

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32

Study Protocol for a Two-Phase Randomized Evaluation of Large Language Models in Adherence to Evidence-Based Health Communication Guidelines for Breast and Prostate Cancer Screening: The Role of User Prompt Specificity and Minimal Interventions (BOOST-AI)

Protocol version: 1.0
Last updated: 16.10.2024

Dr. Felix G. Rebitschek^{1, 2}, rebitschek@uni-potsdam.de, 1st author (corresponding)
M.Sc. Christoph Wilhelm^{1, 3}, christoph.wilhelm@uni-potsdam.de, senior author

¹Harding Center for Risk Literacy, Faculty of Health Sciences Brandenburg, University of Potsdam, Virchowstr. 2, 14482 Potsdam, Germany
²Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany
³Institute of Health and Nursing Sciences, Martin Luther University Halle-Wittenberg, Magdeburger Str. 8, 06112 Halle (Saale), Germany

Trial Registration: This clinical trial is registered in the German Clinical Trials Register (DRKS).
Registration Number: DRKS00035228
Registration Date: October 15, 2024
Permalink to Trial Registration: <https://drks.de/search/de/trial/DRKS00035228>

33 Introduction

34 In Western healthcare systems, there is a growing emphasis on ensuring that health information is
35 communicated in a way that allows individuals to make well-informed decisions about their medical
36 treatments. Central to this is the clear presentation of patient-relevant benefits and risks of medical
37 procedures.¹ To achieve this, evidence-based guidelines, such as the Guideline for the Development of
38 Evidence-based Patient Information² and the Working Group for Good Practice in Health Information
39 (GPHI)³, provide a structured framework for health communication. These guidelines stress the
40 importance of using precise numerical data to present risks and benefits, contextualizing information
41 within a relevant reference class, and offering balanced explanations of possible outcomes. Despite the
42 availability of these comprehensive frameworks, their practical application remains inconsistent. As the
43 internet increasingly becomes the primary source of health information for people in Europe and the
44 U.S.,^{4,5} many online resources fall short of adhering to these evidence-based standards.⁶ As a result,
45 much of the health information available online lacks the necessary accuracy and rigor, and in some
46 cases even spreads misinformation, particularly in areas such as cancer.^{7,8} This leaves individuals
47 vulnerable to receiving misleading information, which can lead to poorly informed health decisions and
48 a compromised ability to evaluate risks and benefits effectively. The rapid emergence of artificial
49 intelligence (AI) tools, particularly large language models (LLMs) such as OpenAI's ChatGPT, Google's
50 Gemini, and Mistral AI's Le Chat, has opened new possibilities for digital health communication and
51 laypeople are increasingly turning to these platforms to seek answers to health-related questions.⁹
52 However, a critical concern remains: Can these AI-driven systems consistently provide information that
53 adheres to established guidelines for communicating patient-relevant benefits and harms, especially
54 when prompted by laypeople?

55 While previous studies have compared LLM responses to those of physicians, exploring aspects such as
56 empathy and accessibility (Ayers et al., 2023; Haver et al., 2024), others have focused on LLM accuracy
57 in specific areas such as lab test interpretation (He et al., 2024), cancer screening recommendations
58 (Huo et al., 2024), and random cancer-related queries (Pan et al., 2023), just to name a few, revealing
59 both the strengths and limitations of these models. However, in all of these studies, the interactions
60 were based on artificial prompts generated by researchers, rather than authentic inquiries from
61 laypeople. This controlled setup allowed for a systematic evaluation but did not capture the variability
62 and complexity of real-world questions relevant to patients in a chat situation.

