Interspeech 2024
1-5 September 2024, Kos, Greece

Speaking of Health: Leveraging Large Language Models to assess Exercise Motivation and Behavior of Rehabilitation Patients

Suhas BN[1], *Amanda Rebar*[2], *Saeed Abdullah*[1]

[1]College of Information Sciences & Technology, Penn State, USA
[2]Central Queensland University, Australia

bnsuhas@psu.edu, a.rebar@cqu.edu.au, saeed@psu.edu

Abstract

This paper aims to establish relationships between conversational markers and health outcomes using data from cardio-pulmonary rehabilitation sessions. Specifically, we used speech and text data from conversations between patients and researchers to assess exercise compliance and psychological well-being. We trained a Multimodal Transformer (MMT) on speech, transcript, and ground-truth labels. We further evaluate MMT's predictive performance by using session summaries generated by three Large Language Models (LLMs), which focused on dialogue characteristics (e.g., sentiment, thematic content, and future planning). Our findings establish the feasibility of augmenting speech and language processing of clinical sessions to improve decision-making and health outcomes.

Index Terms: Clinical Conversational Markers, Multimodal Learning, Large Language Models

1. Introduction

Cardiovascular diseases (CVDs) and Chronic Obstructive Pulmonary Disease (COPD) collectively represent a major public health issue. CVDs is a leading cause of death globally, claiming 17.9 million lives annually [1]. Furthermore, COPD is a major cause of chronic morbidity and mortality worldwide with estimates that it will become the third leading cause of death by 2030 [2]. This dual threat to global health highlights the critical need for research into effective treatments and interventions. Some CVD and COPD patients have access to exercise rehabilitation programs, consisting of 6-12 weeks of weekly in-clinic exercise sessions designed to enhance mobility, functional capacity, and quality of life [3, 4]. Participation in exercise rehabilitation programs can enhance physical and mental health, quality of life, functioning, and fitness [5, 6, 7, 8]. However, maintaining consistent exercise habits can be challenging for this population, which is critical for their recovery and long-term well-being [9]. Traditional methodologies for monitoring their adherence and wellbeing mostly rely on subjective self-reporting measures. While such measures are valuable, they are often limited by inherent biases and inability to capture granular data [10]. In this work, we focus on objective and real-time assessments of patient outcomes and experiences. Specifically, our study explores the integration of advanced linguistic analysis through Large Language Models (LLMs) such as Llama-2 7B [11], Meditron 7B [12], and Microsoft Phi-2 2.7B [13], aiming to provide insights about patients' motivation and likelihood of engaging in exercise rehabilitation. Linguistic analysis has previously been shown to be helpful to understand patient health behaviors and outcomes [14, 15, 16, 17, 18].

In this research, we aim to correlate specific linguistic markers in patient speech, extracted through end-to-end methodologies, with critical indicators of rehabilitation success: motivation, mental health, and exercise behavior. These include self-reported physical activity, measured via the International Physical Activity Questionnaire (IPAQ), psychological health assessments using tools such as the Depression Anxiety Stress Scales (DASS-21) [19, 20], and validated measures of motivation including the Behavioural Regulation in Exercise Questionnaire (BREQ-2) [21]. This approach opens new opportunities for understanding and improving the rehabilitation process [22]. By evaluating the impact of LLM-generated session summaries on the predictive accuracy of a Multimodal transformer for critical clinical outcomes, our research seeks to bridge qualitative insights with quantitative analysis, thereby enhancing the precision of patient motivation and rehabilitation adherence.

This paper makes the following contributions:

- Introducing a novel framework that combines clinician labels, LLM predictions, and embeddings of speech and transcripts, aiming to provide a more comprehensive understanding of patient motivation, mental health, & exercise behavior.
- Developing a loss function for managing the diversity of data types within cardio-pulmonary rehabilitation contexts,
- Enabling insights into parameters influencing patient motivation, mental health, and exercise behavior, areas traditionally challenging to quantify and monitor effectively.

