected from the expert workshops.

Item #	Statement pre-chat	Statement post-chat	I-CVI relevance	I-CVI clarity
Item 1	Discuss and explore how to improve a specific relationship	I have discussed and explored how to improve a specific relationship	0.71	0.83
Item 2	Learn about coping strategies	I learned about coping strategies	1	1
Item 3	Feel heard or understood	I felt heard and understood	0.83	0.83
Item 4	Build my support system	I have found ways to build my support system	1	0.83
Item 5	Learn how to become more accepting of myself	I have learnt how to become more accepting of myself	0.83	1
Item 6	Help with grieving a loss	I had help with grieving a loss	1	1
Item 7	Understand how my physical and mental health could be linked	I understand how my physical and mental health could be linked	0.83	0.66
Item 8	Understand what my values are and how they could shape my actions	I now understand what my values are and how they could shape my actions	1	0.83
Item 9	Work through a specific problem	I worked through a specific problem	1	0.83
Item 10	Feel calmer	I feel calmer	0.86	0.83
Item 11	Talk about my story or my concerns with someone who is not judgemental	I have been able to talk about my story or my concerns with someone who is not judgemental	0.83	0.66
Item 12	Begin to understand unhelpful patterns of behaviour and how to change them	I have begun to understand unhelpful patterns of behaviour and how to change them	1	1
S-CVI			0.91	0.86

'Learn more about mental health' was removed after achieving an I-CVI score of 0.66 for both relevance and clarity and, therefore, did not meet the minimum I-CVI score. The average S-CVI score, which is the average of the I-CVIs for all items in the scale, was 0.91 for relevance and 0.86 for clarity, showing clear evidence that this outcome measure is relevant and clear to both SUs and clinicians when used to understand outcomes in a SST.

4 | DISCUSSION

This study shows the process of developing a single-session outcome measure, the Adult SWAN-OM, using contemporary NLP techniques, providing insight into the wide array of SU wants and needs in SSTs on DMHPs. The application of LLMs proved to be informative when evaluating a large corpus of transcripts from a

DMHP. This approach enabled the analysis of a significantly larger number of transcripts compared with the manual evaluation of therapeutic transcripts, improving the capacity to extrapolate the wants and needs of SUs. This approach enabled other processes, such as workshops and content validity, to happen sooner by speeding up the initial analysis of transcripts and allowing human resources to focus on other process elements.

The development of a data-informed outcome measure has expanded upon prior outcome measure development methodologies which typically take a participatory and qualitative approach, such as focus groups, participatory workshops, semi-structured interviews or the thematic analysis of transcripts (Blais et al., 1999; De Ossorno Garcia et al., 2021; Rose et al., 2011). This qualitative approach to outcome measure development is often limited by the number of focus group or workshop participants, the availability of experts or the number of transcripts that can be analysed and evaluated in the duration of the study. Specifically in the context of SST outcome measure development, De Ossorno Garcia et al. (2021) conducted expert workshops and manually evaluated a small sample of transcripts to develop the SWAN-OM; the present study expands on this work by evaluating 254 transcripts and conducting participatory workshop sessions with SST experts and SUs to develop the Adult SWAN-OM.

This study analyses data from a DMHP, which ensures that the findings can be tailored to the nuances of such platforms, offering more targeted and applicable insights compared with studies which use alternative data sources, such as text data from social media platforms or online forums (Coppersmith et al., 2018; Schwartz et al., 2016; Zeberga et al., 2022). By comparison, previous studies within the broader application of NLP in mental health have evaluated large samples of text data, such as the examination of 1.53 million geo-tagged tweets (Jaidka et al., 2020) or the analysis of 80,885 conversation transcripts (Althoff et al., 2016). However, it is noteworthy that these studies primarily concentrate on detecting specific mental health concerns, such as suicide ideation (Oseguera et al., 2017) and the likelihood of depression (Arachchige et al., 2021), rather than evaluating transcript data for SST outcome measure development. In contrast, this work utilises relevant transcript data and co-creation methods to contextualise the model outputs, providing direct insights for therapeutic practice and the development of SST outcome measures.

