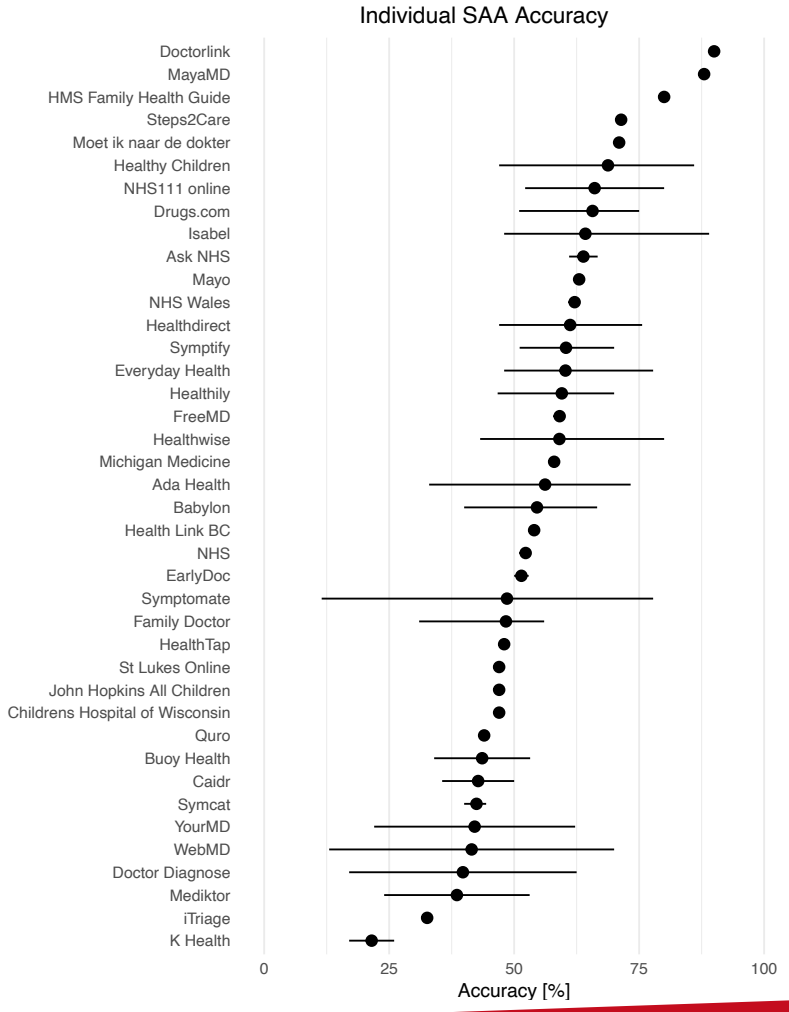
Vergleich von Ersteinschätzungssystemen

Dr. Dr. Marvin Kopka | Urgent Care Conference | 24.06.2026

Fachgebiet Arbeitswissenschaft, Technische Universität Berlin

Genauigkeit von Systemen

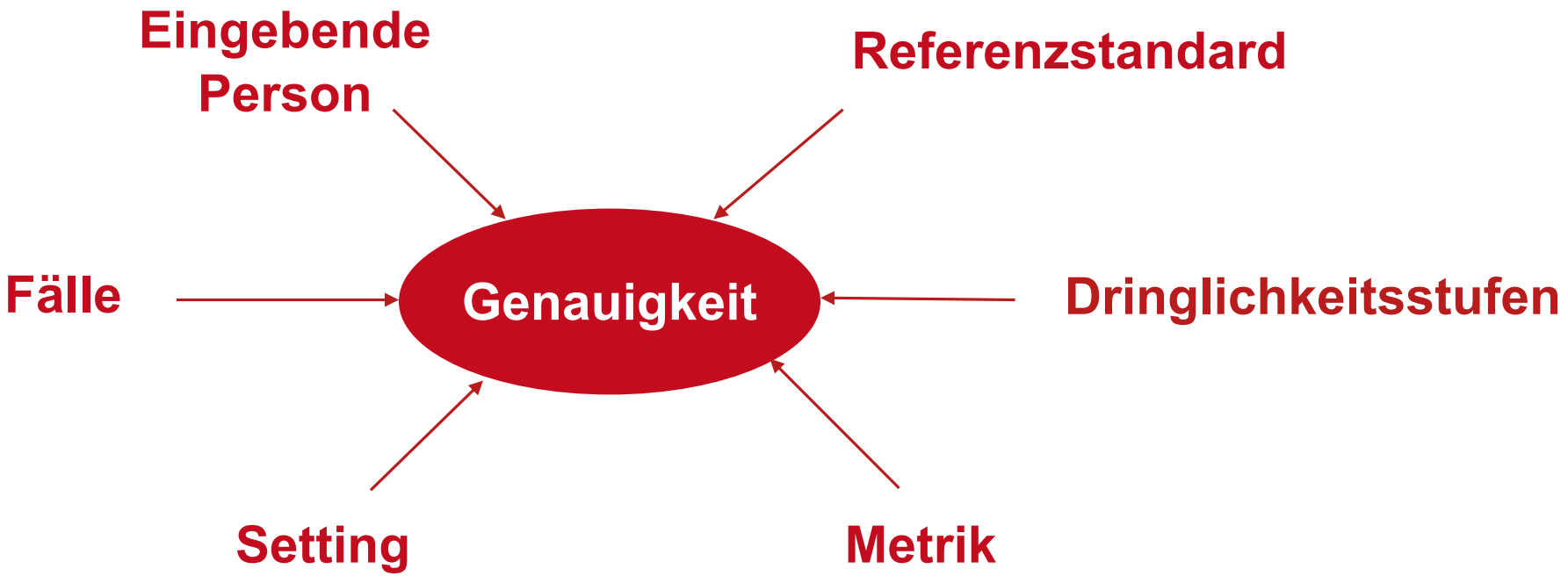


Kopka, M., von Kalckreuth, N., & Feufel, M.A. (2025). Accuracy of online symptom assessment applications, large language models, and laypeople for self-triage decisions. *npj