2. Dataset

The data consists of audio conversations from interviews and corresponding text transcripts. This includes 73 recordings from a total of 31 subjects, with subjects contributing between 1 to 3 recordings each. The participant group comprised individuals with cardiac (n = 15) and pulmonary (n = 16) conditions. The majority of participants were male (71%), Caucasian (96.8%), with an average age of 71.5 years (SD = 9.4, range = 36 – 84 years). The required IRB approvals were obtained for both the data collection as well as the subsequent data analysis. Exercise behavior was measured using the International Physical Activity Questionnaire (IPAQ-SF) [24]. The IPAQ-SF consists of seven items requiring participants to reflect on the past seven days and report the number of days and average time (hours and minutes) spent performing vigorous, moderate and walking activities (e.g., "*During the last 7 days, on how many days did you do vigorous physical activities like heavy lifting, digging, aerobics or fast bicycling?*"; "*How much time did you usually spend doing vigorous physical activities on one of those days?*"). Participants completed this questionnaire twice — assessing exercise behaviors both during and not during the rehabilitation process. Time spent in physical activity was calculated as minutes per week of low, moderate and vigorous physical activity intensities. Two versions of the validated Be-

10.21437/Interspeech.2024-121

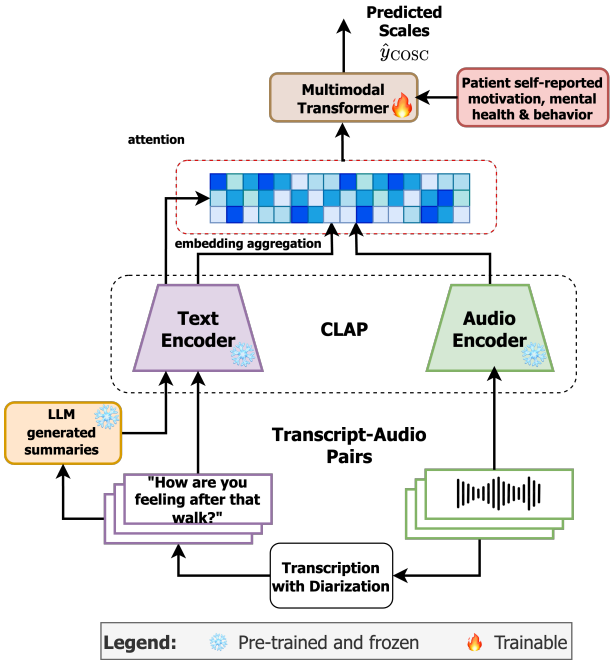

Figure 1: *The diagram illustrates our methodology, starting with the speaker diarization and transcription of clinical conversations. These transcriptions are fed into a CLAP model [23] for generating contextual embeddings. Concurrently, the audio encoder processes the original audio to capture paralinguistic features like intonation and stress, which may be lost in transcription. Both text-derived and audio-derived features are then integrated with interviewer and LLM-generated summaries within a multimodal transformer. In this framework, $\hat{y}_{COSC}$ denotes the Continuous/Ordinal Scale Component. This approach underscores the significance of combining speech and language analysis in healthcare applications.*

havioural Regulation in Exercise Questionnaire (BREQ-2) [21] were used to measure exercise motivation within and outside of rehabilitation. The BREQ-2 uses a 5-point Likert scale (0="*Not true for me*"; 4="*Very true for me*") with items such as "*I exercise because it's fun*", and "*I feel like a failure when I haven't exercised*". Scores were calculated as subscales of intrinsic (4 items), identified (4 items), introjected (3 items), and external (4 items) regulation as well as an overall relative autonomy index score. Interitem reliabilities ranged from α = 0.61 to α = 0.93. Depression, anxiety, and stress symptom severity was assessed with the 21-item Depression, Anxiety, and Stress Scale [25]. Participants reported how much each statement applied to them over the past week using the response scale ranging from 0 (*did not apply to me at all*) to 3 (*applied to me very much, or most of the time*). The depression, anxiety, and stress symptom severity scores were calculated as the sum of responses for the 7-item subscales: depression (e.g., "*I felt that I had nothing to look forward to*"), anxiety (e.g., "*I felt scared without any good reason*"), and stress (e.g., "*I found it hard to wind down*"). Each of the three scores could range from 0 – 21 with higher scores indicating more severe symptoms. These scales have acceptable inter-item reliability in the present study (depression α = 0.83, anxiety α = 0.67, stress α = 0.89). Quality of Life was assessed with the Assessment of Quality of Life 8-D (AQoL) [26, 27]. Participants responded to 35 items and scores were calculated as standard for the 8 dimensions of independent living, relationships, mental health, coping, pain, and senses. Additionally, an overall quality of life score was calculated. Inter-item reliability was acceptable, with α ranging from 0.79 to 0.89. We enforced a strict partitioning strategy, allocating entire subject datasets to either training, validation, or test sets, effectively preventing data leakage and ensuring robust model evaluation. This strategy led to a distribution of 17 subjects (40 sessions) for training, 5 subjects (12 sessions) for validation, and 9 subjects (21 sessions) for testing. We downsampled the audio from 44.1 kHz to 16 kHz for use in the CLAP model [23]. We also extracted turn-by-turn conversation transcripts by using WhisperX [28], which provided us with the start and stop times, speaker information, and the text content. We validated the outputs by comparing a subsection of transcripts generated by another service.

