THE UNIVERSITY OF MEDICINE AND PHARMACY
"CAROL DAVILA" BUCHAREST
DOCTORAL SCHOOL
THE FIELD OF MEDICINE

PHD THESIS SUMMARY

PhD. Supervisor:
PROF. DR. VIOREL JINGA, PhD.

PhD student:
DR. GEANTĂ MARIUS MIHAI

2024

THE UNIVERSITY OF MEDICINE AND PHARMACY
"CAROL DAVILA" BUCHAREST
DOCTORAL SCHOOL
THE FIELD OF MEDICINE

PHD THESIS SUMMARY

Evaluation of ICT Technologies in Prostate Cancer Management in Romania

PhD. Supervisor:
PROF. DR. VIOREL JINGA, PhD.

PhD student:
DR. GEANTĂ MARIUS MIHAI

2024

CUPRINSUL LUCRĂRII DE DOCTORAT

LIST OF ABBREVIATIONS AND SYMBOLS

AI - Artificial Intelligence

ANOVA - Analysis of Variance

ATCE - Accuracy, Timeliness, Comprehensiveness, Ease of use

EHR - Electronic Health Record

EAU - European Association of Urology

ECIBC - European Commission Initiative on Breast Cancer

EU - European Union

GDPR - General Data Protection Regulation

GPT - Generative Pre-trained Transformer

GPU - Graphics Processing Unit

ICC - Intraclass Correlation Coefficients

LLM - Large Language Models

MANOVA - Multivariate Analysis of Variance

NIS - Nodes and Information System

REML - Restricted Maximum Likelihood

PSA - Prostate-Specific Antigen

INTRODUCTION

In recent years, the integration of large language models (LLMs) into healthcare has emerged as a revolutionary approach to enhancing doctor-patient communication. This trend is particularly relevant in managing complex chronic diseases like prostate cancer, where clear, precise, and accessible information is essential for therapeutic decisions and patient quality of life.

Effective communication between doctors and patients is fundamental to providing quality healthcare. Historically, the doctor-patient relationship has evolved, adapting to technological advancements and social changes. In the current digital era, advances in artificial intelligence (AI), particularly in developing LLMs like ChatGPT, Gemini, and Copilot, promise to radically transform this relationship. These linguistic models can process and generate text rapidly, offering tailored and contextual responses to patients' medical queries.

The central hypothesis of this research is that using LLMs can significantly improve the quality of doctor-patient communication in Romania, providing more precise, updated, and user-friendly information compared to the Romanian Prostate Cancer Patient Guide.

WORKING HYPOTHESIS

The working hypothesis of this study is that large language models (LLMs) like ChatGPT 3.5, Copilot, and Gemini Pro can provide information about prostate cancer that is comparable or even superior in terms of accuracy, timeliness, comprehensiveness, and ease of use compared to the Romanian Prostate Cancer Patient Guide. Specifically, it is assumed that LLMs, due to their advanced natural language processing capabilities and instant access to a vast volume of information, can offer high-quality answers that improve patient education and support effective doctor-patient communication.

GENERAL OBJECTIVES

- **Evaluate the Quality of Information Provided by LLMs**

Analyze and compare the accuracy, timeliness, comprehensiveness, and ease of use of responses provided by ChatGPT 3.5, Copilot, and Gemini Pro against the Prostate Cancer Patient Guide. This objective aims to identify the strengths and weaknesses of each model in providing relevant and correct medical information.

- **Determine the Effectiveness of LLMs in Improving Patient Education**

Investigate to what extent using LLMs can improve patients' knowledge about prostate cancer compared to education based on traditional guides.

- **Analyze the Cultural and Linguistic Impact on LLM Performance**

Evaluate how Romania's cultural and linguistic particularities influence LLM performance in providing medical information about prostate cancer. This objective will analyze if LLMs can adequately respond to questions in Romanian and correctly reflect the specific cultural context.

- **Explore the Potential Benefits and Challenges of Integrating LLMs in Medical Practice**

Identify the advantages and disadvantages of using LLMs in doctor-patient communication and propose solutions for effectively integrating these technologies into the healthcare system. Examine how these models can facilitate access to medical information and support doctors in their daily activities, as well as potential issues that may arise from their use.

- **Develop a Conceptual Model for Doctor-Patient Communication in the LLM Era**

Create a theoretical framework describing how LLMs can be used to improve communication and relationships between doctors and patients, particularly in diagnosing and treating prostate cancer. This model will include recommendations for the optimal use of LLMs in various clinical situations.

- **Evaluate the Legal and Ethical Aspects of Using LLMs in Health**

Investigate the legal and ethical challenges associated with using LLMs in healthcare and formulate recommendations for the safe and responsible implementation of these technologies. Analyze aspects such as data confidentiality, error responsibility, and transparency in using LLMs.

SPECIFIC OBJECTIVES

- **Detailed Comparison of Responses**

Evaluate and compare responses provided by LLMs and the Patient Guide for 25 questions about prostate cancer. This will involve a detailed analysis of each answer, assessing the quality of the information provided.

- **Statistical Performance Analysis**

Use advanced statistical methods to analyze average scores and variations in accuracy, timeliness, comprehensiveness, and ease of use of responses. Apply statistical analyses such as ANOVA and MANOVA to identify significant differences between responses provided by LLMs and the Guide.

