

1 **Effects of temperature settings on information quality of ChatGPT-3.5 responses: A**

2 **prospective, single-blind, observational cohort study**

3

4 **Short title:** Temperature settings and ChatGPT-3.5 response information quality

5

6 Akihiko Akamine^{1*}, Daisuke Hayashi², Atsushi Tomizawa³, Yuya Nagasaki³, Chikae

7 Akamine⁴, Takahiro Fukawa⁵, Iori Hirosawa⁶, Ori Saigo⁷, Misa Hayashi⁸, Mitsuru

8 Nanaoya³, Yuka Odate⁹

9

10 ¹Department of Pharmacy, Doujin Hospital, 1-37-12 Gusukuma, Urasoe, Okinawa, 901-

11 2133, Japan

12 ²Department of Pharmaceutical Service, Nippon Medical School Hospital, 1-1-5

13 Sendagi, Bunkyo-ku, Tokyo, 113-8603, Japan

14 ³Department of Pharmacy, Kitasato University Hospital, 1-15-1 Kitasato, Minami-ku,

15 Sagamihara, Kanagawa, 252-0375, Japan

16 ⁴Amity Co., Ltd. Viola Pharmacy, 1-3-32, Uenoya, Naha, Okinawa, 900-0011, Japan

17 ⁵Research and Education Center of Clinical Pharmacy, Kitasato University School of
18 Pharmacy Kanagawa, 1-15-1 Kitasato, Minami-ku, Sagamihara, Kanagawa, 252-0375,
19 Japan

20 ⁶Laboratory of Pharmacy Practice, Showa Pharmaceutical University, 3-3165,
21 Higashitamagawa-Gakuen, Machida, Tokyo, 194-8543, Japan

22 ⁷Department of Pharmacy, Juntendo University Hospital, 3-1-3, Hongo, Bunkyo-ku,
23 Tokyo, 113-0033, Japan

24 ⁸Department of Pharmacy, Japan Community Health Care Organization Tokyo Kamata
25 Medical Center, 2-19-2, Minami Kamata, Ota-ku, Tokyo, 144-0035, Japan

26 ⁹Department of Pharmacy, Nissan Tamagawa Hospital, 4-8-1, Seta, Setagaya-ku,
27 Tokyo, 158-0095, Japan

28

29 **ORCID number**

30 Akihiko Akamine: 0000-0002-2766-4611; Atsushi Tomizawa: 0000-0002-6395-339X;

31 Takahiro Fukawa: 0000-0001-7942-345X

32

33 *Corresponding Author

34 E-mail: a-aka@kitasato-u.ac.jp (AA)

35 **Abstract**

36 **Objective:** The effect of temperature settings on the quality of ChatGPT version 3.5

37 (OpenAI) responses related to drug information remains unclear. We investigated

38 ChatGPT-3.5's response quality on apixaban information with and without the

39 temperature being set to 0.

40 **Methods:** On 6 September 2023, 37 questions regarding apixaban, derived from the

41 frequently asked questions on the Bristol–Myers Squibb's website, were entered into

42 ChatGPT in Japanese. The primary endpoint was the effect of temperature settings on

43 ChatGPT-3.5's responses to apixaban-related questions. The response accuracy, clarity,

44 detail, and adequacy were rated on a 5-point Likert scale by 10 pharmacists, with higher

45 scores indicating higher response quality. Cumulative score means were analyzed using

46 the Mann–Whitney U test. In the subgroup analysis, evaluators were limited to

47 pharmacists at university hospitals. Welch's t-test was employed in sensitivity analysis

48 to validate primary endpoint findings.

49 **Results:** The mean scores for ChatGPT-3.5's apixaban-related responses with (13.08)

50 and without (14.40) the temperature being set to 0 were not significantly different ($p =$

51 0.064). Accuracy differed significantly (3.15 vs. 3.54, $p = 0.045$), whereas clarity,

52 detail, and appropriateness were similar. Subgroup analysis (13.30 vs. 14.21, $p = 0.394$)

53 and sensitivity analysis confirmed similar results (13.45 vs. 14.52, $p = 0.105$).