A challenge in SSTs suggested by Young and Dryden (2019) is the increased need for accessible and responsive service delivery; this work provides an outcome measure that has been designed to cover a range of wants and needs via the analysis of representative conversation-level transcript data and co-design methods. This approach has generated an outcome measure representative of a large cohort of DMHP SUs and incorporating their insight during the design of the Adult SWAN-OM should enable accessible and responsive service delivery. In the context of SSTs, this work has further illuminated the wide array of desired outcomes for adults participating in SST/OAAT interventions, and that there is a significant range of potential wants and needs. The Adult SWAN-OM enables SUs and practitioners to understand what the SU desires from a single session. Hymmen et al. (2013) evaluated the effectiveness of the SST delivery model and the outcome measures used to evaluate SSTs; these studies used non-standardised outcome measures thereby limiting the applicability of SST outcome evaluation. The Adult SWAN-OM was created by specifically evaluating SST transcripts to ensure the outcomes being measured are applicable to the SUs participating in SST, and the outcomes are achievable in a single session regardless of any potential follow-up sessions the SU may have.

4.1 | Limitations

A limitation of this study is that these results were generated using a fine-tuned RoBERTa model; the model evaluation metrics suggest that the model may not fully capture the data and could benefit from improvement through the inclusion of more data to enhance result accuracy. Nevertheless, the model used produced satisfactory results and still enabled the creation of this outcome measure; this limitation is mitigated by incorporating co-creation workshops and content validity surveys.

4.2 | Future work

The next step for this work is implementing the Adult SWAN-OM within the Qwell DMHP for face validity and pilot testing. This study lays the foundation for future work incorporating LLMs in the analysis of transcript data, providing insight into the elicitation of specific presenting issues.

5 | CONCLUSION

A key objective of this study was to create an outcome measure for SSTs on a DMHP, incorporating insights from SUs and clinicians into the design and implementation of the outcome measure. NLP methods were employed to evaluate a large volume of SST transcripts to create an outcome measure that represents the wide range of wants and needs of SUs undertaking SSTs. This has been achieved to a significant degree through the application of LLMs, focus groups and content validity surveys, gathering SU and clinical understanding to improve the relevance and clarity of this outcome measure, ensuring the items are applicable and achievable in a single session.

Including SUs in this process has enabled the creation of an applicable outcome measure, and applying a collaborative and iterative approach to item creation with contemporary machine learning methods demonstrates a strong case for the combination of computational analysis of text data with co-creation methods. A finalised set of 12 outcome items was defined that cover a range of themes that occur in SSTs, including improving relationships, building support systems and having a safe space to talk about concerns. This work provided novel contributions across several fields, including the application of LLMs

in the analysis of DMHP transcript data and the development of SST outcome measures by incorporating insights from people with lived experience of engaging with DMHPs. Ultimately, this work provides a novel, clear and relevant outcome measure tailored to SSTs.

FUNDING INFORMATION

This work has been funded as part of service innovation by Kooth plc. As one of the funders of the project, Kooth plc. provided the resources to plan for, collect and analyse data, run the surveys and workshops, and support with writing the report for this study. This work was supported by the UKRI Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/S023305/1 (EPSRC Centre for Doctoral Training in Horizon: Creating Our Lives in Data) and EP/T022493/1 (Horizon: Trusted Data-Driven Products).

DATA AVAILABILITY STATEMENT

The raw datasets presented in this article are not readily available because they contain information that can compromise the privacy of the research participants. Requests to access the datasets that support the findings of this study should be directed to research@kooth.com.

ETHICS STATEMENT

All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1964 and its later amendments. Ethics approval for this study was granted by the University of Nottingham Business School (Application ID: 202223018). Service user participants provided written informed consent when participating in the conducted workshops.

ORCID

Gregor Milligan https://orcid.org/0000-0003-3357-9330
Louisa Salhi https://orcid.org/0000-0001-6458-1391

ENDNOTE

[i]Tables showing the relevance, clarity and S-CVI scores for the initial 15 and final 12 measure items are found in the supplementary material files.