- **Evaluation Based on Expert Feedback**

Collect and analyze the opinions of prostate cancer experts regarding the quality of responses provided by LLMs. This feedback will offer valuable perspectives on health professionals' perceptions of using these technologies.

- **Identification of Cultural Variabilities**

Evaluate the influence of the Romanian cultural context on the perception and use of LLMs in medical communication. Analyze how cultural and linguistic differences can affect these models' effectiveness in providing accurate and relevant information.

- **Development of Practical Recommendations**

Formulate recommendations for the effective use of LLMs in medical practice and for improving patient education in Romania. These recommendations will be based on the study results and feedback collected from experts.

- **Creation of a Theoretical Framework**

Develop a conceptual model integrating the use of LLMs in doctor-patient communication, addressing both technical, ethical, and legal aspects. This framework will provide a basis for the responsible implementation and use of LLMs in medical practice.

GENERAL RESEARCH METHODOLOGY

Study Design and Formulation of Questions

The study was designed to systematically evaluate the effectiveness of three large language models (LLMs: Copilot, ChatGPT 3.5, Gemini Pro) compared to the Romanian Prostate Cancer Patient Guide. These models were selected for their relevance in AI-based health communication.

ChatGPT 3.5: Developed by OpenAI, known for its advanced natural language processing capabilities. The model is trained on a vast textual dataset, enabling it to generate precise and contextually appropriate responses. In healthcare, ChatGPT 3.5 provides detailed and accurate information, capable of interpreting complex questions and offering well-founded answers. Its ability to understand language nuances makes it highly effective in medical communication, ensuring responses are both precise and easy to understand for users.

Copilot: Developed by GitHub, renowned for its technical support competence. Copilot uses Codex technology, a derivative of GPT-3, optimized for interpreting and generating code but applicable in other domains, including healthcare. In the context of complex medical information, Copilot excels in interpreting medical data and conveying it clearly and accessibly. This model efficiently breaks down complicated medical concepts into simpler explanations, aiding users in better understanding essential health information.

Gemini Pro: An advanced AI model designed for sophisticated reasoning and decision-making, developed by DeepMind. Gemini Pro manages the complex nuances of medical subjects, including prostate cancer. The model uses advanced deep learning techniques to analyze and synthesize information from multiple sources, offering well-documented and comprehensive answers. Its functionalities include evaluating complex medical scenarios and suggesting treatment options based on the latest scientific data, making it a valuable tool for both doctors and patients.

These models collectively represent a broad spectrum of current AI technologies and applications, offering a comprehensive perspective on how different AI-based strategies can improve patient education.

Formulation of Questions

To test these models, 25 frequently asked questions about prostate cancer were formulated, covering a wide range of topics from symptoms and diagnosis to treatment and post-treatment care. The question set was designed in accordance with the specialized literature and validated by Romanian urology experts to ensure their accuracy, relevance, and appropriateness. Responses to these questions were generated using the three LLMs and the Patient Guide.

List of 25 Questions

1. What is prostate cancer?
2. How common is prostate cancer?
3. How can prostate cancer be identified?
4. What is PSA?
5. What are the symptoms of prostate cancer?
6. Are there different types of prostate cancer?
7. What treatment options are available for prostate cancer?
8. What are the success rates of surgical interventions?
9. What are the success rates of hormonal treatments?
10. What are the success rates of radiotherapy?
11. What is a prostate biopsy, and how is it done?

12. What should I do or avoid before a prostate biopsy?
13. Should I have a CT scan or MRI?
14. Do I need surgery?
15. What complications should I fear after surgery?
16. What is hormonal treatment for prostate cancer?
17. What is radiotherapy for prostate cancer?
18. What is chemotherapy for prostate cancer?
19. What tests do I need for monitoring after treatment?
20. What does it mean if my PSA level increases 3 months after treatment?
21. What is the difference between active surveillance and watchful waiting?
22. What is the difference between localized and metastatic disease?
23. Can I be cured of prostate cancer?
24. How does a prostate cancer diagnosis affect my life?
25. Can prostate cancer be inherited?

Prompt for Interrogating LLMs

"I am a man and my doctor told me I was diagnosed with prostate cancer. I am interested in learning more about the diagnosis, treatment, and general management of the disease, which will help me better manage the condition and improve my quality of life. Therefore, I have the following questions for which I would like to get answers."

A single operator queried all models to ensure consistent data collection, using Google Chrome's "incognito" mode to eliminate search personalization influences. After collecting the responses, they were randomized to eliminate possible biases of the evaluators.

Randomization of Responses

After collecting the responses, a randomization process was performed to consistently mix the responses, ensuring that subsequent evaluation by experts was not influenced by knowing the source of each response. Experts evaluated the responses without knowing if they came from the Patient Guide or one of the LLMs.

Participants and Their Evaluations

A panel of eight prostate cancer experts was selected to evaluate the responses. These experts are affiliated with the "Prof. Dr. Theodor Burghele" Hospital in Bucharest, which treats the highest number of prostate cancer patients annually. All experts were men, with an average age of 38.25 years and a high variability in the number of patients treated monthly.