54 **Conclusions:** ChatGPT-3.5 temperature setting does not significantly affect overall

55 responses to apixaban-related inquiries. However, the variance in accuracy suggests that

56 ChatGPT-3.5 is unable to consistently provide precise responses. Hence, it is more

57 suitable as a supplementary tool rather than a primary medical resource.

58

59 **Keywords:** Chatbot, ChatGPT, Drug information services, Large language models

60 **Introduction**

61 Recent advances in artificial intelligence (AI) have led to the development of
62 sophisticated tools, such as ChatGPT [1], which are increasingly utilized in various fields,
63 including pharmaceutical information services. ChatGPT, developed by OpenAI, has the
64 potential to enhance patient care in the medical field by providing accurate information.
65 Its efficacy in predicting drug–drug interactions highlights its important role in healthcare
66 [2]. Furthermore, AI integration into medical safety education, including drug
67 information services, is being actively explored. These investigations focus on addressing
68 ethical and security issues associated with AI integration, ultimately aiming to provide
69 comprehensive and personalized medical services [3]. In life sciences, AI has facilitated
70 advances in research methods, protocols, and data analysis, enabling medical providers
71 to make more effective decisions [4].

72 Despite these advances, the application of ChatGPT in drug information services
73 remains challenging. In a previous real-world study, ChatGPT answered the majority of
74 drug-related questions incorrectly or only partly correctly, highlighting the limitations of
75 applying AI in the drug information field due to issues such as inaccurate content and a
76 lack of references [5–9]. This performance inconsistency raises concerns regarding the
77 reliability and robustness of AI-generated drug information. It is necessary to assess the

78 accuracy and reliability of ChatGPT in providing pharmaceutical information,
79 particularly under different operational settings, such as temperature, which can influence
80 the model's response style and content. A method for adjusting ChatGPT's response
81 quality by setting the temperature has been reported. ChatGPT temperature is a parameter
82 that controls the diversity of the generated text and can be specified as a value between 0
83 and 1. At low temperatures, the generated text is more predictable and monotonous,
84 whereas at high temperatures, the generated text may contain more diverse and random
85 words and expressions [10].

86 The following limitations have been associated with previous studies: (i) a small
87 number of ChatGPT response raters (<10) may have biased the ratings; (ii) drug
88 information validation was lacking; (iii) responses with different temperature settings
89 were not validated; and (iv) despite the evaluation of the response accuracy of ChatGPT,
90 the clarity, detail, and appropriateness of the responses were not verified. Addressing
91 these limitations would facilitate a more comprehensive evaluation of the quality of
92 ChatGPT responses to pharmaceutical information.

93 This study compared and verified the quality of the answers provided by
94 ChatGPT-3.5 regarding drug-related questions with and without a temperature setting of
95 0. The aim of this study was to evaluate the effect of temperature settings on the accuracy,

96 clarity, detail, and appropriateness of ChatGPT-3.5 responses and to elucidate its
97 reliability as a drug information tool. The drug of interest was oral apixaban (tablet;
98 Eliquis®; Bristol–Myers Squibb, New York, NY, USA), the highest-selling oral drug in
99 FY2022, excluding COVID-19 prophylaxis treatments and therapeutic modalities [11].

100

101 **Methods**

102 **Ethics approval**

103 This was an observational study. The Institutional Review Board for Observation and
104 Epidemiological Study at the Doujin Hospital confirmed that no ethics approval was
105 required (Date: 20 May 2023).

106

107 **Consent to participate**

108 Informed consent was not required for this study because it did not involve human
109 subjects.

110

111 **Study design**

112 This prospective, single-blind, observational cohort study was conducted at eight
113 hospitals, pharmacies, and pharmacy schools in Japan (Level of Evidence IV). Ten

114 evaluators (D.H., A.T., Y.N., C.A., F.T., I.H., M.H., M.N., O.S., and Y.O.) participated
115 in this study; among them, three pharmacists specialized in cancer, heart failure, and
116 perioperative patient management. Data collection, creation of questions for ChatGPT,
117 and analysis of ChatGPT responses were conducted between 1st July and 30th November
118 2023.