63 In contrast, our research addresses a critical gap in the evaluation of LLMs by investigating how well
64 these AI tools adhere to evidence-based guidelines when communicating about breast and prostate
65 cancer screenings. The primary research question is: Can LLMs provide accurate, guideline-based
66 information on the risks, benefits, and outcomes of cancer screenings, particularly when prompted by
67 laypeople? This question is crucial, given the previous mentioned increasing reliance on AI-driven
68 platforms for health information

69 Ultimately, previous research has focused on the general capabilities of LLMs, but our study aims to
70 determine whether these models could support informed decision-making in high-stakes medical
71 contexts, focusing on improving both user interaction and model reliability—by providing clear,
72 accurate, and evidence-based health communication. LLMs like ChatGPT, Google Gemini, and Mistral
73 AI will be tested to determine if they can be reliable sources of education, particularly in scenarios
74 where individuals seek information about cancer screening. The ultimate goal is to assess whether
75 these AI systems can be effectively integrated into healthcare systems to complement traditional health
76 communication tools, enhancing patient understanding and supporting informed health decisions
77 without replacing healthcare professionals.

78 Finally, we need to evaluate the effectiveness of basic evidence-based strategies of information search.
79 Consequently, we study different prompting strategies on the quality of LLM responses. Specifically, we
80 will compare standard prompting (control group) with enhanced prompting (intervention group who
81 receives a brief premise for an evidence-based search) to assess how the specificity of user input affects
82 the adherence of LLM-generated responses to evidence-based health communication guidelines.

83 **Objective**

84 The primary objective is to evaluate how well LLMs, including OpenAI’s ChatGPT, Google Gemini, and
85 Mistral AI, adhere to evidence-based guidelines when responding to breast and prostate cancer
86 screening prompts. Specifically, we aim to:

87 1. Assess the evidence-based quality of LLM responses by evaluating the extent to which the
88 health risk communication provided by LLMs aligns with established evidence-based
89 guidelines. This includes assessing whether LLM-generated responses clearly communicate
90 risks, benefits, and outcomes of breast and prostate cancer screening.

91 RQ1: Is the health risk communication provided by LLMs evidence-based? The hypothesis is
92 that LLMs frequently fail to meet standard criteria for evidence-based health risk
93 communication, with more than 50% of responses deviating from these guidelines across
94 multiple presentation criteria.

95 2. Analyze the effect of prompt specificity on LLM response quality by determining whether more
96 specific, well-informed prompts lead to a higher quality of evidence-based responses from
97 LLMs. This objective focuses on how the inclusion of key decision-making information in
98 prompts affects the clarity and accuracy of LLM outputs.

99 RQ2: Does more informed prompting, particularly in terms of health decision-making
100 preparation, result in better evidence-based health risk communication from LLMs? We
101 hypothesize that there is a moderate to strong positive correlation between the specificity of
102 the prompts and the quality of evidence-based health risk communication provided by LLMs.

103 3. Compare the evidence-based quality of responses generated by laypeople with varying levels
104 of informed prompting by comparing the quality of LLM responses based on prompts
105 generated by laypeople who provide varying levels of detail and information. We will assess
106 whether layperson-generated prompts result in more evidence-based responses than low-
107 informed prompts, but less so than moderately informed prompts.

108 RQ3: Are LLM responses generated by laypeople more evidence-based than those from low-
109 informed prompting, but less so than moderately informed prompting? We hypothesize that
110 Layperson-generated prompts produce more evidence-based responses compared to low-
111 informed prompts, but less so than moderately informed prompts that include about 50% of
112 key health decision-making information.

113 4. Evaluate the impact of a minimal boosting intervention on the quality of LLM responses by
114 assessing whether a minimal boosting intervention—where users are encouraged to provide
115 more specific prompts by considering the consequences of their health decisions—can
116 improve the evidence-based quality of LLM responses.

117 RQ4: Does a minimal boosting intervention increase evidence-based responses? We
118 hypothesize that reminding users to consider the consequences of their choices will lead to an
119 increase in evidence-based responses, improving the overall adherence of LLM outputs to
120 guideline-based communication standards.

121 5. Test the digital native hypothesis: RQ5: Without boosting intervention, digital natives (defined
122 as participants under 30 years of age) do not prompt a higher quality of LLM responses than
123 people from the age of 30 years and more.

124 **Trial Design**

125 The study is divided into two phases. Phase 1 uses a content analysis design to evaluate LLM responses
126 to standardized prompts. Phase 2 is a randomized between-subjects experiment with a 1:1 allocation
127 ratio, comparing standard prompting (control group) and enhanced prompting (intervention group).
128 Each group will interact with LLMs under similar conditions, but with differing levels of instruction
129 specificity.

