Check for updates

Statistical refinement of patient-centered case vignettes for digital health research

Marvin Kopka* and Markus A. Feufel

Division of Ergonomics, Department of Psychology and Ergonomics (IPA), Technische Universität Berlin, Berlin, Germany

Digital health research often relies on case vignettes (descriptions of fictitious or real patients) to navigate ethical and practical challenges. Despite their utility, the quality and lack of standardization of these vignettes has often been criticized, especially in studies on symptom-assessment applications (SAAs) and self-triage decision-making. To address this, our paper introduces a method to refine an existing set of vignettes, drawing on principles from classical test theory. First, we removed any vignette with an item difficulty of zero and an item-total correlation below zero. Second, we stratified the remaining vignettes to reflect the natural base rates of symptoms that SAAs are typically approached with, selecting those vignettes with the highest item-total correlation in each quota. Although this two-step procedure reduced the size of the original vignette set by 40%, comparing self-triage performance on the reduced and the original vignette sets, we found a strong correlation ($r = 0.747$ to $r = 0.997$, $p < .001$). This indicates that using our refinement method helps identifying vignettes with high predictive power of an agent's self-triage performance while simultaneously increasing cost-efficiency of vignette-based evaluation studies. This might ultimately lead to higher research quality and more reliable results.

1 Introduction

In the field of digital health research, short case vignettes—that involve either fictitious or real medical scenarios—have become a widely accepted methodology (1–3). In contrast to detailed vignettes used in medical education, vignettes in digital health research are often brief and designed for specific tools or research objectives. The reliance on vignettes is primarily due to practical constraints: direct involvement of patients may be unfeasible and challenging due to ethical concerns, comparability across patients can be limited, and specific research scenarios may present additional barriers to using real patients (4, 5). To mitigate these constraints, researchers frequently use case vignettes as proxies to conduct these studies. However, the vignettes used in digital health research are often developed in an unstandardized way and without theoretical foundation (6–9).

Particularly in research focused on symptom-assessment applications (SAAs), many studies have adopted a set of vignettes developed by Semigran et al. in 2015 (6). This vignette set—which was derived from diverse medical resources—has not only been used in studies examining the self-triage (deciding if and where to seek care) performance of SAAs and laypeople but also in the evaluation of large language models

(LLMs) (10–13). Although these vignettes marked a significant step in evaluating SAAs, they have also been criticized. Key concerns include the development process (e.g., scenarios based on medical textbooks may not reflect ecologically valid descriptions of real patients), and the lacking validation of these vignettes (i.e., all developed vignettes are used without any quality assessment) (8, 9, 14). This criticism raised not only questions about the suitability of vignettes for accurately estimating the self-triage performance of human and digital agents, but also whether some vignettes might be better suited for evaluations than others (9). For example, it may be unclear which cases are easier or more difficult to solve and whether vignettes have incremental predictive power or could be omitted. It is also unclear if the predictive power of a vignette differs between human and digital agents and if different vignette sets might be needed for each agent. Thus, it is no surprise that Painter et al. recommend creating guidelines to identify which vignettes to include in accuracy evaluations (8).

The problems associated with existing vignettes highlight an urgent need for a more systematic and theory-driven approach to developing case vignettes. Ecological psychology and test theory provide a framework for addressing these challenges (15). By applying test-theoretical approaches, researchers can identify vignettes with high predictive power for assessing the performance of SAAs and other digital tools. In a previous study, we have shown that test-theoretic metrics can be readily applied to case vignettes in digital health research and that the current sets of vignettes [e.g., the one suggested by Semigran et al. (6)] are problematic in this regard (9). In another study we have outlined a method based on Egon Brunswik's concept of representative design to develop vignettes with high ecological validity (16).

The current paper takes these efforts a step further by detailing how to refine any existing vignette set using test-theoretical metrics to increase internal validity. This procedure allows selecting only those vignette subsets that best satisfy test-theoretical criteria and can thus predict performance. This refined set aims to make case vignettes studies more cost-efficient by reducing the number of vignettes required while simultaneously identifying and maintaining the vignettes that most effectively predict the performance of different diagnostic agents.