119

120 **Eligibility criteria of the pharmacists**

121 We included pharmacists with at least 3 years of experience in hospital or pharmacy
122 practice, who were employed in a facility utilizing apixaban. We excluded pharmacists
123 who, according to the principal researcher, were deemed incapable of adequately
124 evaluating drug information as well as pharmacists who used ChatGPT to obtain content
125 relevant to the study during the evaluation period.

126

127 **Creating questions for ChatGPT**

128 The principal researcher developed a total of 37 questions dissecting 35 frequently asked
129 questions regarding apixaban posted in Japanese on the Bristol–Myers Squibb’s website,
130 separating combined questions into individual queries, if necessary [12]. Three core
131 researchers (D.H., A.T., and Y.N.) evaluated and approved the questions. In cases of

132 disagreement, inputs were obtained from a fourth researcher (C.A.), and the principal
133 investigator made the final decision.

134

135 **Evaluation of ChatGPT’s responses**

136 On 6 September 2023, the primary researcher posted questions to ChatGPT-3.5 in
137 Japanese. To maintain novelty, each question was asked from the same account using a
138 “New Chat.” To maintain the temperature setting of ChatGPT at 0, we noted “Please
139 use temperature of 0” at the end of each question. The reproducibility of responses was
140 not considered in this study, and the first response obtained was evaluated. All questions
141 required written responses, and no multiple-choice questions were included. All the
142 textual prompts are provided in S1 File. The principal researcher used a Google form
143 with one question that did not specify the temperature setting and two options (set and
144 not set). These responses were randomized, blinded, and presented to the evaluators.
145 The evaluators were provided with answers with and without the temperature setting, in
146 a random order. They were instructed to read the questions and their answers prior to
147 the evaluation. Evaluators were asked to rate 296 responses using a 5-point Likert scale
148 using four dimensions, (accuracy, clarity, detail, and appropriateness) [13]. Each
149 question was answered only once, and the response options were rated on a scale of 1–

150 5, with higher numbers indicating higher quality. The evaluation criteria are presented
151 in S2 File. Evaluators consulted reliable sources of apixaban drug information (e.g.,
152 package inserts, interview forms, and Bristol–Myers Squibb’s website), as needed.

153

154 **Main analysis**

155 The primary endpoint of the study was the quality of ChatGPT-3.5’s answers to apixaban-
156 related questions with and without the temperature being set to 0. The accuracy, clarity,
157 detail, and appropriateness of the responses were rated on a 5-point Likert scale (1–5),
158 and the mean scores for all questions (4–20) were analyzed using Mann–Whitney U test.

159 In addition, as secondary endpoints, the accuracy, clarity, detail, and appropriateness of
160 ChatGPT-3.5’s answers were individually rated on a 5-point Likert scale (1–5) with and
161 without a temperature setting of 0. The mean score for each category for all questions was
162 analyzed using Mann–Whitney U test.

163

164 **Subgroup analysis**

165 To confirm the robustness of the primary outcomes, we conducted a subgroup analysis
166 that included only pharmacists affiliated with university hospitals, to evaluate the
167 influence of affiliations of pharmacists on outcomes similar to the main analysis.

168 University hospitals have reported higher patient satisfaction than general hospitals [14],
169 but no difference in mortality or readmission rates by disease has been noted. However,
170 since no studies have compared the quality of pharmacists at different institutions, only
171 pharmacists from university hospitals were included in this analysis.

172

173 **Sensitivity analysis**

174 To determine the effect of the statistical analysis method on the primary outcome, a
175 sensitivity analysis was performed by changing the statistical analysis method to Welch's
176 t-test.