130 Given that this study is entirely web-based, there are no onsite requirements for integrating the AI
131 intervention into the trial setting. All interactions with the LLMs will occur remotely, facilitated through
132 the SoSci Survey platform, hosted by the University of Potsdam. The integration of the AI systems (Open
133 AI ChatGPT, Google Gemini, and Mistral AI Le Chat) is managed through an API, ensuring seamless
134 communication between the survey platform and the LLMs.

135 For offsite requirements, participants will need access to a digital device (computer, tablet) with a
136 stable internet connection to interact with the LLMs and complete the survey.

137 There is no need for any physical infrastructure or on-premises installations, as the entire intervention
138 process, from user prompts to LLM responses, will be executed in the cloud through secure, encrypted
139 connections.

140 **Methods**

141 This protocol was developed in adherence to the Guidelines for clinical trial protocols for interventions
142 involving artificial intelligence: the SPIRIT-AI extension.¹⁰

143 **Phase 1: Systematic Evaluation of LLM Responses to Predefined Prompts**

144 In the first phase of the study, we will systematically assess the quality of outputs generated by
145 OpenAI's ChatGPT, Google Gemini, and Mistral AI. The LLMs will respond to a set of predefined,
146 standardized English-language prompts related to BC and PC screening. These prompts will encompass
147 a range of typical patient inquiries, such as the risks and benefits of cancer screenings, the
148 interpretation of screening results, and the overall recommendation for undergoing screening. Each
149 LLM will respond to multiple iterations of the same prompts, allowing for an assessment of both
150 consistency and variability in the responses.

151 In total, six (decision preparation elements) times three distinct (in specificity varied) prompts will be
152 repeated 20 times across three LLMs and two screening topics, resulting in 2,160 total responses to be
153 scored. An additional scoring across six prompts (full decision preparation) will be based on nine
154 evaluation criteria of the compliance with context information standards (e.g. declaration of conflict of
155 interest). Each LLM response will be evaluated against predefined quality metrics, by independent
156 human raters using the validated MappInfo tool for the evidence-based quality assessment of digital
157 health information¹¹ and a new checklist derived from the Guideline for the Development of Evidence-
158 based Patient Information². This process will help identify patterns of strengths and weaknesses in how
159 LLMs communicate crucial health information.

160 **Phase 2: User-Generated Prompts and Evaluation with a Minimal Boosting Intervention**

161 The second phase of the study shifts from predefined prompts to user-generated prompts to explore
162 how laypeople's input affects the quality of LLM-generated health information. In this phase,
163 participants will be asked to generate their own prompts regarding breast and prostate cancer
164 screening. Participants will be randomized into two groups:

165 Control group: Participants will generate prompts with no specific guidance, reflecting a typical
166 layperson's inquiry based on their own knowledge or concerns.

167 Intervention group: Participants will receive the minimal intervention (boosting) - a brief premise for
168 an evidence-based search: "Please consider the OARS rule: You need to know your options, the
169 advantages and risks of each, and how steady they are to happen.", which implies simple instructions
170 encouraging them to consider the possible consequences of their health decisions and to provide more
171 specific, detailed prompts.

172 We will follow a similar approach to Phase 1, using predefined quality metrics to assess the responses.
173 Furthermore, this phase will analyze whether the boosting intervention improves the overall quality of
174 LLM responses by increasing the specificity and relevance of the user-generated prompts. Specifically,
175 the evaluation will focus on two key aspects: (1) the proportion of evidence-based responses,
176 measuring how often LLMs provide information aligned with established guidelines; (2) the impact of
177 the boosting intervention, determining whether simple guidance leads to significant improvements in
178 LLM outputs compared to unguided prompts.