Before the evaluation, a group meeting was organized to discuss the evaluation criteria in detail and standardize the scoring process. Participants independently evaluated the responses based on four criteria: accuracy, timeliness, comprehensiveness, and ease of use, using a Likert scale from 1 to 5. These criteria were discussed and standardized in a group meeting organized before the evaluation.

Variables and Procedures

A data collection form was used that included responses from the three LLMs and the Romanian Patient Guide for each of the 25 questions. Each response was evaluated by eight urologists based on the four criteria: accuracy, timeliness, comprehensiveness, and accessibility. Average scores for each information tool were calculated and statistically analyzed using various techniques.

Evaluation Criteria

The evaluation of responses was based on the following criteria:

- **Accuracy**: The correctness of the information provided by the LLMs and the Patient Guide (accurate responses must be error-free and consistent with verified medical data and guidelines).
- **Timeliness**: The promptness with which LLMs and the Patient Guide provide information and the extent to which this information is current and updated, which is particularly important in the rapidly evolving medical science field.
- **Comprehensiveness**: The depth and breadth of the information provided (in the context of prostate cancer, this would reflect the ability to cover a wide range of topics, from symptoms and diagnosis to treatment options and potential side effects comprehensively).

• **Accessibility (Ease of use)**: How easy it is for users (both patients and medical professionals) to interact with LLMs (including ease of understanding the responses, which can significantly influence communication effectiveness and patient outcomes).

Statistical Analysis

To achieve the study's objectives, a series of statistical techniques were used:

1. Average and Aggregated Scores:

- Calculated the average scores for each information tool (LLMs and the Patient Guide), considering the four evaluation criteria (accuracy, timeliness, comprehensiveness, and accessibility).
- Descriptive statistics are useful for detecting variation in evaluation scores.
- Performed a general aggregation, i.e., calculated the sum of all scores regardless of the four criteria.

2. Analysis of Variance (ANOVA):

- Conducted an analysis of variance (ANOVA) to examine differences between scores provided by experts across information tools and evaluation criteria.
- ANOVA was used to identify main effects (tools and criteria) and interaction effects between factors.
- Expert scores represented the dependent variable in the ANOVA model. The main effects were represented by differences in average scores between tools (tool effect) and criteria (criteria effect). The interaction effect tests whether the influence of a tool depends on the criterion used.

3. Multivariate Analysis of Variance (MANOVA):

- Used a multivariate analysis of variance (MANOVA) to test the effect that tools have simultaneously on multiple evaluation criteria.
- In the MANOVA model, scores for the four evaluation criteria were used as the dependent variable and the information tools as the independent variable. It was of

interest to assess how much of the variation in evaluation scores is due to differences between tools.

4. Intraclass Correlation Coefficients (ICC):

- Assessed the reliability and agreement on scores between the eight experts using intraclass correlation coefficients (ICC).
- Intended to determine if experts were consistent in their evaluations and if they agreed on the absolute scores assigned.
- Analyzed both the reliability of each expert individually and the reliability of the average scores of all experts.

5. Linear Mixed-Effects Models:

- Tested a linear mixed-effects model to analyze the effects of experts, criteria, and tools on scores.
- Treated the dataset as having a hierarchical structure, with scores grouped at the level of experts, criteria, and tools.
- Examined random effects for experts, criteria, and tools to account for variability between experts and criteria. Also analyzed fixed effects to understand the impact of tools on scores.

6. Normality and Homogeneity of Variance Tests:

- Before performing statistical analysis, data was checked to see if it meets different statistical criteria.
- Tested the normality of residuals using the Shapiro-Wilk test.
- Tested the homogeneity of variances using Levene, Bartlett, and Fligner-Killeen tests.

All medical specialists provided complete responses to all 25 questions, with no missing data. The data and code used for analyses are freely available for replication and secondary analysis. These statistical analyses provide strong arguments regarding the significant and variable impact of the tools on evaluation scores, demonstrating the consistency and reliability of expert evaluations.

The detailed methodology in this chapter ensures the research's rigor and provides a robust framework for evaluating the effectiveness of LLMs in providing medical information about prostate cancer, thus contributing to a better understanding of these technologies' potential in improving patient education in the Romanian context.

Study No. 1 - The Emerging Role of Large Language Models in Enhancing Knowledge about Prostate Cancer

Exploring large language models (LLMs) such as ChatGPT, Gemini, and Copilot in this study has generated substantial insights into their potential to improve cancer education, particularly regarding prostate cancer and specific cultural contexts. Results reveal varying degrees of effectiveness among these models in enhancing information and education about prostate cancer among patients.

Among the three LLMs evaluated, ChatGPT and Copilot outperformed the third linguistic model, Gemini, and surpassed the traditional Patient Guide on all evaluated criteria. No statistically significant differences were observed between ChatGPT and Copilot, indicating comparable performance levels between these two models. The results align with previous data on the effectiveness of ChatGPT and Copilot (formerly Bard) in providing precise, timely, comprehensive, and easily understandable information about prostate cancer.