177

178 **Statistical analysis**

179 The Shapiro–Wilk test was used to test the normality of the distribution of the age and
180 career data of the participating pharmacists as continuous variables. Continuous variables
181 were expressed as medians and means, whereas categorical data were expressed as
182 absolute values and percentages. Welch's t-test was used to analyze the means of
183 continuous variables, and Mann–Whitney U test was used to analyze the medians [15,16].
184 Pharmacists with missing study data were excluded from the univariate analyses.
185 However, when $\geq 20\%$ of the data were missing, multiple imputations were planned using

186 chained equations to create 100 sets of corresponding data. All statistical analyses were
187 performed using the EZR version 1.36 software (Saitama Medical Center, Jichi Medical
188 University, Saitama, Japan) [17]. All the tests were two-tailed. Statistical significance was
189 set at $p < 0.050$. As this was an exploratory study, the sample size was not calculated.
190 Nominal p -values were used to account for the multiplicity of analyses.

191

192 **Results**

193 **Characteristics of pharmacists**

194 Ten pharmacists evaluated all the responses, and none of the evaluators met the exclusion
195 criteria. The age ($p = 0.649$) and career length ($p = 0.551$) of the pharmacists showed
196 normal distributions. Pharmacist characteristics are listed in Table 1.

197

198 **Table 1. Pharmacist characteristics evaluated at baseline (n = 10).**

Characteristic	Total (n = 10)
Age	
Median (IQR), years	38.5 (32.8–42.8)

Mean (SD), years	38.2 (6.3)
Sex, No. (%)	
Male	5 (50.0)
Female	5 (50.0)
Career as a pharmacist	
Median (IQR), years	14.5 (9.0–20.3)
Mean (SD), years	14.6 (6.7)
Academic history, No. (%)	
Doctor	1 (10.0)
Master	4 (40.0)
Bachelor	5 (50.0)
Affiliation	
University hospital	6 (60.0)
General hospital	2 (20.0)
Pharmacy	1 (10.0)
Faculty	1 (10.0)

199 Abbreviations: IQR, interquartile range; SD, standard deviation

200

201 **Primary outcome**

202 All ChatGPT responses were in Japanese, eliminating the need to translate the responses.

203 With the temperature set to 0, the median of the answers was 13.08 (interquartile range:

204 12.50–14.03), whereas it was 14.40 (interquartile range: 13.84–15.32) without a

205 temperature setting, demonstrating no significant differences ($p = 0.064$). Answers with

206 the temperature set at 0 had a lower rate of total scores of ≥ 16 (maximum: 20) than those

207 without a temperature setting (7/37 [18.92%] vs. 17/37 [45.95%]; Fisher’s exact test, $p =$

208 0.024). The results of Mann–Whitney U test showed a significant difference between the

209 mean scores for answers with and without the temperature being set to 0 (accuracy: 3.15

210 [interquartile range: 3.06–3.30] and 3.54 [interquartile range: 3.29–3.65]; $p = 0.045$).

211 However, clarity, detail, and adequacy of answers were similar between groups (Table

212 2).

213

214 **Table 2. Comparisons of the scores of each endpoint for questions on apixaban drug**

215 **information with and without the ChatGPT temperature setting (n = 74).**

Score ^a	Temperature set to 0	No temperature setting	p value
	(n = 37)	(n = 37)	

Accuracy

Median	of	mean	3.15	3.54	0.045 ^b
(IQR)			(3.06–3.29)	(3.30–3.65)	
Clarity					
Median	of	mean	3.46	3.56	0.384
(IQR)			(3.00–3.80)	(3.40–4.13)	
Detail					
Median	of	mean	2.97	3.30	0.054
(IQR)			(2.85–3.25)	(3.16–3.59)	
Appropriateness					
Median	of	mean	3.67	3.91	0.121
(IQR)			(3.55–4.00)	(3.85–4.08)	