2 Method

2.1 Study design

This study presents a two-step procedure based on test theory to refine an existing set of case vignettes to test the self-triage performance of different agents. Specifically, our goal is to refine the full set and arrive at a validated subset of vignettes for each agent. To validate the presented vignette-refinement procedure, we compare the performance of the original and the refined set of vignettes based on data collected from laypeople, SAAs and LLMs in a previous study (16).

2.2 The original vignette set and data

The original vignettes were developed with the RepVig Framework according to the principles of representative design as outlined in Kopka et al. (16). That is, the vignettes were selected through random sampling of actual patient descriptions, where individuals presented symptoms in their own words and asked whether and where to seek medical care. These vignettes were left unaltered to reflect patients' real experiences, rather than modifying them for medical plausibility. The full vignette set was developed to reflect the natural base rate of symptoms that SAAs are approached with, using symptom clusters for stratification as reported by Arellano Carmona et al. (17).

A total of 198 laypeople (from Germany with no medical training, who were sampled from the online platform Prolific) evaluated the urgency of these case vignettes (20 vignettes each, resulting in a total of 3,960 assessments). Additionally, the dataset encompasses evaluations from 13 SAAs, each tested across all vignettes by two research assistants without a professional medical background. Since not every SAA gave a recommendation for each case, only cases where the SAA provided advice were included in the analysis. Furthermore, for LLMs, the lead author collected data on five LLMs (GPT-4, Llama 2, PaLM 2, Pi, Claude 2) that were openly available and offered a chat interface. For obtaining advice from these LLMs, we used a one-shot prompt developed by Levine et al. (13).

2.3 Statistical refinement

To refine the vignettes, we applied test-theoretical metrics in a first step: *item difficulty* (ID, how difficulty vignettes were to solve for any one agent) and *item-total-correlations* (ITC, how solving a given vignette correlates with solving other vignettes of the same acuity) (9, 18, 19). These metrics were calculated for the full vignette set. For each agent, vignettes with an ID of zero were excluded because they were unsolvable for the corresponding agent, and thus offered no insight into differences in self-triage performance. Vignettes with a low ID (i.e., solved by only few agents) were not excluded, as their inclusion provides insights into the capabilities of those few agents that do solve them. Similarly, vignettes with negative ITC values were removed because they negatively correlate with self-triage performance, rendering them unsuitable for performance evaluation according to test theory standards (18).

In a refined vignette set, the natural base rates should reflect real-world conditions to ensure a representative distribution of symptom clusters, preventing any one cluster from disproportionately influencing evaluation results. If the natural base rates are representative in the original set already (e.g., in this study), the base rates should be maintained; otherwise, they should be adjusted. In a second step, we thus evaluated the relative proportion of the smallest remaining symptom cluster and adjusted all other clusters to match this proportion to ensure that the base rates in the refined vignette set remained unaltered. For instance, if the smallest quota retained 3 out of 5 vignettes, this corresponds to 60% of the original size. Consequently, we adjusted the size of all

other quotas to reflect this proportion and included 60% of vignettes in the other quotas as well. In each quota, we retained the vignettes with the highest ITCs as these have the highest predictive power.

2.4 Validation

Finally, to validate the refined set, we compared the degree of association between performance estimates derived from the subset with those derived from the full set. We used common metrics for self-triage performance evaluations: accuracy, accuracy for each self-triage level, safety of advice, inclination to overtriage, and—this metric can only be calculated for SAAs—the capability comparison score, which is a score that allows capability comparisons between SAAs that were tested with differing vignette sets (9). We calculated these metrics using the symptomcheckR package (20) for every person, every SAA, and every LLM using both the full and the refined vignette sets and assessed the degree of similarity between these outcomes using

Pearson correlation. Following a review by Overholser and Sowinski (21), we interpreted a correlation above 0.90 as very high, between 0.70 and 0.90 as high, between 0.40 and 0.70 as moderate and below 0.40 as low or negligible.

3 Results

3.1 Refinement

Step 1 for laypeople: All (45/45) vignettes from the full vignette set have an ID greater than zero. Thus, no vignettes were excluded because of the ID. Six vignettes have an ITC below zero and were excluded, see Figure 1.