Digital Medicine*, *8*, 178. https://doi.org/10.1038/s41746-025-01566-6

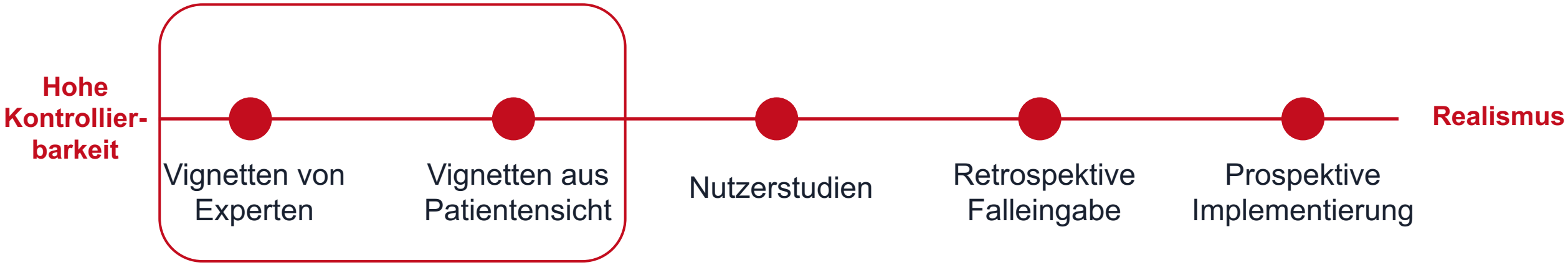
Genauigkeit von Systemen

Kopka, M., von Kalckreuth, N., & Feufel, M.A. (2025). Accuracy of online symptom assessment applications, large language models, and laypeople for self-triage decisions. *npj Digital Medicine*, *8*, 178. https://doi.org/10.1038/s41746-025-01566-6

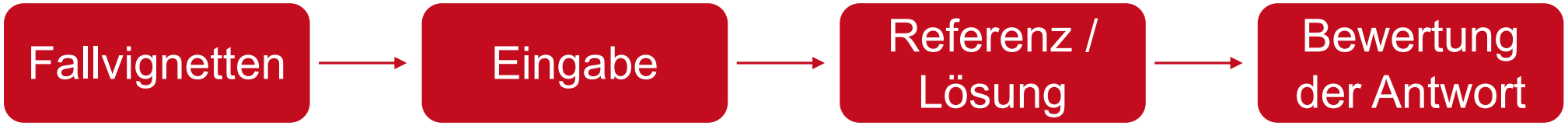
Genauigkeit?