LLM Generated Summary Based on Rehabilitation Session

Format: The overall sentiment of the conversation is {*sentiment*}. The primary subject matter or theme revolves around {*theme*}. The level of engagement exhibited by the patient is {*engagement*}, and the clinician's approach is {*approach*}. Actionable insight provided includes {*insight*}. The narrative structure is {*narrative*}, with a {*complexity*} level of complexity. Emotional support is {*support*}, and the conversation identifies {*barrier*} as a compliance barrier. Future planning is {*planning*}.
Example: "The overall sentiment of the conversation is positive, indicating a generally upbeat and constructive interaction. The primary subject matter or theme of the conversation revolves around emotional well-being, focusing on the mental and emotional state of the patient. The level of engagement exhibited by the patient is high, suggesting the patient was fully participative and responsive, and the clinician's approach is supportive, offering empathy and understanding. Actionable insight provided includes follow-up actions, outlining next steps for continued care. The narrative structure is unresolved, leaving some questions or issues open, with a high complexity, involving multiple layers or nuances level of complexity. Emotional support is moderate emotional support, providing a balanced level of understanding, and the conversation identifies resource accessibility issues, pointing to external limitations as a compliance barrier. Future planning is goal-oriented planning, with clear objectives set for the future."

Figure 2: *This figure shows an example of how the LLM generates summaries from rehabilitation sessions. It illustrates the model's ability to condense complex dialogues into key themes such as sentiment, engagement, and the clinician's approach. These themes can then provide actionable insights to improve clinical care and patient outcome.*

3. Methodology

Loss Function: In this work, we propose the Adaptive Huber Loss (AHL), a refined version of the traditional Huber loss function. AHL is specifically designed for regression tasks involving continuous and ordinal variables, commonly encountered in health assessments such as DASS-21, IPAQ, and BREQ-2. The objective of AHL is to enhance the accuracy of predictive models dealing with cardiorespiratory fitness and behavioral data by more effectively managing the error sensitivity unique to each variable. We formulate AHL as follows:

$$L_{\text{AHL}} = \alpha \cdot L_{\text{Huber-COSC}}, \tag{1}$$

where $L_{\text{Huber-COSC}}$ extends the traditional Huber loss to all M continuous or ordinal variables by integrating:

$$L_{\text{Huber-COSC}} = \frac{1}{M} \sum_{j=1}^{M} \left(\text{Huber}(\hat{z}_j, z_j; \delta_j, w_j)\right), \tag{2}$$

where, $\hat{z}_j$ and z_j denote the predicted and actual values, respectively, while δ_j and w_j are the variable-specific thresholds and weights tailored to account for the distinct sensitivity and distribution of errors across different variables. The parameter α serves as an adaptive scaling factor, enabling fine-tuning of

Table 1: *Framework for Evaluating Clinical Conversations. This table outlines the analysis criteria for evaluating patient-clinician dialogues, including conversation sentiment, themes, engagement levels, among others. It provides a structured methodology for extracting insights from rehabilitation sessions, aiming to quantitatively link linguistic markers with patient health outcomes.*