Step 2 for laypeople: The biggest reduction occurred within the "other pain" symptom cluster, which was reduced from 5 cases to 3 cases. This reduction corresponds to a 60% retention rate. Consequently, we adjusted the size of all quotas to reflect this new size, reducing them to 60% of their original

FIGURE 1
ID (left) and ITC (right) for each full vignette and agent. Red points represent vignettes that do not satisfy test-theoretic criteria and were excluded. Green points show vignettes that are included in the refined vignette set.
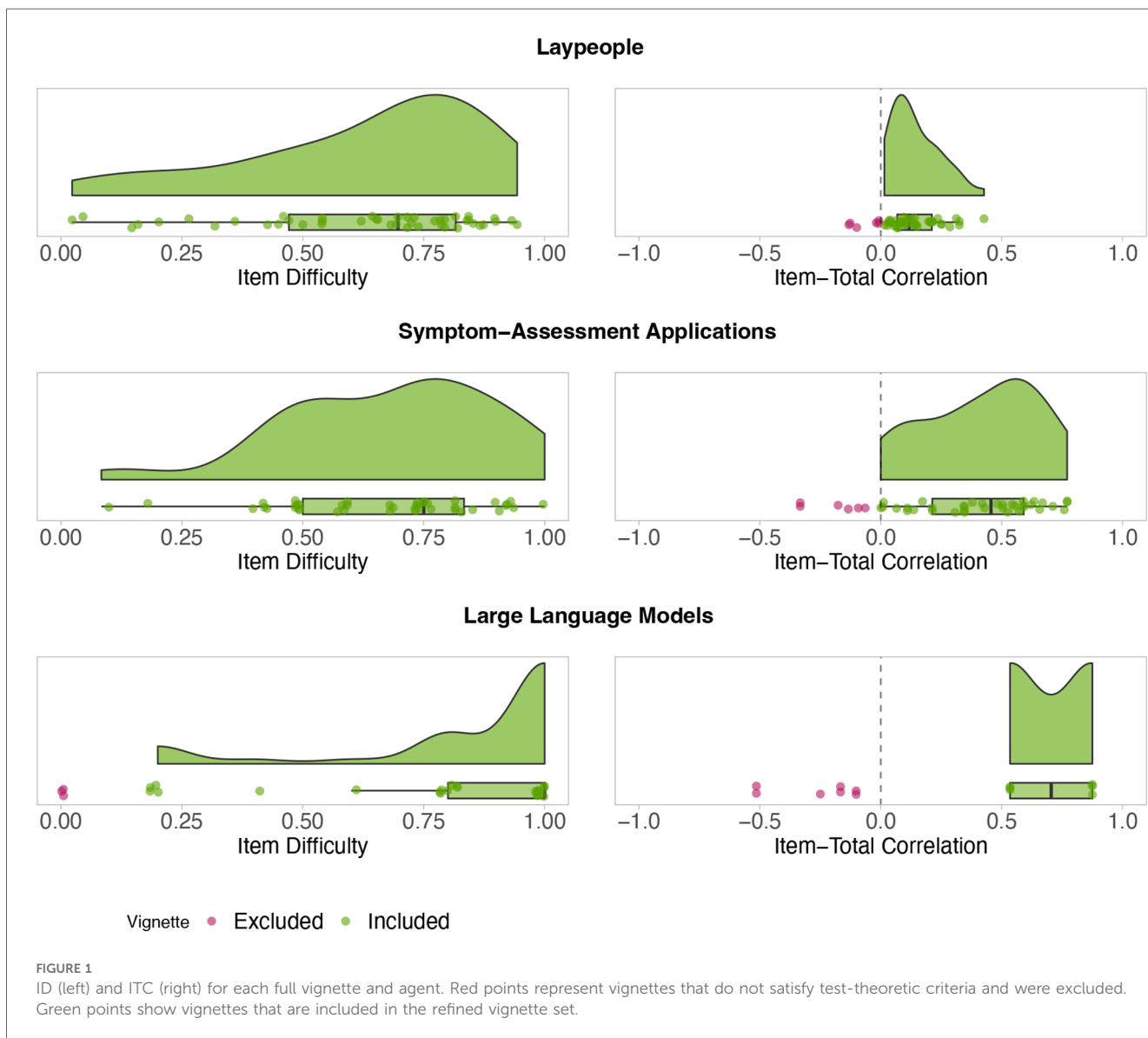
TABLE 1 New quotas for refined vignette sets for laypeople and SAAs after excluding vignettes that did not satisfy test-theoretic criteria. The clusters are based on Arellano Carmona et al. (17).