216 Notes: The p value was calculated using Mann–Whitney U test.

217 ^a Scores for accuracy, clarity, detail, and appropriateness (1–5 points each)

218 ^b Significant difference ($p < 0.050$)

219 Abbreviations: IQR, interquartile range

220

221 **Results of the subgroup analysis**

222 The university hospital pharmacist subgroup analysis yielded median values of 13.30

223 (interquartile range; 12.00–15.02) and 14.21 (interquartile range: 13.45–15.32) for
224 answers with and without the temperature being set at 0, respectively, demonstrating no
225 significant differences ($p = 0.394$). Hence, the subgroup analysis yielded results similar
226 to those of the primary analysis.

227

228 **Results of the sensitivity analysis**

229 When the primary analysis method was revised to Welch's t-test, the mean scores of the
230 answers with and without the temperature being set to 0 were 13.45 (standard deviation;
231 1.51) and 14.52 (standard deviation; 1.28), respectively, which were not significantly
232 different ($p = 0.105$). Thus, the sensitivity analysis yielded the same results as those of
233 the primary analysis.

234

235 **Discussion**

236 **Summary of key findings**

237 This study yielded two important findings. First, the overall quality of ChatGPT-3.5's
238 responses in terms of accuracy, clarity, detail, and adequacy was consistent, regardless of
239 the temperature setting, as evidenced by similar results across the primary endpoints,
240 subgroup analyses, and sensitivity analyses. Second, responses with a temperature setting

241 of 0 were less likely to have a total score ≥ 16 than those with no temperature setting
242 (18.92% vs. 45.95%, Fisher's exact test, $p = 0.024$). These findings provide a basis for
243 further discussion on the implications of temperature settings on AI-generated drug
244 information.

245 Although a previous study has shown that AI-based chatbots, including this
246 version of ChatGPT, have robust search and information integration capabilities,
247 particularly in clinical pharmacy [18], our study found a lower percentage of high-quality
248 responses when the temperature was set to 0 than those without temperature settings. This
249 finding is particularly interesting because it suggests a subtle effect of temperature setting
250 on response quality, which has not been previously explored. Furthermore, it underscores
251 the life-threatening consequences of using medication based on incorrect information.
252 Thus, addressing and resolving this issue promptly is crucial. Additionally, users should
253 ask detailed questions because the quality of ChatGPT answers depends on the phrasing
254 of the questions.

255

256 **Strengths and weaknesses**

257 This study has several strengths. It contributes markedly to the field of clinical
258 pharmacy and AI-based tools as it provides unique insights into the impact of

259 temperature setting on the quality of pharmaceutical information provided by
260 ChatGPT-3.5. This specific focus on temperature settings and their influence on AI
261 response quality has not been extensively explored in previous research. Second, the
262 study employed a robust methodology, including a clear primary endpoint and
263 comprehensive statistical analyses.

264 In addition, the study included 10 pharmacists in the evaluation, whose
265 diversity provided a broader perspective, thereby reducing bias and enhancing the
266 representativeness of the ratings. During the evaluation, the evaluators were not
267 informed about whether the responses from ChatGPT-3.5 were temperature-adjusted,
268 thus reducing potential order bias. Subgroup and sensitivity analyses were performed
269 to ensure the consistency of results for the primary endpoints, thereby enhancing result
270 reliability. This enabled us to generalize the results across different settings and rater
271 profiles. Moreover, this study acknowledges ChatGPT-3.5's limitations, particularly its
272 lack of internet search capabilities and reliance on preexisting datasets. Our critical
273 evaluation highlights the importance of continuous updates and improvements in AI
274 tools to ensure their effective use in healthcare settings, particularly in domains where
275 current and accurate information, such as drug data, is crucial. Finally, this study
276 focused on a specific drug, apixaban, allowing for a detailed and focused analysis of

277 AI performance in providing medication information. Although this approach limits
278 the generalizability of the findings, it enables a more in-depth understanding of AI
279 capabilities and limitations in the context of a single, widely used medication.