LLM generated:	Overall Conversation Sentiment (Positive, Neutral, Negative), Conversation Themes (Exercise Motivation and Challenges, Health and Treatment Updates, Personal and Emotional Well-being, Rehabilitation Program Details, Future Health Goals and Planning), Patient Engagement Level (Highly Engaged, Moderately Engaged, Minimally Engaged), Clinician's Approach (Informative, Motivational, Supportive, Directive), Actionable Insights (Exercise Regimen Adjustments, Lifestyle Changes Recommendations, Medication or Treatment Modifications, Follow-up or Referral Suggestions), Narrative Structure (Cohesive, Fragmented, Resolved, Unresolved), Conversation Complexity (Simple, Moderate, Complex), Emotional Support and Empathy (Frequently Offered, Occasionally Offered, Rarely Offered), Barriers to Compliance (None Identified, Personal, Environmental, Physical), Future Orientation and Planning (Goal-Oriented, Progress-Focused, Adjustment-Oriented, Limited).
Interviewer:	Clinical Assessments (Health Status, Progress Evaluation), Previous Rehabilitation Programs, Current Rehabilitation Program Weeks, Gym Session Count, Decisional Intention for Physical Activity Outside Rehabilitation (Frequency, Strength), Decisional Intention for Physical Activity During Rehabilitation (Frequency, Strength), Physical Activity Engagement (Time Spent Outside, Time Spent Within Rehabilitation), Self-Determined Motivation (Outside, Within Rehabilitation), Mental Health Symptoms (Depression, Anxiety, Stress), Quality of Life Assessment.

the loss function's sensitivity to discrepancies in the data. This approach to AHL, with its emphasis on variable-specific customization, aims to address the diverse error sensitivities and distributions inherent in each variable, seeking to improve the robustness and accuracy of model predictions. Variable weights were determined by predictive importance & variability using feature importance. Thresholds were set iteratively based on validation error distribution to balance noise reduction & variation capture. The normalization process involved scaling the mean absolute deviation (MAD) of each variable against the median MAD across all variables, ensuring that weights reflect both the relative variability and significance of each variable in predicting health outcomes.

Training: Our MultimodalTransformer utilizes speech, textual data, and LLM-generated session summaries for continuous outcome prediction in therapy sessions. We have the labels from validated self-report measures (ex., how.often.1 or stress) provided for a subject per session. We utilized 5-fold cross-validation and reported avg. performance across folds to ensure model stability. For a detailed representation of the LLM summaries, refer to Fig. 2. Inputs are dynamically prioritized through an AttentionModule, and the resulting embeddings are processed by a TransformerEncoder, configured with n=64 attention heads and d=6 layers. The model's training utilizes the Adam optimizer, starting at a learning rate of 5e-3, and employs a ReduceLROnPlateau scheduler for learning rate adjustments based on validation loss. Performance is evaluated by calculating the Mean Absolute Error (MAE) for each outcome. We incorporate zero-shot classification capabilities from three LLMs [11, 12, 13]. Analyzing entire session transcripts allows for the identification of critical dialogue dynamics and markers indicative of patient outcomes. Further, clinician annotations augment this analysis, correlating speech and language patterns with factors influencing patient compliance and mental health. The integration of LLM analysis and clinical insights for dialogue evaluation is shown in Table 1 and the overall methodology is illustrated in Fig. 1.

4. Discussion

We begin with an analysis of pairwise correlations, providing insights into the correlations between mental health, exercise motivation, and behavior both within and outside rehabilitation sessions, as depicted in Fig. 3. While correlation does not equate to causation, these findings offer valuable indicators for potential improvements in treatment strategies.

• **Intention and Commitment in Physical Activity:** We observed strong correlations between the intention for physical activity outside rehabilitation sessions (how.often.1) and the strength of that intention (what.extent.1; r = 0.925), suggesting a close alignment between the frequency of intended physical activity and the commitment level. This pattern underscores the link between patients' exercise intentions and their commitment to fulfilling those intentions.

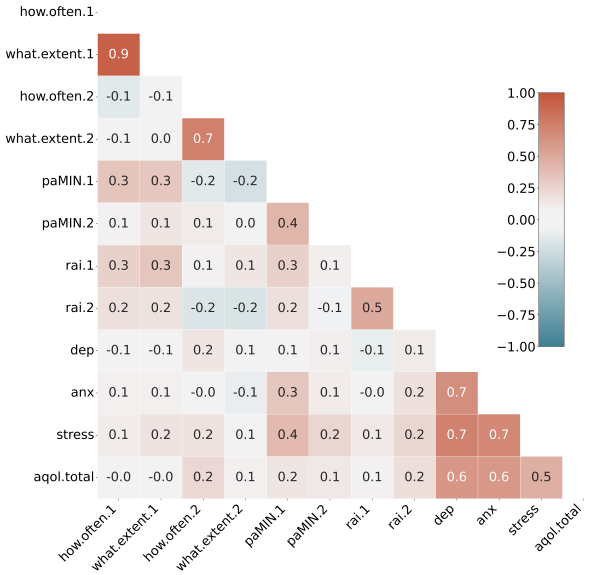

Figure 3: *The heatmap presents the relationships between variables such as exercise motivation, mental health, and activity levels.*