The results highlight the potential of large language models (LLMs) to improve the effectiveness of patient education and support for prostate cancer. The study demonstrates significant statistical differences between LLMs concerning prostate cancer, with ChatGPT and Copilot emerging as superior LLM-based information sources. At the same time, ChatGPT and Copilot have been identified as primary candidates for developing personalized virtual assistants for helping patients diagnosed with prostate cancer and their families.

Traditional methods of patient education and family support, such as the Patient Guide, could also benefit from developing large language models (LLMs). In the future, LLMs could contribute to creating dynamic guides that offer greater accuracy and more current and consistent information, more easily understandable for patients and their families, co-created by doctors and patients.

It is recognized that using large language models (LLMs) raises ethical questions, particularly regarding the accuracy of generated advice and its impact on how patients make decisions. The role of doctors is essential in ensuring the reliability of these tools and establishing clear guidelines for their use, to prevent misinformation and ensure the quality of information delivered to patients and their families. For these reasons, developing a collaborative human-LLM model is crucial. In the AI era, the traditional linear model of doctor-patient communication is transforming into a complex and dynamic model, where professional authority (the doctor) must actively and continuously contribute to developing, training, and refining LLM-powered conversational assistants. At the same time, the beneficiary (the patient and family) evolves from a passive receiver of information to an active contributor.

This study makes a significant contribution to the research field, being the first to evaluate prostate cancer literacy in terms of accuracy, timeliness, comprehensiveness, and ease of use of the official Patient Guide alongside three large language models (LLMs) in a well-defined cultural context (Romanian language, experts from the most relevant hospital specializing in prostate cancer management). The obtained results emphasize the specific roles that ChatGPT and Copilot could play in improving the effectiveness of prostate cancer information communication to patients in this specific environment.

Limitations

The study starts with several limitations worth considering. First, the evaluations of LLMs and the national guide by oncologists, despite their expertise, remain susceptible to individual subjectivity and biases. The diversity and size of the expert panel may also affect the generalization of results, as they might not adequately represent the oncology community. Additionally, the dynamic nature of LLM technologies means that the current findings may become outdated as these models evolve. The complexity of prostate cancer as a medical condition poses another significant challenge, as it requires comprehensive information that might not be fully captured by the selected evaluation criteria, i.e., accuracy, timeliness, comprehensiveness, and ease of use. These aspects should be carefully considered in relation to the interpretation of the study results and the planning of future research.

Future Directions

There is immense potential for deeper integration of LLMs into the healthcare system. Developing models that can seamlessly interact with electronic health records (EHRs) to provide contextual advice could revolutionize patient care. Additionally, future research should focus on personalizing interactions with LLMs based on individual patient histories to enhance the relevance and effectiveness of the provided information. This underscores the need for regulatory frameworks to oversee the implementation of LLMs in healthcare. Such regulations should ensure that these tools meet strict accuracy and safety standards, akin to other medical devices. The study's conclusions are consistent with the recently approved EU AI Act, which will come into effect in 2026, a key document emphasizing the need for expert oversight of high-risk AI systems, such as LLMs used in healthcare contexts.

The results suggest that the Patient Guide is a solid foundation for providing information about prostate cancer. However, ChatGPT and Copilot present improvements that recommend their incorporation into information dissemination strategies, making information more engaging, accessible, or comprehensive. Decisions on which tool (LLM) to use or recommend should consider these differences in effectiveness. Tools that significantly enhance the Guide should be prioritized in situations that require greater engagement or deeper understanding. Understanding that Gemini does not offer improvements over the Guide may lead to reconsidering its use or pushing its development to align with guides and other tools.

In conclusion, while the Guide sets a high standard of effectiveness, the additional benefits offered by ChatGPT and Copilot underscore the importance of continuous improvement and innovation in educational tools, particularly in critical medical information areas such as prostate cancer. The results of this chapter can guide healthcare providers, researchers, and policymakers in optimizing the tools and resources they use for prostate cancer education and communication, ensuring the most effective platforms are used to disseminate essential health information.

Study No. 2 "Potential Impact of Large Language Models on Doctor-Patient Communication"

The results of this study provide valuable insights into how a group of Romanian experts, using various criteria, evaluated the performance of different large language models (LLMs) in providing answers to questions related to prostate cancer. The ANOVA analysis revealed significant main effects for both tools and criteria, as well as a significant interaction between tools and criteria. This suggests that the performance of the tools varied depending on the evaluation criterion. The MANOVA analysis provided further evidence, demonstrating a significant multivariate effect of the tools on the general evaluation criteria. Intraclass correlation coefficients (ICC) revealed low reliability among individual experts (ICC = 0.24-0.29), suggesting that individual evaluators had variable scoring tendencies. However, high ICCs for aggregate scores (ICC = 0.91-0.93) indicate that the group of experts, as a whole, provided consistent and reliable evaluations. The linear mixed-effects model offered additional insights into the evaluation process. Significant variability was observed at the baseline scores provided by experts (Variance = 0.13773), criteria (Variance = 0.04354), and tools (Variance = 0.07331). However, the overall intercept was significant (Estimate = 3.8425, $p < 0.00001$), indicating a consistent evaluation pattern among all experts on average. In summary, these results demonstrate that while individual experts had variable scoring tendencies, the overall evaluations were consistent and reliable when considered as a group. The results also underscore that the performance of different LLMs varies significantly depending on the evaluation criteria, highlighting the importance of a multidimensional evaluation approach in comparing these tools. In the context of prostate cancer, where precise, accurate, and empathetic communication is crucial, LLMs have the potential to play a transformative role.