Symptom cluster	# of vignettes in full set	# of vignettes in refined set
Musculoskeletal pain	5	3
Joint pain	3	2
Headache	1	<1, joined with other clusters
Chest pain	1	<1, joined with other clusters
Other pain	5	3
Gynecological	5	3
Tumors/lumps/masses	4	2
Edema	2	1
Skin issues	2	1
Gastrointestinal	2	1
Impaired sensations	3	2
Urinary tract problems	3	2
Upper respiratory symptoms	1	<1, joined with other clusters
Other	8	5

TABLE 2 Correlations of metrics for laypeople for the refined vs. the original vignette set.

Metric	Correlation coefficient	p value
Average accuracy	0.846	<.001
Accuracy for emergencies[a]	1	<.001
Accuracy for non-emergencies	0.747	<.001
Accuracy for self-care	0.951	<.001
Safety of advice	0.845	<.001
Inclination to overtriage	0.834	<.001

[a]Included the same two cases as the full set.

TABLE 3 Correlations of metrics for SAAs for the refined vs. the original vignette set.

Metric	Correlation coefficient	p value
Average accuracy	0.995	<.001
Accuracy for emergencies[a]	0.577	.04933
Accuracy for non-emergencies	0.989	<.001
Accuracy for self-care	0.986	<.001
Safety of advice	0.970	<.001
Inclination to overtriage	0.974	<.001
Capability comparison score	0.997	<.001

[a]Included only one out of two original cases. Correlation is unreliable, because estimates are either 1 or 0.

sizes. The new quota sizes are shown in Table 1. In each symptom cluster, these quotas are filled beginning with the vignettes that have the highest ITC. For example, in the cluster "musculoskeletal pain", this corresponds to the vignettes 1 ($ITC_1 = 0.312$), 4 ($ITC_4 = 0.249$), and 2 ($ITC_2 = 0.204$), as the other vignettes had lower ITCs with $ITC_5 = 0.151$ and $ITC_3 = 0.064$. Clusters that only had 1 vignette originally ($n = 3$ clusters) are joined together and the two vignettes with the highest ITC in this joined cluster remain in the filtered set. The refined vignette set can be found in the Supplementary Table S1.

Step 1 for SAAs: All (45/45) vignettes have an ID greater than zero and remained in the set, while six vignettes had an ITC below zero, see Figure 1. These cases were excluded.

Step 2 for SAAs: The biggest reduction (2 vignettes) occurred in the "other pain" symptom cluster again. Because this represents the smallest new cluster now, all quotas are reduced to 60% of their original size (with the same retention rate as for laypeople, see Table 1). Those quotas are filled again with vignettes with the highest ITC in each cluster. For "musculoskeletal pain", this corresponds to the vignettes 5 ($ITC_5 = 0.634$), 2 ($ITC_2 = 0.586$), 4 ($ITC_4 = 0.504$), as the other vignettes had a lower ITC with $ITC_1 = 0.307$ and $ITC_3 = 0.110$. The refined vignette set for SAAs can be found in the Supplementary Table S2.

For LLMs, we identified nine vignettes with an ID of zero and seven vignettes with a negative ITC in step 1, all of which must be excluded, see Figure 1. However, only six vignettes have a positive ITC and for the remainder of the vignettes, an ITC could not be determined. Given the high number of exclusions and the inability to assess ITC for many vignettes, refining the vignette set for LLMs is unfeasible. Consequently, the full set must be retained for future evaluations.

3.2 Comparing the refined with the original vignette set

The metrics obtained for each person using the refined vignette set showed a very high correlation with metrics obtained using the original

vignette set, see Table 2. Similarly, the metrics obtained for each SAA using the refined vignette set showed very high correlations with metrics obtained using the original vignette set, see Table 3.

4 Discussion

Overall, our analysis demonstrates that refining a vignette set as outlined in this study proves feasible. Results using the refined set show a very high correlation with performance based on the complete original set, indicating minimal loss of predictive power despite using fewer vignettes. This approach not only makes evaluations more cost-efficient by using fewer vignettes, but also ensures that only vignettes accurately predicting overall performance are included, thereby yielding more reliable performance estimates and higher internal validity. This effort aligns with the call for standardized vignettes and their refinement (7–9, 22). Answering Painter et al.'s call for guidance on which vignettes to include in self-triage evaluation studies (8), our two-step procedure offers a systematic, theory-driven way to refine an initial set of vignettes and select the most predictive vignettes out of a full set.