Ebenen der Evaluation



Vorgehen bei Fallvignettenstudien

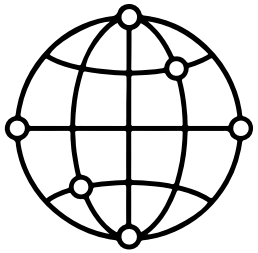


Probleme mit Fallvignetten

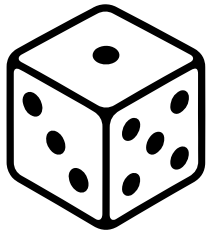
- Beschreibung von echten Fällen?
- Beschreibung von Ärzten oder von Patienten selbst?
- Menge an Informationen? Laborergebnisse?
- Prävalenz von Symptomen repräsentativ?
- Anzahl an Vignetten?

Kopka, M., & Feufel, M.A. (2026). How to Evaluate the Accuracy of Symptom Checkers and Diagnostic Decision Support Systems: Symptom Checker Accuracy Reporting Framework (SCARF). *JMIR Human Factors*, *13*. https://doi.org/10.2196/76168

Lösungen für Fallvignetten - Repräsentative Patientenfälle



r/AskDocs
(n = 8.794)



Zufallsziehung
& Bewertung

Prävalenz

Vorläufiges
Vignettenset
(n = 45)

statistische
Kriterien

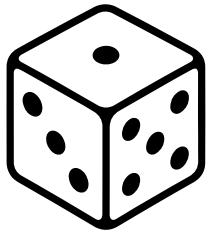
Finales
Vignettenset
(n = 27)

Kopka, M., Napierala, H., Privoznik, M. et al. The RepVig framework for designing use-case specific representative vignettes and evaluating triage accuracy of laypeople and symptom assessment applications. *Scientific Reports* 14, 30614 (2024). https://doi.org/10.1038/s41598-024-83844-z

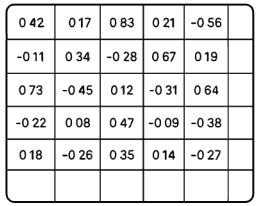
Lösungen für Fallvignetten - Repräsentative Nutzerfälle

116
117

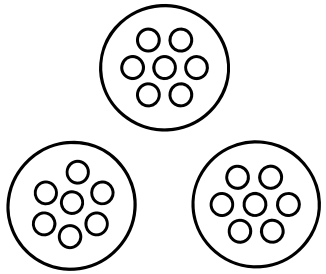
Triage Anfragen
(n = 3.2 Millionen)



Zufallsziehung
(n = 50.000)



Semantic
Embeddings



Cluster
(n = 514)

LLM-basierte Rückübersetzung in Fallvignetten
(n = 48.400)

Seidel, A., Steiger, E., Schuster, J., & Kroll, L. E. (2026). A Data-Driven Framework for Generating Population-Linked Case Vignettes from Nationwide Triage Data. *medRxiv*. https://doi.org/10.64898/2026.06.08.26354886

Probleme mit Eingabe

- Wer gibt Fälle ein? Laien? Experten?
- Welche Annahmen werden getroffen?
- Fehlende Information durch klinische Expertise gefüllt?
- Wie viele Personen sollten Fälle eingeben?