179 **Analysis**

180 Phase 1:

181 Descriptive statistics will be used to summarize the quality scores of the LLM responses across the
182 different levels of prompt specificity. The mean scores with standard deviations will be calculated for
183 each LLM and prompt category. Additionally, the percentage of responses meeting the predefined
184 evidence-based criteria will be computed. We test the assumption of prompt specificity and explore
185 potential differences between LLMs with the help of ANOVA across repetitions. Two independent
186 researchers will code the data. Interrater reliability will be assessed using Cohen's kappa coefficient to
187 evaluate agreement between the coders. Any discrepancies will be resolved through consensus
188 discussions to ensure consistent and accurate evaluation.

189 Phase 2:

190 Descriptive statistics will be calculated for the demographic variables gender, age, and education, as
191 well as for LLM usage frequency, prior experience, and preference for shared decision-making.

192 The primary analysis will focus on comparing the quality of LLM-generated responses across groups
193 using inference statistical methods for assessing differences between participants assigned to the
194 prompting conditions (independent variable). To evaluate the responses (the dependent variable) the
195 responses will be coded using two predefined scales to assess its adherence to evidence-based (EB)
196 health communication standards and pooled additionally (so, three dependent variables for
197 independent analyses). The coding will be based on the MappInfo tool for the evidence-based quality
198 assessment of digital health information¹¹ and the Guideline for the Development of Evidence-based
199 Patient Information². The coding will be done by two independent raters, and inter-rater reliability will
200 be assessed using Cohen's kappa. Any discrepancies between raters will be resolved through discussion
201 and consensus. The code book will be provided in the appendix of the final study.

202 Secondary analyses will include a statistical test based on age split (<30 vs 30+ years) on the difference
203 in the quality of LLM responses prompted by participants of the control condition and there will be an
204 assessment of participant preferences for informed decision-making processes and their experiences
205 with LLMs, using descriptive statistics and mean comparisons. A subgroup analysis will explore
206 potential differences in responses based on demographic characteristics.

207 All statistical tests will be conducted at a significance level of $p < .050$ (and adjusted downwards
208 according to Bonferroni procedure for multiple testing), and effect sizes will be calculated to determine
209 the meaning of findings. Results will be presented in detail with supporting tables and figures.

210 **Inclusion Criteria**

211 Phase 1, LLMs:

212 The inclusion criteria for selecting the LLMs in the study will be the following:

- 213 - State-of-the-Art Performance: Only LLMs that represent current, state-of-the-art models in
214 natural language processing, such as OpenAI's ChatGPT, Google Gemini, and Mistral AI, will be
215 considered.
- 216 - Accessibility for Public Use: The LLMs must be accessible to the general public, ensuring that
217 their capabilities reflect real-world use cases and that the findings can be generalized to typical
218 interactions by laypeople.
- 219 - Multidomain Knowledge: The LLMs must demonstrate the ability to handle a wide range of
220 topics, including healthcare and cancer-related information, ensuring their relevance for
221 answering complex, domain-specific queries.

222 Phase 2, Participants:

- 223 - Age: Participants must be 18 years or older to ensure they have the capacity to make
224 informed decisions and engage meaningfully with health-related prompts.
- 225 - Language Proficiency: Participants must have proficiency in English, as the study involves
226 interacting with LLMs in English and understanding health-related information presented in
227 this language.
- 228 - Access to Digital Devices: Participants must have access to and be able to use digital devices
229 (e.g., computers, tablets) with internet access, as the study involves generating and
230 submitting prompts to LLMs online.
- 231 - Geographic Location: Participants should reside in regions where access to healthcare is
232 comparable to international standards, such as the U.K., to ensure that the health
233 information provided by LLMs is relevant and applicable to their context.

234 **Exclusion Criteria**

235 Phase 1, LLMs:

236 The exclusion criteria for selecting the LLMs in the study will be the following:

- 237 - Limited Access or Restricted Use: LLMs that are not publicly accessible or require proprietary
238 access for specialized use will be excluded, as they do not represent general-use cases for
239 laypeople.
- 240 - Domain-Specific Models: LLMs that are specifically trained or fine-tuned for niche domains
241 (e.g., exclusively healthcare-specific models) will be excluded, as they do not reflect the
242 broader, general-purpose models used by the public.
- 243 - Non-English Language Proficiency: LLMs that primarily operate in languages other than English
244 or demonstrate limited proficiency in understanding and generating English-language
245 responses will be excluded.