In our data, the ID was less relevant for refining vignettes for laypeople and SAAs, because at least one person or SAA managed to solve each case. The ID proved more meaningful for LLMs, however, which could rarely solve the self-care cases. This thwarted the refinement process for LLMs, because an ITC value was impossible to calculate for those vignettes that could not be solved by LLMs. In the current dataset, this problem might be due to the small number of LLMs (only five) included and the resulting low variance. With a higher number of LLMs, a refinement might be possible, but the number of different LLMs is currently limited.

The generated vignette sets vary between laypeople and SAAs. That is, a set validated for SAAs might not be suitable for assessing laypeople's performance. So, researchers must refine the vignette set and collect data to validate it for each agent they wish to generalize to. Specifically, research should initially collect data using the full set and then refine and validate a subset to be used in follow-up studies with the same agent.

5 Limitations and future directions

Our study comes with limitations that should be addressed in future studies. For example, we entered data into LLMs only once and could not account for output variability due to the lack of established methods. Future studies might explore pooling algorithms to code varying outputs into a consolidated datapoint for evaluation studies. Further, we focused on classical test theory to refine our vignette set, but item-response theory (IRT) could offer an alternative theoretical framework for the refinement. However, most models would require bigger sample sizes, which are often not available due to the limited number of SAAs (15). Until larger samples of SAAs become available, test theory is the best choice for refining vignettes for all agents. For laypeople, where larger sample sizes of more than 200 participants are feasible (23), IRT might yield different refined vignette sets. Comparing vignette sets refined through classical test theory and IRT would be a valuable next step. Additionally, assessing the validity (e.g., convergent and divergent validity) of case vignette sets presents a further research opportunity, for example, through external validation.

6 Conclusions

Our two-step vignette refinement procedure offers a significant advancement for the evaluation of self-triage decision-making and digital health research at large. By systematically excluding vignettes that are unsuitable for measuring the constructs (e.g., self-triage accuracy) researchers wish to assess, they can avoid arbitrary selection of vignettes and ensure that only statistically validated vignettes are included in the test set. This approach can enhance the internal validity and quality and reduce the costs of digital health research, lead to more reliable results, and enable more precise inferences in the long run. If more researchers apply the presented refinement method, methodological rigor and research quality will increase, which helps move the field forward and ultimately contributes to the development of more effective digital health tools and interventions.

Data availability statement

The dataset analyzed in this study is available upon reasonable request. Requests to access these datasets should be directed to marvin.kopka@tu-berlin.de.

Ethics statement

The studies involving humans were approved by Department of Psychology and Ergonomics (IPA) at Technische Universität Berlin. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2024.1411924/full#supplementary-material

References

1. Riboli-Sasco E, El-Osta A, Alaa A, Webber I, Karki M, El Asmar ML, et al. Triage and diagnostic accuracy of online symptom checkers: systematic review. *J Med Internet Res.* (2023) 25(e43803):1–18. doi: 10.1177/20552076231194929

2. McInroy LB, Beer OWJ. Adapting vignettes for internet-based research: eliciting realistic responses to the digital milieu. *Int J Soc Res Methodol.* (2022) 25(3):335–47. doi: 10.1080/13645579.2021.1901440

3. Riley AH, Critchlow E, Birkenstock L, Itzoe M, Senter K, Holmes NM, et al. Vignettes as research tools in global health communication: a systematic review of the literature from 2000 to 2020. *J Commun Healthc.* (2021) 14(4):283–92. doi: 10.1080/17538068.2021.1945766

4. Converse L, Barrett K, Rich E, Reschovsky J. Methods of observing variations in physicians' decisions: the opportunities of clinical vignettes. *J Gen Intern Med.* (2015) 30(S3):586–94. doi: 10.1007/s11606-015-3365-8

5. Matza LS, Stewart KD, Lloyd AJ, Rowen D, Brazier JE. Vignette-based utilities: usefulness, limitations, and methodological recommendations. *Value Health.* (2021) 24(6):812–21. doi: 10.1016/j.jval.2020.12.017

6. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. *Br Med J.* (2015) 351:1–9. doi: 10.1136/bmj.h3480