280 The study limitations must also be acknowledged. First, evaluators were limited
281 to pharmacists, primarily those working at university hospitals. This specific professional
282 background may have influenced the perception and evaluation of AI-generated
283 responses. Second, we used the Japanese version of ChatGPT-3.5, and the results may
284 vary for other languages due to differences in language processing and available datasets
285 in the AI model.

286

287 **Interpretation**

288 Consistent with our findings, the limitations of AI chatbots in effectively handling
289 complex medical information have been highlighted in previous studies that have cited a
290 lack of medicine-specific datasets and challenges in advanced reasoning [19].
291 Temperature settings designed to control response randomness may inadvertently affect
292 the chatbots' ability to access and integrate complex medical information effectively.

293 Although no significant difference was detected in the overall quality of
294 responses from ChatGPT-3.5 across temperature settings, a lower percentage of high-

295 quality responses was observed when the temperature was set to 0, thereby warranting
296 further investigation. This emphasizes the importance of careful consideration of the AI
297 chatbot settings in clinical applications and settings, ensuring they are optimized to
298 provide accurate and relevant information.

299 The subgroup and sensitivity analyses conducted in our study provided
300 additional insights into the robustness of ChatGPT-3.5's responses to pharmaceutical
301 inquiries. In the subgroup analysis limited to university hospital pharmacists, our findings
302 remained consistent with the primary outcome. This consistency across different groups
303 of raters reinforces the quality of ChatGPT-3.5 responses in a professional academic
304 setting.

305 Furthermore, the sensitivity analysis performed using a different statistical test
306 also supported the primary findings, showing non-significant differences in response
307 quality. This methodological robustness enhances the credibility of our results, suggesting
308 that the observed variance in accuracy is not a statistical anomaly but a characteristic of
309 the AI model's performance. However, the slight variance in accuracy observed in the
310 primary analysis remains a matter of concern. Although this variance is not statistically
311 significant, it could have implications in clinical settings where precise drug information
312 is crucial. Previous studies have also indicated variability in AI responses in clinical

313 scenarios, suggesting the need for the cautious application and continuous monitoring of
314 AI tools in healthcare [2,20].

315 ChatGPT-3.5 does not have an internet search capability, constricting its ability
316 to provide responses integrating the latest information, which is crucial in the drug
317 information field. For example, although the package insert recommends the
318 administration of Ondexxya in the event of life-threatening or difficult-to-staunch
319 bleeding when consuming apixaban, no responses related to the administration of
320 andexanet alfa (injection) (Ondexxya®; AstraZeneca, London, UK) were noted. This
321 could be attributed to the fact that Ondexxya was not available in Japan until May 2022,
322 and the ChatGPT data only extended until September 2021. This limitation is particularly
323 relevant in clinical pharmacy practice where accurate and up-to-date information is
324 paramount for patient safety. ChatGPT-3.5, however, contains limited learning data,
325 which is a crucial factor to be considered in the field of drug information. Our findings
326 underscore the importance of regularly updating and improving AI chatbots for their
327 effective utilization in clinical pharmacies and healthcare settings.

328

329 **Future research**

330 Although we found no significant difference in the overall quality of responses from

331 ChatGPT-3.5 across temperature settings, a lower percentage of high-quality responses
332 was observed when the temperature was set to 0, suggesting the need for further
333 investigation and careful consideration of AI chatbot settings in clinical applications.
334 Settings that optimize the information accuracy and relevance should be provided. Future
335 studies should focus on the inclusion of various drugs to obtain a deeper understanding
336 of the capabilities and limitations of ChatGPT-3.5 in relation to various drug classes and
337 their respective complexities.