246 Phase 2, Participants:

248 The exclusion criteria for selecting the Participants in the study will be the following:

- 249 - Health Professionals: Individuals with professional expertise in healthcare, particularly in
250 cancer screening or health communication, will be excluded to avoid bias and ensure that
251 participants reflect the general lay population.
- 252 - Previous Experience with LLM Studies: Participants who have taken part in similar studies
253 involving LLMs will be excluded to prevent familiarity with the technology from influencing the
254 results.

255 **Procedure and material**

256 Phase 1:

257 Researchers will prompt LLMs via API and collect the responses.

258 Phase 2:

259 Participants will receive a brief introduction outlining the study's purpose, which involves interacting
260 with one of three LLMs [OpenAI ChatGPT (gpt-3.5-turbo)¹², Google Gemini (1.5-Flash)¹³, or Mistral AI
261 Le Chat (mistral-large-2402)]¹⁴, all preset as a "helpful assistant" to obtain health information of either
262 BC or PC screening. Following informed consent, participants will be asked to provide basic
263 demographic details such as gender, age, and education level through a questionnaire interface built
264 with SoSci Survey. They will then be given the choice to receive information about either BC or PC
265 screening. Participants will engage with the chatbot by entering their queries (prompts) through the
266 LLM's API, and the generated responses will be collected alongside the prompts for systematic analysis.
267 At the end of the session, participants will be asked about their frequency of LLM usage, their prior
268 experience with such models, and their perspectives on how an informed decision-making process
269 should ideally be conducted for them, by choosing one from four options.

270 A pre-test with n=20 participants will be conducted to verify the effectiveness of the randomization
271 procedures, ensuring balanced group assignment, the technical functionality of the survey platform
272 and its integration with the LLM APIs. Additionally, the average completion time will be measured and
273 potential issues related to participant burden will be identified. Feedback on user experience will be
274 collected to address any usability concerns. Based on these findings, necessary adjustments will be
275 made at the protocol for the main trial.

276 **Tests and Outcomes**

277 Primary Outcome:

278 Expert scoring of how evidence-based LLM communication is (mappInfo tool in addition to a novel
279 score based on the guideline EB health information)

280 Secondary Outcome:

281 Self-reported use and experience with LLM; preference in shared decision-making

282 **Sample Size**

283 Phase 1:

284 There are no human participants involved.

285 Phase 2:

286 To analyze a moderate ANOVA main effect (partial eta squared = .06) comparing two between-subjects
287 conditions (with and without intervention), we are aiming for a minimum sample size of n= 237

288 participants. To ensure a representative sample reflecting the simplified census data of Great Britain in
289 terms of sex, age, and ethnicity, we will recruit $n=300$ participants from a diverse pool via Prolific¹⁵.

290 **Recruitment**

291 Participants will be recruited via the online platform Prolific¹⁵.

292 **Assignment of Interventions**

293 Participants will be assigned to either the control group or the intervention group via block
294 randomization. In the intervention group, participants will receive additional instructions designed to
295 improve the quality of the prompts they generate for the LLMs.

296 **Allocation**

297 To achieve a balanced distribution of participants across the study groups, we will employ a block
298 randomization (1:1 allocation ratio) with the BLOCKRAND() function in SoSci Survey. The assignment
299 is concealed by this function.

300 **Blinding**

301 Participants and researchers will be blinded to both the assigned LLM and the type of prompting
302 instructions to prevent any bias in interaction and response evaluation. Coding will happen
303 independently by two researchers.

304 **Data Management**

305 Data will be collected via the SoSci Survey platform, and all datasets will be anonymized and stored
306 securely on password-protected servers to ensure participant confidentiality. Contact details will not
307 be collected, as Prolific—responsible for participant recruitment—operates with an anonymized data
308 collection model. Participant demographic data, such as gender and age, will be collected separately
309 within the study, despite these being part of the quota set by Prolific.