• **Contextual Variations in Patients' Exercise Intentions:** Contrarily, the intention for physical activity during rehabilitation (how.often.2) showed no significant relationship with intentions outside of rehabilitation (how.often.1; r = -0.083) and even a negative association with intention strength during rehabilitation (what.extent.2; r = -0.104). These results highlight how rehabilitation environments and external settings may influence motivational dynamics, highlighting the need for tailored motivational strategies across different contexts.

• **Exercise Duration and Rehabilitation:** A moderate correlation (r = 0.437) was found between the duration of physical activity conducted inside (paMIN.2) and outside (paMIN.1) rehabilitation sessions. This suggests distinct factors may influence exercise duration in different settings, with an important finding being the absence of a negative correlation, indicating that engagement in rehabilitation exercises complements rather than substitutes for external physical activities.

Table 2: *The Mean Absolute Error (MAE) values for different Large Language Models (LLMs) across various prediction tasks. Best performing method is shown in bold. The scores highlight the relative effectiveness of different models to accurately forecast outcomes.*

	h.o.1	w.e.1	h.o.2	w.e.2	paMIN.1	paMIN.2	rai.1	rai.2	dep	anx	stress	aqol.total
No LLM	1.82	**1.28**	0.66	0.51	**49.74**	**36.70**	4.53	**4.10**	3.75	3.03	5.63	10.42
Phi-2	1.55	1.35	**0.56**	**0.46**	53.45	57.14	**3.75**	5.52	3.33	2.75	6.63	**9.72**
Meditron	1.92	1.92	1.03	0.54	74.39	42.50	4.57	5.65	2.78	2.99	6.06	10.16
Llama-2	**1.37**	1.78	0.96	0.57	50.73	51.57	7.51	5.15	**2.77**	**1.20**	**4.69**	12.64

• **Self-Determination and Motivation Across Different Environments:** We noted a moderate correlation (r = 0.501) between self-determination levels in activities conducted during & outside rehab (rai.1 vs. rai.2), pointing to a degree of continuity in motivational factors across various contexts but also highlighting the importance of tailored motivational strategies.

• **Mental Health Indicators and Quality of Life:** Our analysis confirmed the strong interconnectedness of mental health indicators (depression, anxiety, and stress) with each showing significant correlations (r ranges from 0.659 to 0.731). These indicators inversely relate to quality of life (aqol.total; r = -0.612 for depression), reinforcing the critical role of mental health support in rehabilitation to enhance overall well-being.

• **Exercise, Mental Health, and Quality of Life:** Insights suggest that while exercise (both during and outside of rehabilitation) correlates with improved quality of life (aqol.total; r = 0.188), it also relates to varying levels of depression, anxiety, and stress, indicating subtle effects of exercise on mental health.

We next extend our analysis to evaluating the training effectiveness of models augmented with LLM-generated labels, specifically, the effect of LLMs' zero-shot classification in enriching therapy session analysis. This examination reveals how different LLMs (Phi-2, Meditron, Llama-2) contribute uniquely to understanding therapeutic dialogues, their capability for generating structured summaries, and their impact on predicting self-reported outcomes. Analysis of LLM summaries using VADER sentiment analysis & NLTK narrative complexity provides useful insights into patient engagement & help capture contextual information and dialogue dynamics. The models' performance, detailed by MAE values in Table 2, shows the benefits of LLM insights in evaluating patient progress and therapy effectiveness. In our comparative analysis of LLM-augmented models against a baseline without LLM summaries (keeping all other parameters constant), we observed that both the LLM-enhanced models (Phi-2 and Llama-2) and the baseline (No LLM) each secured the best performance in 4 out of 12 features. This parity across models highlights the role of LLMs in therapy outcome prediction and the complexity of leveraging natural language processing (NLP) techniques in clinical settings. The observed MAE values for the two variables (paMIN.1, and paMIN.2), ranging between 0-540, fall between 36-75, which initially may seem high. This high value can be attributed to the small dataset size and outlier presence. In such datasets, outliers disproportionately affect the MAE while still not affecting the overall comparison.