338 Additionally, the participation of a diverse group of healthcare professionals is
339 essential for future evaluations to obtain a broader perspective on AI performance and its
340 utility across the healthcare ecosystem. Given the global applicability of AI tools, it is
341 critical to conduct similar studies in various linguistic and cultural contexts. This
342 approach will aid in understanding the impact of language processing and cultural
343 nuances on ChatGPT-3.5, aiming to evaluate the global validity and reliability of this
344 tool. Further investigation is also required to determine the reasons for the variation in AI
345 response quality, particularly under different temperature settings. Understanding the
346 mechanisms that lead to this variability will guide the development of more consistent
347 and reliable AI tools for clinical use.

348 Research focusing on the impact of AI tools on patient safety is critical. Finally,

349 studies should examine the ethical and legal aspects of AI in healthcare, particularly
350 regarding privacy, data security, and liabilities. Understanding these implications is
351 essential for the responsible incorporation of AI tools into clinical practice.

352

353 **Conclusions**

354 The use of temperature settings of ChatGPT-3.5 did not result in significant differences
355 in the overall quality of responses to drug queries, specifically those related to apixaban.

356 This suggests that ChatGPT-3.5 responses are not significantly affected by this setting.

357 However, the variability in accuracy highlights the need for careful consideration when
358 using this tool in clinical settings. Despite its potential as a supportive tool in

359 pharmaceutical information retrieval, its limitations, including the lack of real-time

360 internet access and the potential for the use of outdated information, must be

361 acknowledged. Healthcare professionals should use ChatGPT-3.5 as a supplementary

362 source, always verifying its output against current, evidence-based medical literature.

363 Future research should aim to evaluate chatbot performance across a broader range of

364 medications with larger and more diverse groups of healthcare professionals for a

365 comprehensive understanding of the capabilities and limitations of chatbots in the

366 context of clinical pharmacy.

367

368 **Acknowledgments**

369 We would like to thank Hiroki Yoshida for advice on statistical analysis and Editage
370 (www.editage.com) for English language editing and journal submission support. The
371 authors have authorized the submission of this manuscript through Editage.

372

373 **References**

- 374 1. Hassani H, Silva ES. The role of ChatGPT in data science: How AI-assisted
375 conversational interfaces are revolutionizing the field. *Big Data Cogn Comput.*
376 2023;7: 62. doi: 10.3390/bdcc7020062.
- 377 2. Al-Ashwal FY, Zawiah M, Gharaibeh L, Abu-Farha R, Bitar AN. Evaluating the
378 sensitivity, specificity, and accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and
379 Bard Against Conventional Drug-Drug Interactions Clinical Tools. *Drug Healthc*
380 *Patient Saf.* 2023;15: 137-147.
- 381 3. Wang X, Liu XQ. Potential and limitations of ChatGPT and generative artificial
382 intelligence in medical safety education. *World J Clin Cases.* 2023;11: 7935-
383 7939. doi: 10.12998/wjcc.v11.i32.7935.
- 384 4. Heck TG. What artificial intelligence knows about 70 kDa heat shock proteins,

- 385 and how we will face this ChatGPT era. *Cell Stress Chaperones*. 2023;28: 225-
386 229. doi: 10.1007/s12192-023-01340-1.
- 387 5. Morath B, Chiriac U, Jaszowski E, Deiß C, Nürnberg H, Hörth K, et al.
388 Performance and risks of ChatGPT used in drug information: An exploratory real-
389 world analysis. *Eur J Hosp Pharm*. 2023;31: 85-86. doi: 10.1136/ejhpharm-2023-
390 003750.
- 391 6. Kusunose K, Kashima S, Sata M. Evaluation of the accuracy of ChatGPT in
392 answering clinical questions on the Japanese Society of Hypertension Guidelines.
393 *Circ J*. 2023;87: 1030-1033. doi: 10.1253/circj.CJ-23-0308.
- 394 7. Lahat A, Shachar E, Avidan B, Glicksberg B, Klang E. Evaluating the utility of a
395 large language model in answering common patients' gastrointestinal health-
396 related questions: Are we there yet? *Diagnostics (Basel)*. 2023;13: 1950. doi:
397 10.3390/diagnostics13111950.
- 398 8. Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, et al. Assessing the
399 accuracy of responses by the language Model ChatGPT to questions regarding
400 bariatric surgery. *Obes Surg*. 2023;33: 1790-1796. doi: 10.1007/s11695-023-
401 06603-5.
- 402 9. Wagner MW, Ertl-Wagner BB. Accuracy of information and references using