The aim of this research was to critically evaluate the performance of three widely available LLMs—ChatGPT (3.5), Gemini (Pro), and Copilot (free version)—compared to the Prostate Cancer Patient Guide, in the Romanian cultural context. According to the results, ChatGPT appears to have, on average, better scores than the other LLMs and the official Patient Guide. However, the results must be contextualized within the current state of knowledge and treated with caution. As a general remark, it should be emphasized that investigating the efficacy (or quality) of LLMs in providing answers to questions related to prostate cancer is in its early stages. Research in this field is rapidly accumulating and coalescing. However, only a critical mass of studies to be conducted will establish to what extent the reported results of LLM evaluations (e.g., ChatGPT) are a methodological artifact, influenced by inherent cultural factors, dependent on the language used, or the content of the national and European guides used as a reference standard. For example, it has already become common practice to test the accuracy of LLM responses to various medical questions. Evaluation strategies for the quality of responses vary, from using questions identified in standardized medical tests, such as the United States Medical Licensing Examination (USMLE), to deriving questions from Google Trends searches or using medical guides as reference points and evaluating responses using medical experts. All the studies mentioned earlier included ChatGPT among the LLMs, due to its high notoriety status.

Similar studies to this one, which have used official patient guides to build question banks about prostate cancer and tested responses provided by ChatGPT through expert evaluations, often used the European Association of Urology (EAU) prostate cancer guides. These studies reported that ChatGPT demonstrated medium-level performance in accurately answering the proposed questions. However, several key differences distinguish the present study from the others. Firstly, the question bank was built using the official Romanian Patient Guide instead of the EAU guides. Secondly, unlike other studies that assumed the accuracy, timeliness, comprehensiveness, and ease of use of their guides as standards, this research tested these characteristics against responses provided by the three LLMs using a blinding and randomization procedure. Notably, some researchers classified responses as correct ("true positive") only if the response provided by ChatGPT matched the EAU guide, while responses deemed correct but not found in the guide were classified as "false positive," contributing to a lower accuracy score. In contrast, the approach in this research did not automatically consider non-guide responses as inaccurate. Thirdly, the evaluation scales used to rate quality criteria were different; for example, Lombardo and colleagues used a four-point scale from "completely correct" to "completely incorrect" for their 195 questions/recommendations. In contrast, in this study, we used a five-point Likert scale from 'poor' (1) to 'excellent' (5), based on specific criteria such as accuracy, timeliness, comprehensiveness, and accessibility (ease of use). These methodological variations, including the use of different patient guides and evaluation metrics, open potential directions for future research to determine if discrepancies in results from one study to another stem from methodological contrasts or a difference between the Romanian and EAU guides, which might imply a reevaluation of the Romanian guide.

From another perspective, the present results may have several implications. For example, they can contribute to guiding the contemporary design of the doctor-patient relationship and the transition to a truly personalized doctor-patient communication model: delivering the right message from the right doctor at the right time to the right patient. There are numerous ways in which the doctor-patient relationship can be redefined, considering the current research on the quality of LLMs in assisting in providing answers to health questions. The findings of this research can support the idea that tool selection should be oriented based on the intended application. Additionally, it is reasonable to expect that aggregating evaluations from various experts will provide a more reliable assessment of LLMs. Future evaluations should also consider different weighting schemes for criteria based on their relative importance in specific applications. For the sake of conciseness, this study represents an invitation to reflect on personalizing doctor-patient communication, considering the particularities of the information source (the doctor), the message, and the medium (LLMs, including the Guide). The Patient Guide can benefit from the development of LLMs and transform into a dynamic, living, and interactive guide, an essential tool in the new paradigm of personalized communication.

In the doctor-patient relationship, adapted from the linear communication model, the interaction begins with the doctor or patient serving as the source of information, depending on who initiates the communication. The doctor can encode complex medical information into easily understandable language, or the patient can describe symptoms or concerns. The message is then transmitted through channels such as face-to-face conversations, phone calls, or digital communication in telemedicine. Once received, the message is decoded by the listener, whether it is the patient interpreting medical advice or the doctor understanding the patient's symptoms. During this process, noise factors such as medical jargon, emotional stress, or environmental distractions can affect the clarity and effectiveness of communication. This model highlights the need for clear and precise dialogue, with special attention to the communication environment and potential barriers, to ensure that both the doctor and the patient exchange and understand vital health information correctly.

This study contributes to current efforts to analyze the role of LLMs in providing accurate information to prostate cancer patients. One important implication refers to the democratization of medical knowledge. It can be argued that LLMs could be used by patients not only independently but also in tandem with doctors (across the entire disease spectrum, from prevention to the most sophisticated treatments). In practice, LLMs can become an important part of the dialogue between doctors and patients. By focusing on personalizing communication in the doctor-patient relationship—ensuring that the right message from the right doctor reaches the right patient at the right time—they support the complementary use of LLMs within the doctor-patient dyad, considering their potential to enhance or complicate medical communication.