REFERENCES

Althoff, T., Clark, K., & Leskovec, J. (2016). Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, *4*, 463–476. https://doi.org/10.1162/tacl_a_00111

Arachchige, N., Sandanapitchai, P., Weerasinghe, R., & Isuri, A. (2021). Investigating machine learning & natural language processing techniques applied for predicting depression disorder from online support forums: A systematic literature review. *Information*, *12*(11), 444. https://doi.org/10.3390/info12110444

Ashworth, M., Guerra, D., & Kordowicz, M. (2019). Individualised or standardised outcome measures: A co-habitation? *Administration and Policy in Mental Health and Mental Health Services Research*, *46*(4), 425–428. https://doi.org/10.1007/s10488-019-00928-z

Bianchi, F., Terragni, S., & Hovy, D. (2021). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (volume 2: Short papers)* (pp. 759–766). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-short.96

Blais, M. A., Lenderking, W. R., Baer, L., deLorell, A., Peets, K., Leahy, L., & Burns, C. (1999). Development and initial validation of a brief mental health outcome measure. *Journal of Personality Assessment*, *73*(3), 359–373. https://doi.org/10.1207/S15327752JPA7303_5

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning—ICML '06* (pp. 113–120). ACM. https://doi.org/10.1145/1143844.1143859

Clarke, A., Friede, T., Putz, R., Ashdown, J., Martin, S., Blake, A., Adi, Y., Parkinson, J., Flynn, P., Platt, S., & Stewart-Brown, S. (2011). Warwick-Edinburgh mental well-being scale (WEMWBS): Validated for teenage school students in England and Scotland. A mixed methods assessment. *BMC Public Health*, *11*(1), 487. https://doi.org/10.1186/1471-2458-11-487

Cook, B. L., Progovac, A. M., Chen, P., Mullin, B., Hou, S., & Baca-Garcia, E. (2016). Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. *Computational and Mathematical Methods in Medicine*, *2016*, 1–8. https://doi.org/10.1155/2016/8708434

Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights*, *10*, 1–11. https://doi.org/10.1177/1178222618792860

De Ossorno Garcia, S., Edbrooke-Childs, J., Salhi, L., Ruby, F., Sefi, A., & Jacob, J. (2022). *'Examining concurrent validity and item selection of the wants and needs outcome measure (SWAN-OM) in a web-based therapy service' (preprint).* JMIR Mental Health. https://doi.org/10.2196/preprints.40122

De Ossorno Garcia, S., Salhi, L., Sefi, A., & Hanley, T. (2021). The session wants and need outcome measure: The development of a brief outcome measure for single-sessions of web-based support. *Frontiers in Psychology*, *12*, 748145. https://doi.org/10.3389/fpsyg.2021.748145

Denner, S., & Reeves, S. (1997). Single session assessment and therapy for new referrals to CMHTS. *Journal of Mental Health*, *6*(3), 275–280. https://doi.org/10.1080/09638239718806

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, *2019*, 16. https://doi.org/10.48550/arXiv.1810.04805

Dryden, W. (2020a). Single-session one-At-a-time therapy: A personal approach. *Australian and New Zealand Journal of Family Therapy*, *41*(3), 283–301. https://doi.org/10.1002/anzf.1424

Dryden, W. (2020b). *What is single-session therapy?* Windy Dryden. https://www.windydryden.com/post/what-is-single-session-therapy

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J., & Audin, K. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE–OM. *British Journal of Psychiatry*, *180*(1), 51–60. https://doi.org/10.1192/bjp.180.1.51

Goodman, R. (1997). The strengths and difficulties questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38*(5), 581–586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Artificial Intelligence*, *349*(6245), 7–266. https://doi.org/10.1126/science.aaa8685

Honary, M., Fisher, N. R., McNaney, R., & Lobban, F. (2018). A web-based intervention for relatives of people experiencing psychosis or bipolar disorder: Design study using a user-centered approach. *JMIR Mental Health*, *5*(4), e11473. https://doi.org/10.2196/11473