- 403 ChatGPT-3 for retrieval of clinical radiological information. *Can Assoc Radiol*
404 *J.* 2024;75: 69-73. doi: 10.1177/08465371231171125.
- 405 10. Open AI. API reference. [Cited 4 May 2023]. Available from:
406 [https://platform.openai.com/docs/api-](https://platform.openai.com/docs/api-reference/completions/create#completions/create-temperature)
407 [reference/completions/create#completions/create-temperature.](https://platform.openai.com/docs/api-reference/completions/create#completions/create-temperature)
- 408 11. DRUG DISCOVERY: The 50 best-selling pharmaceuticals of 2022: COVID-19
409 vaccines poised to take a step back. [Cited 4 May 2023]. Available from:
410 [https://www.drugdiscoverytrends.com/50-of-2022s-best-selling-](https://www.drugdiscoverytrends.com/50-of-2022s-best-selling-pharmaceuticals/)
411 [pharmaceuticals/.](https://www.drugdiscoverytrends.com/50-of-2022s-best-selling-pharmaceuticals/)
- 412 12. BMS. HEALTHCARE Japan: Inquiries about our products. [Cited 4 May 2023].
413 Available from: <https://www.bmshealthcare.jp/medical/faq/pheq> (in Japanese).
- 414 13. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing
415 physician and artificial intelligence chatbot responses to patient questions posted
416 to a public social media forum. *JAMA Intern Med.* 2023;183: 589-596. doi:
417 10.1001/jamainternmed.2023.1838.
- 418 14. Chen AS, Revere L, Ratanatawan A, Beck CL, Allo JA. A comparative analysis
419 of academic and nonacademic hospitals on outcome measures and patient
420 satisfaction. *Am J Med Qual.* 2019;34: 367-375. doi:

- 421 10.1177/1062860618800586.
- 422 15. West RM. Best practice in statistics: Use the Welch t-test when testing the
423 difference between two groups. *Ann Clin Biochem.* 2021;58: 267-269. doi:
424 10.1177/0004563221992088.
- 425 16. MacFarland TW, Yates JM. Mann–Whitney U test. In: MacFarland TW, Yates
426 JM, editors. *Introduction to nonparametric statistics for the Biological Sciences*
427 *using R.* Cham. Springer International Publishing; 2016. pp. 103-132.
- 428 17. Kanda Y. Investigation of the freely available easy-to-use software ‘EZR’ for
429 medical statistics. *Bone Marrow Transplant.* 2013;48: 452-458. doi:
430 10.1038/bmt.2012.244.
- 431 18. Huang X, Estau D, Liu X, Yu Y, Qin J, Li Z. Evaluating the performance of
432 ChatGPT in clinical pharmacy: A comparative study of ChatGPT and clinical
433 pharmacists. *Br J Clin Pharmacol.* 2024;90: 232-238. doi: 10.1111/bcp.15896.
- 434 19. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb
435 SN, et al. Revolutionizing healthcare: The role of artificial intelligence in
436 clinical practice. *BMC Med Educ.* 2023;23: 689. doi: 10.1186/s12909-023-
437 04698-z.
- 438 20. Al-Dujaili Z, Omari S, Pillai J, Al Faraj A. Assessing the accuracy and consistency

439 of ChatGPT in clinical pharmacy management: A preliminary analysis with
440 clinical pharmacy experts worldwide. Res Social Adm Pharm. 2023;19: 1590-
441 1594. doi: 10.1016/j.sapharm.2023.08.012.

442

443 **Supporting information**

444 **S1 File. Questions entered in Japanese were translated into English for this study.**

445 **S2 File. Criteria for evaluating the quality of responses used in this study.**

446