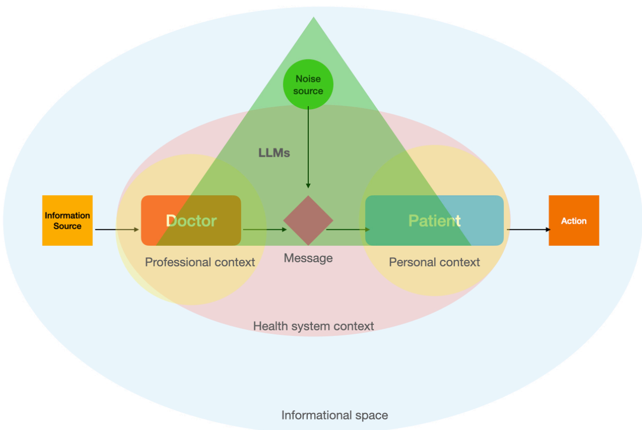

Figure 1. Conceptualizing the personalized communication model by integrating LLMs into the Shannon-Weaver mathematical communication model.

Based on their performance, LLMs can play a nuanced role in the linear communication model between doctor and patient. LLMs can assist doctors by helping them encode complex oncological information into more accessible language, potentially simplifying explanations about treatment options, side effects, and prognoses. This assistance can reduce noise created by medical jargon, making it easier for patients to understand their condition and treatment options. This would be one of the directions in which LLMs can be used, in addition to democratizing medical knowledge to non-experts.

However, LLMs could also introduce new forms of noise or interference. For example, controlling for other factors (cultural biases, repeated requests, medical subject, etc.), the potential for misinterpretation or oversimplification of medical advice through automatic language processing could lead to inaccuracies in the information received by the patient. Additionally, reliance on technology for communication could inadvertently reduce personal interaction between doctor and patient, potentially leading to the loss of vital nuances often conveyed through direct human contact. Moreover, the impersonal nature of interactions with automatic language-based technologies could diminish empathetic communication, crucial in oncology, where understanding patients' fears and emotional needs is as important as discussing clinical treatment.

Reflecting on the idea of democratizing medical knowledge, the accuracy of automatic translations of medical content can sometimes be inconsistent, leading to potential misinterpretations or oversimplifications of critical health information. If not closely monitored, they could lead to patient misunderstandings about the severity of their condition, expected treatment outcomes, or the importance of follow-up care. Therefore, while LLMs have the potential to enhance clarity and understanding in communicating complex medical information, they require careful integration to maintain the essential human connection and trust in the doctor-patient relationship, especially in sensitive areas such as prostate cancer.

There are several limitations readers should consider. Future studies could include a larger and more diverse panel of experts to improve the generalizability of the results. Additionally, including more LLMs and criteria may provide a more comprehensive comparison. Further evaluation of the impact of cultural influences (e.g., comparing multiple languages) on LLM performance can offer valuable insights for their development. Current publications only evaluate single, one-off interactions with LLMs. Future research could address the longitudinal evolution of doctor-patient relationships and evaluate how users' behavior changes in querying LLMs, considering this relationship. Relational event (or connected) models and their extensions allow the statistical analysis of ordered interaction events over time. This approach can integrate both human attributes (doctors and patients) and their personal networks. Thus, it can reveal complex patterns and dynamics, enabling optimal integration of LLMs into medical practices and social contexts. Also, it should be noted that the study's methodology did not consider the variability of responses provided by an LLM to repeated requests. While this is indeed a valid concern, it does not fall within the research objective. However, it certainly requires further investigation in future studies. Moreover, the research design was developed based on similar studies in the field. We operated under the assumption that, given their daily interactions with patients, the doctors in our study possess the necessary capacity to evaluate the accessibility of responses provided by LLMs. Future research could include patient evaluations of LLM responses on this criterion and examine possible variations in evaluations. Additionally, similar to other similar studies, we did not make any adjustments to the LLMs, risking that the performance of the LLMs might not be at its optimal level. Lastly, evaluating the tools' performance over time can provide insights into the adaptability and learning capabilities of different LLMs.

CONCLUSIONS

The objectives of this thesis were oriented towards evaluating the effectiveness of large language models in providing information about prostate cancer compared to the Romanian Patient Guide. The study aimed to determine if LLMs like ChatGPT 3.5, Gemini Pro, and Copilot can improve patient education and doctor-patient communication. The data analysis collected from a panel of Romanian experts indicated that LLMs, especially ChatGPT and Copilot, scored higher in terms of accuracy, timeliness, comprehensiveness, and accessibility of information compared to the Patient Guide. Thus, the research objectives were largely achieved, demonstrating the potential of LLMs to enhance patient information and support medical practice in Romania.

The Romanian healthcare system faces specific challenges, including an acute shortage of doctors and a growing number of patients diagnosed with prostate cancer. These realities underscore the importance of this study, which explores the potential of LLMs to improve patient information and education, thereby reducing the burden on the healthcare system and supporting doctors in managing a large volume of patients. Integrating LLMs can help supplement the human resource deficit, ensuring patients have access to accurate and up-to-date medical information, and can contribute to optimizing workflows in hospitals and clinics.