With just 2.7B parameters, Phi-2 has been demonstrated as the best-performing model in the sub-13B parameter model category [29]. The analysis of its summaries reveals an emphasis on positive sentiment and complex narrative generation, suggesting that a detailed and forward-looking analysis of therapy sessions can be pivotal in understanding patient progress. This is reflected in its top performance in 4 specific areas. Such alignment with therapeutic goal-setting may show the potential of positive, complex summaries to enhance prediction accuracy for certain patient outcomes. Conversely, Meditron, despite being a Llama-2 7B fine-tuned on medical data, did not outperform its counterparts in more than 4 areas. This observation **might challenge the assumption that domain-specific finetuning universally improves model performance in clinical applications**. Instead, it shows the importance of how the model's summarization approach aligns with the predictive tasks at hand, with Meditron's balanced sentiment and narrative complexity providing a practical but perhaps less predictive angle on therapeutic dialogues. However, the less consistent future orientation in the summaries may omit long-term patient objectives, which could be critical for prediction accuracy and could possibly result in a higher MAE compared to the other two LLM's. Llama-2's focus on identifying negative sentiments and presenting a mixed narrative complexity demonstrates a keen sensitivity to session challenges. Its best performance in 4 areas suggests that recognizing negative sentiments and complexities in therapy sessions may correlate with specific outcomes, offering valuable insights into patient struggles that are crucial for prediction. However, this approach's varied recognition of therapeutic barriers emphasizes the challenge of ensuring consistency in predictive performance. The equal distribution of top performances across models, including the baseline, points to a broader indication: the integration of LLMs into predictive modeling for therapy outcomes is not a straightforward enhancement but a delicate augmentation that requires careful consideration of the specific features and contexts where LLMs can provide the most significant benefits. The quality of LLM-generated summaries, characterized by sentiment, narrative complexity, and future orientation, plays a crucial role in their predictive utility. This prompts further exploration into when LLMs best enhance outcome prediction. The integration of LLMs offers a complementary tool rather than a universal solution, enhancing traditional methods where their capabilities align most closely with the predictive task's demands.

5. Conclusion

This study shows the relationships between exercise, mental health, and well-being in rehabilitation settings. We found that motivational strategies tailored to the patient's environment and personal motivations can significantly improve outcomes. Our analysis of Large Language Models (LLMs) in therapy sessions indicates their potential to enhance our understanding of patient dialogues. However, the effectiveness of LLMs depends on their alignment with the specific therapeutic needs, particularly in their sentiment analysis and narrative complexity. While LLMs have potential in improving health outcome predictions, their practical application in clinical settings requires careful implementation to match the unique demands of each case. Our research highlights that incorporating LLM-generated summaries can lead to more personalized and effective treatment plans in rehabilitation. This approach promises to enhance both clinical decision-making and patient care.

6. References

[1] WHO, "Cardiovascular diseases (cvds)," https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), June 2021.

[2] ——, "Chronic obstructive pulmonary disease (copd)," https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd), March 2023.

[3] C. J. Lavie and R. V. Milani, "Effects of cardiac rehabilitation, exercise training, and weight reduction on exercise capacity, coronary risk factors, behavioral characteristics, and quality of life in obese coronary patients," *The American journal of cardiology*, vol. 79, no. 4, pp. 397–401, 1997.

[4] F. Pitta, T. Troosters, V. S. Probst, D. Langer, M. Decramer, and R. Gosselink, "Are patients with copd more active after pulmonary rehabilitation?" *Chest*, vol. 134, no. 2, pp. 273–280, 2008.

[5] A. A. Chatziefstratiou, K. Giakoumidakis, and H. Brokalaki, "Cardiac rehabilitation outcomes: modifiable risk factors," *British Journal of Nursing*, vol. 22, no. 4, pp. 200–207, 2013.

[6] N. M. Hamburg and G. J. Balady, "Exercise rehabilitation in peripheral artery disease: functional impact and mechanisms of benefits," *Circulation*, vol. 123, no. 1, pp. 87–97, 2011.