Hoyt, M. F., Bobele, M., Slive, A., Young, J., & Talmon, M. (2018). Single-session/one-at-a-time walk-in therapy. In *Single-session therapy*

by walk-in or appointment: Administrative, clinical, and supervisory aspects of one-at-a-time services (pp. 3–24). Routledge/Taylor & Francis Group. https://doi.org/10.4324/9781351112437-1

Hymmen, P., Stalker, C. A., & Cait, C.-A. (2013). The case for single-session therapy: Does the empirical evidence support the increased prevalence of this service delivery model? *Journal of Mental Health*, *22*(1), 60–71. https://doi.org/10.3109/09638237.2012.670880

Jaidka, K., Giorgi, S., Schwartz, H. A., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2020). Estimating geographic subjective well-being from twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*, *117*(19), 10165–10171. https://doi.org/10.1073/pnas.1906364117

Janizek, J., Sturmfels, P., & Lee, S.-I. (2021). Explaining explanations: Axiomatic feature interactions for deep networks. *The Journal of Machine Learning Research*, *22*(1), 4687–4740. https://doi.org/10.5555/3546258.3546362

Kessler, R. C., Barker, P. R., Colpe, L. J., Epstein, J. F., Gfroerer, J. C., Hiripi, E., Howes, M. J., Normand, S.-L. T., Manderscheid, R. W., Walters, E. E., & Zaslavsky, A. M. (2003). Screening for serious mental illness in the general population. *Archives of General Psychiatry*, *60*(2), 184. https://doi.org/10.1001/archpsyc.60.2.184

Kherwa, P., & Bansal, P. (2018). Topic modeling: A comprehensive review. *ICST Transactions on Scalable Information Systems*, *7*(24), 159623. https://doi.org/10.4108/eai.13-7-2018.159623

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kwan, B., & Rickwood, D. J. (2015). A systematic review of mental health outcome measures for young people aged 12 to 25 years. *BMC Psychiatry*, *15*(1), 279. https://doi.org/10.1186/s12888-015-0664-x

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Liu, Z., Peach, R. L., Lawrance, E. L., Noble, A., Ungless, M. A., & Barahona, M. (2021). Listening to mental health crisis needs at scale: Using natural language processing to understand and evaluate a mental health crisis text messaging service. *Frontiers in Digital Health*, *3*, 779091. https://doi.org/10.3389/fdgth.2021.779091

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*(6), 382–385. https://journals.lww.com/nursingresearchonline/fulltext/1986/11000/determination_and_quantification_of_content.17.aspx

Mindel, C., Oppong, C., Rothwell, E., Sefi, A., & Jacob, J. (2021). Assessing the need of young people using online counselling services: How useful are standardised measures? *Child and Adolescent Mental Health*, *26*(4), 339–346. https://doi.org/10.1111/camh.12456

Oseguera, O., Rinaldi, A., Tuazon, J., & Cruz, A. C. (2017). Automatic quantification of the veracity of suicidal ideation in counseling transcripts. In C. Stephanidis (Ed.), *HCI international 2017—Posters' extended abstracts* (pp. 473–479). Springer International Publishing. https://doi.org/10.1007/978-3-319-58750-9_66

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? critique and recommendations. *Research in Nursing & Health*, *29*(5), 489–497. https://doi.org/10.1002/nur.20147

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, *4*(1), 86. https://doi.org/10.1038/s41746-021-00455-y

Rose, D., Evans, J., Sweeney, A., & Wykes, T. (2011). A model for developing outcome measures from the perspectives of mental health service users. *International Review of Psychiatry*, *23*(1), 41–46. https://doi.org/10.3109/09540261.2010.545990

Schwartz, H. A., Sap, M., Kern, M. L., Eichstaedt, J. C., Kapelner, A., Agrawal, M., Blanco, E., Dziurzynski, L., Park, G., Stillwell, D., Kosinski, M., Seligman, M. E. P., & Ungar, L. H. (2016). Predicting individual well-being through the language of social media. *Biocomputing*, *2016*, 516–527. https://doi.org/10.1142/9789814749411_0047

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092. https://doi.org/10.1001/archinte.166.10.1092