310 Data handling is fully anonymized. Prolific manages recruitment, while SoSci Survey, hosted by the
311 University of Potsdam, handles the questionnaires. Communication with the LLMs is facilitated through
312 an API within SoSci Survey, and LLM providers do not have access to the study data. Payment processing
313 is managed entirely by prolific.co, and no financial data is collected or stored on our end. Additionally,
314 no names, addresses, birth dates, or IP addresses will be recorded. We will not record the recruitment
315 IDs generated by prolific.co, and due to the large number of simultaneous participants and the
316 variability in time between recruitment and survey completion, it will not be feasible to link
317 questionnaire responses with recruitment IDs.

318 The anonymized research data will initially be stored on a password-protected account under the
319 control of the project lead on the SoSci Survey server (hosted by the University of Potsdam). For further
320 analysis, the data will be transferred to password-protected computers under the supervision of project
321 team members at the Harding Center for Risk Literacy. Prolific will not have access to the research data
322 at any point.

323 **Statistical Methods**

324 Data will be analyzed using descriptive statistics and inferential statistical tests, including t-tests and
325 ANOVA for group comparisons, and linear regression models for more detailed analyses of prompt
326 specificity effects. All statistical tests will be conducted at a significance level of $p<0.05$, and effect sizes
327 will be calculated to determine the practical significance of findings.

328 **Harms**

329 No physical or psychological risks are anticipated for participants. Anticipated survey completion time
330 is six minutes. Participation is entirely voluntary, and participants can withdraw at any time.

331 Errors of the used LLMs will be identified through content analysis conducted by independent human
332 raters using predefined quality metrics, and categorized based on their nature and impact on the
333 reliability of the information provided. The errors will then be analyzed to determine patterns of LLM
334 performance issues. If errors are not analyzed in specific instances, it will be due to limitations in the
335 scope of the study, such as focusing primarily on guideline adherence rather than linguistic or technical
336 issues outside the study's objectives.

337 **Monitoring**

338 The study will be monitored by the research team at the Harding Center for Risk Literacy. Regular audits
339 will be conducted to ensure compliance with the study protocol and data management procedures.

340 **Ethics and Dissemination**

341 Ethics approval was obtained from the University of Potsdam's Ethical Committee (Approval No.
342 52/2024) on August 29, 2024. The complete study protocol is publicly available on the website of the
343 Harding Center (<https://www.hardingcenter.de/de/forschung/projekt-eb-llm>). Findings from the study
344 will be disseminated through academic publications, conference presentations, and open-access
345 databases to contribute to the field of digital health communication. A de-identified, aggregate-level
346 data will be made available upon reasonable request to qualified researchers who agree to comply with
347 ethical guidelines for data sharing and usage.

348 **Protocol Amendments**

349 All amendments to the protocol will be submitted to the ethical committee for review and approval.
350 Any major changes will be communicated to participants and will be listed, along with justifications, in
351 a new version of the study protocol and the final study publication.

352 **Consent**

353 Participants will provide informed consent electronically prior to participating in the study. The consent
354 form includes detailed information about the study, participant rights, and data protection measures.

355 **Confidentiality**

356 All participant data will be anonymized, and no personally identifiable information will be collected or
357 stored. Data will be stored securely, and access will be restricted to the research team.

358 **Funding and Conflicts of Interest**

359 This study will not receive any external funding. There are no financial conflicts of interest to declare.
360 Furthermore, we confirm that neither the research team nor any individual involved in this study will
361 receive any monetary or personal benefit from the engagement of Prolific for participant recruitment.
362 The commissioning of funds for various studies using Prolific is managed through the relevant
363 department for procurement of the University of Potsdam.

364 **Role of Study Sponsor and Funders**

365 As this study is not funded by any external sponsor, there is no external influence on the study design,
366 data collection, management, analysis, interpretation of data, or the writing of the final report. The
367 decision to submit the report for publication rests solely with the research team of the Harding Center.