[7] M. M. McDermott, "Exercise rehabilitation for peripheral artery disease: a review," *Journal of cardiopulmonary rehabilitation and prevention*, vol. 38, no. 2, p. 63, 2018.

[8] C. L. Rochester, "Exercise training in chronic obstructive pulmonary disease." *Journal of Rehabilitation research & development*, vol. 40, no. 5, 2003.

[9] T. M. Edenfield and J. A. Blumenthal, "Exercise and stress reduction," *The handbook of stress science: Biology, psychology, and health*, pp. 301–319, 2011.

[10] A. K. Banerjee, S. Okun *et al.*, "Patient-reported outcome measures in safety event reporting: Prosper consortium guidance," *Drug safety*, vol. 36, pp. 1129–1149, 2013.

[11] H. Touvron, L. Martin *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[12] Z. Chen, A. Hernández-Cano *et al.*, "Meditron-70b: Scaling medical pretraining for large language models," November 2023. [Online]. Available: https://github.com/epfLLM/meditron

[13] M. Javaheripi, S. Bubeck *et al.*, "Phi-2: The surprising power of small language models, 2023," *URL https://www. microsoft. com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models*.

[14] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[15] B. N. Suhas, S. Rajtmajer, and S. Abdullah, "Differential Privacy enabled Dementia Classification: An Exploration of the Privacy-Accuracy Trade-off in Speech Signal Data," in *Proc. INTERSPEECH 2023*, 2023, pp. 346–350.

[16] T. Bickmore and T. Giorgino, "Health dialog systems for patients and consumers," *Journal of biomedical informatics*, vol. 39, no. 5, pp. 556–571, 2006.

[17] B. N. Suhas and S. Abdullah, "Privacy sensitive speech analysis using federated learning to assess depression," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6272–6276.

[18] A. Udvardi, "The role of linguistics in improving the evidence base of healthcare communication," *Patient Education and Counseling*, vol. 102, no. 2, pp. 388–393, 2019.

[19] Y. Netz, M.-J. Wu, B. J. Becker, and G. Tenenbaum, "Physical activity and psychological well-being in advanced age: a meta-analysis of intervention studies." *Psychology and aging*, vol. 20, no. 2, p. 272, 2005.

[20] P. Enderby, A. John, and B. Petheram, *Therapy outcome measures for rehabilitation professionals: speech and language therapy, physiotherapy, occupational therapy*. John Wiley & Sons, 2013.

[21] D. Markland and V. Tobin, "A modification to the behavioural regulation in exercise questionnaire to include an assessment of amotivation," *Journal of Sport and Exercise Psychology*, vol. 26, no. 2, pp. 191–196, 2004.

[22] M. H. Lee, D. P. Siewiorek *et al.*, "A human-ai collaborative approach for clinical decision making on rehabilitation assessment," in *Proceedings of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–14.

[23] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[24] C. L. Craig, A. L. Marshall, M. Sjöström, A. E. Bauman, M. L. Booth, B. E. Ainsworth, M. Pratt, U. Ekelund, A. Yngve, J. F. Sallis *et al.*, "International physical activity questionnaire: 12-country reliability and validity," *Medicine & science in sports & exercise*, vol. 35, no. 8, pp. 1381–1395, 2003.

[25] J. D. Henry and J. R. Crawford, "The short-form version of the depression anxiety stress scales (dass-21): Construct validity and normative data in a large non-clinical sample," *British journal of clinical psychology*, vol. 44, no. 2, pp. 227–239, 2005.

[26] A. Maxwell, M. Özmen, A. Iezzi, and J. Richardson, "Deriving population norms for the aqol-6d and aqol-8d multi-attribute utility instruments from web-based data," *Quality of life research*, vol. 25, pp. 3209–3219, 2016.

[27] J. Richardson, A. Iezzi, M. A. Khan, and A. Maxwell, "Validity and reliability of the assessment of quality of life (aqol)-8d multi-attribute utility instrument," *The Patient-Patient-Centered Outcomes Research*, vol. 7, pp. 85–96, 2014.

[28] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio," in *Proc. INTERSPEECH 2023*, 2023, pp. 4489–4493.

[29] M. Javaheripi and S. Bubeck, "Phi-2: The surprising power of small language models," https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/, Microsoft Research Blog, December 2023, accessed: 2023-03-10.