Strous, R. D., Koppel, M., Fine, J., Nachliel, S., Shaked, G., & Zivotofsky, A. Z. (2009). Automated characterization and identification of schizophrenia in writing. *The Journal of Nervous and Mental Disease*, *197*(8), 585–588. https://doi.org/10.1097/NMD.0b013e3181b09068

Twigg, E., Barkham, M., Bewick, B. M., Mulhern, B., Connell, J., & Cooper, M. (2009). The Young Person's CORE: Development of a brief outcome measure for young people. *Counselling and Psychotherapy Research*, *9*(3), 160–168. https://doi.org/10.1080/14733140902979722

Young, J., & Dryden, W. (2019). Single-session therapy—Past and future: An interview. *British Journal of Guidance and Counselling*, *47*(5), 645–654. https://doi.org/10.1080/03069885.2019.1581129

Yusoff, M. S. B. (2019). ABC of content validation and content validity index calculation. *Education in Medicine Journal*, *11*(2), 49–54. https://doi.org/10.21315/eimj2019.11.2.6

Zeberga, K., Attique, M., Shah, B., Ali, F., Jembre, Y. Z., & Chung, T.-S. (2022). A novel text mining approach for mental health prediction using Bi-LSTM and BERT model. *Computational Intelligence and Neuroscience*, *2022*, 1–18. https://doi.org/10.1155/2022/7893775

AUTHOR BIOGRAPHIES

Gregor Milligan is a PhD student in the Horizon Centre for Doctoral Training at the University of Nottingham, working with the N/Lab to develop computational research into the insights obtainable about population well-being through proxy data streams. His research interest centres on understanding the relationship between deprivation and social isolation.

Aynsley Bernard is a Principal Data Scientist who joined Kooth in March 2021 to introduce Data Science and Machine Learning to the business. She leads responsible innovation projects that help Kooth understand their service users, measure practitioner support and improve service efficacy. She is an experienced data scientist with a lifelong commitment to learning, growth and positive social impact; she has an academic background in machine learning, applied statistics and mathematics.

Liz Dowthwaite is a Senior Research Fellow in the Horizon Digital Economy Research Institute at the University of Nottingham, working within the UKRI Trustworthy Autonomous Systems (TAS) Hub and Responsible AI UK. She has a research background grounded in psychology and human factors. Her research interests revolve around how motivation, attitudes, and human values relate to behaviour and well-being, particularly surrounding online participation, interaction with technology, and how human autonomy can be enhanced or undermined by autonomous systems.

Elvira Perez Vallejos is a Professor of Mental Health and Technology at the School of Medicine and School of Computer Science (50/50 appointment) at the University of Nottingham (UoN). Elvira's research portfolio (>£80 M) is highly interdisciplinary with funding from UKRI (AHRC EPSRC, MRC) and NIHR. She specialises in assessing the impact that technology has on the mental well-being of groups with protected characteristics (children, young people and older adults) by applying co-design and participatory methods (e.g., Youth Juries). She is driving world-leading research on digital mental health, including issues of data ethics and privacy, user and stakeholder engagement, co-production and responsible innovation.

Jamie Davis is currently a junior data scientist at Kooth. They have an academic background in health, statistics and machine learning. Their interests centre around the responsible use of AI and in exploring the nexus between mental health and data science in order to better understand, predict and respond to the needs of individuals.

Louisa Salhi is the Head of Research at Kooth; her team often bridge users, clinical practitioners and experts as well as commissioners. She leads research projects for Kooth where she works with academics and universities across the world to understand how digital mental health can better support people, including how to increase access and evidence its effectiveness.

James Goulding is an Associate Professor in Data Analytics, with over 100 internationally peer-reviewed publications. He is a trusted coordinator of large, international research projects with over £19 m of funding since 2015, working in a range of research fields, including COVID, Respiratory Diseases, Poverty Analysis, FGM & Modern Slavery, Food Insecurity Transport and Mental Health.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Milligan, G., Bernard, A., Dowthwaite, L., Vallejos, E. P., Davis, J., Salhi, L., & Goulding, J. (2024). Developing a single-session outcome measure using natural language processing on digital mental health transcripts. *Counselling and Psychotherapy Research*, *24*, 1057–1068. https://doi.org/10.1002/capr.12766