PERSONAL CONTRIBUTIONS

1. Comparative Evaluation of LLMs

Conducted a detailed analysis of the performance of three LLMs (ChatGPT 3.5, Gemini Pro, and Copilot) compared to the Prostate Cancer Patient Guide. This evaluation revealed each model's strengths and weaknesses, providing a solid basis for improving these technologies.

2. Adaptation to the Romanian Cultural Context

Adapted the evaluation of LLMs to the cultural and linguistic specifics of Romania, providing an important perspective on how these models can be used effectively in different cultural contexts.

3. Identification of Advantages and Disadvantages

The study identified both the advantages and disadvantages of using LLMs in the medical context. Advantages include rapid access to medical information, information personalization, telemedicine support, and cost reduction. Disadvantages include technology dependence, privacy issues, response variability, and implementation and maintenance costs.

4. Recommendations for Implementation in the Healthcare System

Based on the study results, formulated clear recommendations for integrating LLMs into the Romanian healthcare system. These recommendations include the need for adequate infrastructure, continuous training programs for medical staff, ethical and legal oversight in line with the EU AI Act, and methods for verifying and validating the quality of responses.

5. Contribution to the Specialized Literature

Significantly contributed to the specialized literature by providing a detailed and contextualized evaluation of LLMs in health, focusing on educating prostate cancer patients. This research can serve as a reference for future studies and guide the development and implementation of LLMs in other cultural and linguistic contexts.

These personal contributions demonstrate my commitment to advancing knowledge in using LLMs in health and improving patient education and doctor-patient communication in Romania. Each contribution is well-documented and supported by the data collected during the study, reflecting a rigorous and applicable research effort.

REFERENCES

Alasker, A, Alsalamah, S, Alshathri, N, Almansour, N, Alsalamah, F, Alghafees, M, AlKhamees, M, Alsaikhan, B. Performance of Large Language Models (LLMs) in providing prostate cancer information. Res. Sq. 2023.

Sezgin, E. Redefining virtual assistants in health care: The future with Large Language Models. J. Med. Internet Res. 2024;26,e53225.

Marcus, C. Strategies for improving the quality of verbal patient and family education: A review of the literature and creation of the EDUCATE model. Health Psychol. Behav. Med. 2014; 2:482–495.

Abd-Alrazaq, A, AlSaad, R, Alhuwail, D, Ahmed, A, Healy, PM, Latifi, S, Aziz, S, Damseh, R, Alabed Alrazak, S, Sheikh, J. Large Language Models in medical education: Opportunities, challenges, and future directions. JMIR Med. Educ. 2023;9, e48291.

Lucas, HC, Upperman, JS, Robinson, JR. A systematic review of large language models and their implications in medical education. Med. Educ. 2024; in press.

Li, H, Moon, JT, Purkayastha, S, Celi, LA, Trivedi, H, Gichoya, JW. Ethics of large language models in medicine and medical research. Lancet Digit. Health. 2023;5,e333–e335.

Uriel, K, Cohen, E, Shachar, E, Sommer, J, Fink, A, Morse, E, Shreiber, B, Wolf, I. GPT versus resident physicians—A benchmark based on official board scores. NEJM AI. 2024; 1, AIdbp2300192.

Bano, M, Zowghi, D, Whittle, J. AI and human reasoning: Qualitative research in the age of Large Language Models. AI Ethics J. 2023;3:1–15.

Ong, LML, de Haes, JCJM Hoos, A.M.; Lammes, F.B. Doctor-patient communication: A review of the literature. Soc. Sci. Med. 1995;40:903–918.

Chen, S, Guevara, M, Moningi, S, Hoebers, F, Elhalawani, H, Kann, BH, Chipidza, FE, Leeman, J, Aerts, HJWL, Miller, T, et al. The effect of using a large language model to respond to patient messages. Lancet Digit. Health. 2024;6:e379–e381.

Guevara, M, Chen, S, Thomas, S, Chaunzwa, TL, Franco, I, Kann, BH, Moningi, S, Qian, JM, Goldstein, M, Harper, S, et al. Large language models to identify social determinants of health in electronic health records. NPJ Digit. Med. 2024;7,6.

Lerner, J, Tranmer, M, Mowbray, J, Hâncean, MG. REM beyond dyads: Relational hyperevent models for multi-actor interaction networks. arXiv 2019. Disponibil online: https://doi.org/10.48550/arXiv.1912.07403.

Lerner, J, Hâncean, MG. Micro-level network dynamics of scientific collaboration and impact: Relational hyperevent models for the analysis of coauthor networks. Netw. Sci. 2023;11:5–35.

Meskó, B, Topol, EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit. Med. 2023;6,120.

European Parliament. EU AI Act: first regulation on artificial intelligence. 2023. Disponibil la: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (accesat pe 14 Mai 2024).

Zhu, L, Mou, W, Chen, R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? J. Transl. Med. 2023;21,269.

Hershenhouse, JS, Mokhtar, D, Eppler, MB, Rodler, S, Ramacciotti, LR, Ganjavi, C, Hom, B, Davis, RJ, Tran, J, Russo, GI, Cocci, A, Abreu, A, Gill, I, Desai, M, Cacciamani, GE, Accuracy, readability, and understandability of large language models for prostate cancer information to the public. Prostate Cancer Prostatic Dis. 2024.