7. Wallace W, Chan C, Chidambaram S, Hanna L, Iqbal FM, Acharya A, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. *NPJ Digit Med.* (2022) 5(1):118. doi: 10.1038/s41746-022-00667-w

8. Painter A, Hayhoe B, Riboli-Sasco E, El-Osta A. Online symptom checkers: recommendations for a vignette-based clinical evaluation standard. *J Med Internet Res.* (2022) 24(10):e37408. doi: 10.2196/37408

9. Kopka M, Feufel MA, Berner ES, Schmieding ML. How suitable are clinical vignettes for the evaluation of symptom checker apps? A test theoretical perspective. *Digit Health.* (2023) 9(20552076231194929):1–17. doi: 10.1177/20552076231194929

10. Hill MG, Sim M, Mills B. The quality of diagnosis and triage advice provided by free online symptom checkers and apps in Australia. *Med J Aust.* (2020) 212 (11):514–9. doi: 10.5694/mja2.50600

11. Schmieding ML, Kopka M, Schmidt K, Schulz-Niethammer S, Balzer F, Feufel MA. Triage accuracy of symptom checker apps: 5-year follow-up evaluation. *J Med Internet Res.* (2022) 24(5):e31810. doi: 10.2196/31810

12. Schmieding ML, Mörgeli R, Schmieding MAL, Feufel MA, Balzer F. Benchmarking triage capability of symptom checkers against that of medical laypersons: survey study. *J Med Internet Res.* (2021) 23(3):e24475. doi: 10.2196/24475

13. Levine DM, Tuwani R, Kompa B, Varma A, Finlayson SG, Mehrotra A, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. *Lancet Digit Health.* (2024) 6(8):e555–61. doi: 10.1016/S2589-7500(24)00097-9

14. El-Osta A, Webber I, Alaa A, Bagkeris E, Mian S, Taghavi Azar Sharabiani M, et al. What is the suitability of clinical vignettes in benchmarking the performance of online symptom checkers? An audit study. *BMJ Open.* (2022) 12(4):e053566. doi: 10.1136/bmjopen-2021-053566

15. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. *Med Educ.* (2010) 44(1):109–17. doi: 10.1111/j.1365-2923.2009.03425.x

16. Kopka M, Napierala H, Privoznik M, Sapunova D, Zhang S, Feufel M. *Evaluating self-triage accuracy of laypeople, symptom-assessment apps, and large language models: a framework for case vignette development using a representative design approach (RepVig).* medRxiv. (2024). p. 2024.04.02.24305193. doi: 10.1101/2024.04.02.24305193v1 (cited April 3, 2024).

17. Carmona KA, Chittamuru D, Kravitz RL, Ramondt S, Ramírez AS. Health information seeking from an intelligent web-based symptom checker: cross-sectional questionnaire study. *J Med Internet Res.* (2022) 24(8):e36322. doi: 10.2196/36322

18. Kelava A, Moosbrugger H. Deskriptivstatistische evaluation von items (itemanalyse) und testwertverteilungen. In: Moosbrugger H, Kelava A, editors. *Testtheorie und Fragebogenkonstruktion.* Berlin, Heidelberg: Springer (2012). p. 75–102. (Springer-Lehrbuch). doi: 10.1007/978-3-642-20072-4_4 (cited September 3, 2021).

19. Möltner A, Schellberg D, Jünger J. Grundlegende quantitative analysen medizinischer prüfungen. *GMS J Med Educ.* (2006) 23(3):1–11.

20. Kopka M, Feufel MA. Software symptom check R: an R package for analyzing and visualizing symptom checker triage performance. *BMC Digit Health.* (2024) 2 (1):43. doi: 10.1186/s44247-024-00096-7

21. Overholser BR, Sowinski KM. Biostatistics primer: part 2. *Nutr Clin Pract.* (2008) 23(1):76–84. doi: 10.1177/011542650802300176

22. Kopka M, Feufel MA, Balzer F, Schmieding ML. The triage capability of laypersons: retrospective exploratory analysis. *JMIR Form Res.* (2022) 6(10):e38977. doi: 10.2196/38977

23. Chen WH, Lenderking W, Jin Y, Wyrwich KW, Gelhorn H, Revicki DA. Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Qual Life Res.* (2014) 23(2):485–93. doi: 10.1007/s11136-013-0487-5