368 The research will be conducted independently, without involvement from any third-party sponsor or
369 funder.

370 **Composition, Roles, and Responsibilities**

371 The study will be coordinated by the research team at the Harding Center, with oversight by the project
372 lead Dr. Felix G. Rebitschek. Given the nature of this trial and the lack of external funding, there is no
373 steering committee or endpoint adjudication committee involved. Data management will be handled
374 by the research team, ensuring compliance with data privacy regulations and ethical standards. There
375 is no need for a separate data monitoring committee, as the study presents minimal risk and involves
376 anonymized, non-invasive interactions with AI-based language models.

377 **Contribution**

378 Christoph Wilhelm (CW) and Dr. Felix G. Rebitschek (FGR) were responsible for the conceptualization
379 and development of the methodology. CW wrote this protocol. CW is the guarantor of the manuscript.
380 FGR contributed equally, provided key resources, and oversaw the project administration. He reviewed,
381 edited, and supervised this protocol. Both authors significantly contributed to the writing, reviewing,
382 and editing of this manuscript.

383

- 384 1. Klein WMP, Chou W-YS, Vanderpool RC. Health Information in 2023 (and Beyond):
385 Confronting Emergent Realities With Health Communication Science. *JAMA* 2023; **330**(12): 1131-
386 2.
- 387 2. Luhn J, Albrecht M, Hanssen K, Hildebrandt J, Steckelberg A. [Guideline for the
388 Development of Evidence-based Patient Information: insights into the methods and
389 implementation of evidence-based health information]. *Z Evid Fortbild Qual Gesundheitswes* 2015;
390 **109**(2): 159-65.
- 391 3. Arbeitsgruppe G. [Good practice guidelines for health information]. *Z Evid Fortbild Qual*
392 *Gesundhwes* 2016; **110-111**: 85-92.
- 393 4. Andreassen HK, Bujnowska-Fedak MM, Chronaki CE, et al. European citizens' use of E-
394 health services: A study of seven countries. *BMC Public Health* 2007; **7**(1): 53.
- 395 5. Calixte R, Rivera A, Oridota O, Beauchamp W, Camacho-Rivera M. Social and
396 Demographic Patterns of Health-Related Internet Use Among Adults in the United States: A
397 Secondary Data Analysis of the Health Information National Trends Survey. *Int J Environ Res Public*
398 *Health* 2020; **17**(18).
- 399 6. Riera R, de Oliveira Cruz Latorraca C, Padovez RCM, et al. Strategies for communicating
400 scientific evidence on healthcare to managers and the population: a scoping review. *Health*
401 *Research Policy and Systems* 2023; **21**(1): 71.
- 402 7. Loeb S, Langford AT, Bragg MA, Sherman R, Chan JM. Cancer misinformation on social
403 media. *CA Cancer J Clin* 2024; **74**(5): 453-64.
- 404 8. Johnson SB, Parsons M, Dorff T, et al. Cancer Misinformation and Harmful Information on
405 Facebook and Other Social Media: A Brief Report. *JNCI: Journal of the National Cancer Institute*
406 2022; **114**(7): 1036-9.
- 407 9. Park Y-J, Pillai A, Deng J, et al. Assessing the research landscape and clinical utility of large
408 language models: a scoping review. *BMC Medical Informatics and Decision Making* 2024; **24**(1):
409 72.
- 410 10. Cruz Rivera S, Liu X, Chan A-W, et al. Guidelines for clinical trial protocols for interventions
411 involving artificial intelligence: the SPIRIT-AI extension. *The Lancet Digital Health* 2020; **2**(10):
412 e549-e60.
- 413 11. Kasper J, Luhn J, Hinneburg J, et al. MAPPinfo - mapping quality of health information:
414 Validation study of an assessment instrument. *PLoS One* 2023; **18**(10): e0290027.
- 415 12. OpenAI. GPT-4 Overview2023. <https://openai.com/index/gpt-4/> (accessed 23.09.2024).
- 416 13. Google. Google Gemini Overview2023. <https://gemini.google.com/> (accessed
417 24.09.2024).
- 418 14. AI M. Mistral AI Overview2023. <https://mistral.ai/> (accessed 24.09.2024).
- 419 15. Prolific. Prolific: Online research participant platform2024. <https://www.prolific.com/>
420 (accessed 25.09.2024).

421