Coskun, B, Ocakoglu, G, Yetemen, M, Kaygisiz, O. Can ChatGPT, an artificial intelligence language model, provide accurate and high-quality patient information on prostate cancer? Urology. 2023;180:35 – 58.

Lombardo, R, Gallo, G, Stira, J, Turchi, B, Santoro, G, Riolo, S, Romagnoli, M, Cicione, A, Tema, G, Pastore, A, Al Salhi, Y, Fuschi, A, Franco, G, Nacchia, A, Tubaro, A, De Nunzio, C. Quality of information and appropriateness of Open AI outputs for prostate cancer. Prostate Cancer Prostatic Dis. 2024.

Longwell, JB, Hirsch, I, Binder, F, Conchas, GAG, Mau, D, Jang, R, Krishnan, RG, Grant, RC. Performance of large language models on medical oncology examination questions. JAMA Netw Open. 2024;7, e2417641.

Kung, TH, Cheat, M, Medenilla, A, Sillos, C, De Leon, L, Elepano, C, Madriaga, M, Aggabao, R, Diaz-Candido, G, Maningo, J, Tseng, V. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023; 2, e0000198.

Bol, NT, Dienlin, T, Kruikemeier, S, Sax, M, Boerman, S.C, Strycharz, J, Helberger, N, de Vreese, CH. Understanding the effects of personalization as a privacy calculus: analyzing self-disclosure across health, news, and commerce contexts. Journal of Computer-Mediated Communication. 2018;23:370–388.

Geantă, M, Boată, A, Brand, A, Cioroboiu, C, Cucoș, B. Personalized medicine literacy. In Precision Medicine in Clinical Practice, 1st ed. Hasanzat, M, Eds. Springer, Singapore, 2022:197-219.

Kianian, R, Sun, D, Crowell, EL, Tsui, E. The use of large language models to generate education materials about uveitis. Ophthalmology Retina. 2024; 8:195-201.

Chen, S, Guevara, M, Moningi, S, Hoebers, F, Elhalawani, H, Kann, BH, Chipidza, FE, Leeman, J, Aerts, HJWL, Miller, T, Savova, GK, Gallifant, J, Celi, LA, Mak, RH, Lustberg, M, Afshar, M, Bitterman, D. The effect of using a large language model to respond to patient messages. The Lancet Digital Health. 2023; 5, e333-e335.

Leyh-Bannurah, SR, Tian, Z, Karakiewicz, PI, Wolffgang, U, Sauter, G, Fisch, M, Pehrke, D, Huland, H, Graefen, M, Budäus, L. Deep learning for natural language processing in urology: state-of-the-art automated extraction of detailed pathologic prostate cancer data from narratively written electronic health records. JCO Clin Cancer Inform. 2018;2:1 – 9.

Koranteng, E, Rao, A, Flores, E, Lev, M, Landman, A, Dreyer, K, Succi, M. Empathy and Equity: Key Considerations for Large Language Model Adoption in Health Care. JMIR Med Educ. 2023; 9, e51199.

Stoltzfus, M, Kaur, A, Chawla, A, Gupta, V, Anamika, FNU, Jain, R. The role of telemedicine in healthcare: an overview and update. Egypt J Intern Med. 2023; 35,49.

Tong, W, Guan, Y, Chen, J, Huang, X, Zhong, Y, Zhang, C, Zhang, H. Artificial intelligence in global health equity: an evaluation and discussion on the application of ChatGPT, in the Chinese National Medical Licensing Examination. Front Med. 2023;10, 1237432.

Hâncean, MG, Lerner, J, Perc, M, Ghiţă, MC, Bunaciu, DA Stoica, AA, Mihăilă, BE. The role of age in the spreading of COVID-19 across a social network in Bucharest, Journal of Complex Networks. 2021; 9, cnab026.

Phatak, A, Mago, VK, Agrawal, A, Inbasekaran, A, Giabbanelli, PJ. Narrating causal graphs with large language models. 2024. Disponibil online: https://doi.org/10.48550/arXiv.2403.07118 (accesat pe 4 Iulie 2024).

LIST OF PUBLISHED SCIENTIFIC WORKS

Geantă, M, Bădescu, D, Chirca, N, Nechita, OC, Radu, CG, Rascu, S, Rădăvoi, D, Sima, C, Toma, C, Jinga, V. The Emerging Role of Large Language Models in Improving Prostate Cancer Literacy. Bioengineering. 2024; 11, 654. https://doi.org/10.3390/bioengineering11070654 (Capitolele 1, 2, 3, 4, 5, 7). Indexare ISI. Factor Impact: 3,8

Geantă, M, Bădescu, D, Chirca, N, Nechita, OC, Radu, CG, Rascu, S, Rădăvoi, D, Sima, C, Toma, C, Jinga, V. The Potential Impact of Large Language Models on Doctor–Patient Communication: A Case Study in Prostate Cancer. Healthcare. 2024; 12, 1548. https://doi.org/10.3390/healthcare12151548 (Capitolele 1, 2, 3, 4, 6, 7). Indexare ISI. Factor Impact: 2,4