Kopka, M., & Feufel, M.A. (2026). How to Evaluate the Accuracy of Symptom Checkers and Diagnostic Decision Support Systems: Symptom Checker Accuracy Reporting Framework (SCARF). *JMIR Human Factors*, *13*. https://doi.org/10.2196/76168

Bildquelle: Marvin Kopka

Lösungen für Eingabe - Adjustierte Metriken & Eingabeprotokoll

- Eingabe von mind. 3 Personen
- Meta-Genauigkeit: Genauigkeit bei der Eingabe von Symptomen
- Standardisierte Eingabeprotokolle

Meczner, A., Cohen, N., Qureshi, A., Reza, M., Sutaria, S., Blount, E., Bagyura, Z., & Malak, T. (2024). Controlling Inputter Variability in Vignette Studies Assessing Web-Based Symptom Checkers: Evaluation of Current Practice and Recommendations for Isolated Accuracy Metrics. *JMIR Formative Research*, 8, e49907. https://doi.org/10.2196/49907

Lösungen für Eingabe - Adjustierte Metriken & Eingabeprotokoll

Chief complaint(s)	- Copy and paste the prescribed chief complaint(s) word by word (if only one symptom can be added at the time, add one by one) - If none of the symptoms are accepted by OSC, stop the vignette entry - In case of multiple chief complaints, if at least one of the symptoms is successfully inputted, continue the consultation. - If not all the symptoms are successfully inputted but there is an option to add more, attempt to do so. If unsuccessful, continue with the consultation without adding them.
Demographics	- Copy over/input age and gender of the persona of the vignette

Instruction for answering questions during the question-answer part of the consultation	- Only select symptoms that are mentioned in the vignette
	- Do not to select any symptom that is not described in the vignette even if you feel it might be missing from the vignette
	- Caveats:
	  - If synonyms offered for symptom described in the vignette, select it e.g. “abdominal pain” for “tummy pain”
	  - If a wider logical category of a symptom is offered, select it e.g. if “knee pain” is in the vignette, then select “leg pain” if offered
	  - If the sentence in the vignette is worded differently but has the same meaning, select it.
	  - If a sentence in a vignette indirectly could lead to a symptom or vice versa, then do not select that symptom. E.g. if someone is 'waking up at night with pain', but the symptom checker asks if the 'pain is severe' - do not select that symptom as it is unknown whether the symptom is severe, just they are waking up at night with it. However, if the vignette states that someone is 'crying with the pain' then you can select it, as it may be understood as a ”synonym” of 'pain is severe'
	  - If further characteristics of a confirmed symptom is asked in one question with multiple answers, and there is an option to decline all of them, do so. If this is not possible but there is an option not to answer e.g. “I do not know”/”Not prefer to say”, choose that, otherwise, stop the consultation and mark the case as “INCOMPLETE”
	- If some of the chief complaint(s) are not recognised but asked later during the consultation, confirm them

Documentation	- Free text entry and other entered symptoms via drop-down - Confirmed chief complaint(s) - Outcome conditions - Triage advice - All the data points (including confirmed and declined symptoms) - depending on symptom checker, if possible copy the consultation report with as much detail as possible

Meczner, A., Cohen, N., Qureshi, A., Reza, M., Sutaria, S., Blount, E., Bagyura, Z., & Malak, T. (2024). Controlling Inputter Variability in Vignette Studies Assessing Web-Based Symptom Checkers: Evaluation of Current Practice and Recommendations for Isolated Accuracy Metrics. *JMIR Formative Research*, 8, e49907. https://doi.org/10.2196/49907

Lösungen für Eingabe - LLMs als Patientensimulation

- „Patient Agent“
- LLM erhält Vignette
- Antwortet nur auf die gestellte Frage auf Basis von Vignettendaten
- Vermeidet klinische Ausdrucksweise
- Füllt fehlende Informationen mit plausiblen Daten

Rashidian, S., Li, N., Amar, J., Lee, J. H., Pugh, S., Yang, E., Masterson, G., Cha, M., Jia, Y., & Vaid, A. (2025). AI Agents for Conversational Patient Triage: Preliminary Simulation-Based Evaluation with Real-World EHR Data. *arXiv*. https://doi.org/10.48550/ARXIV.2506.04032

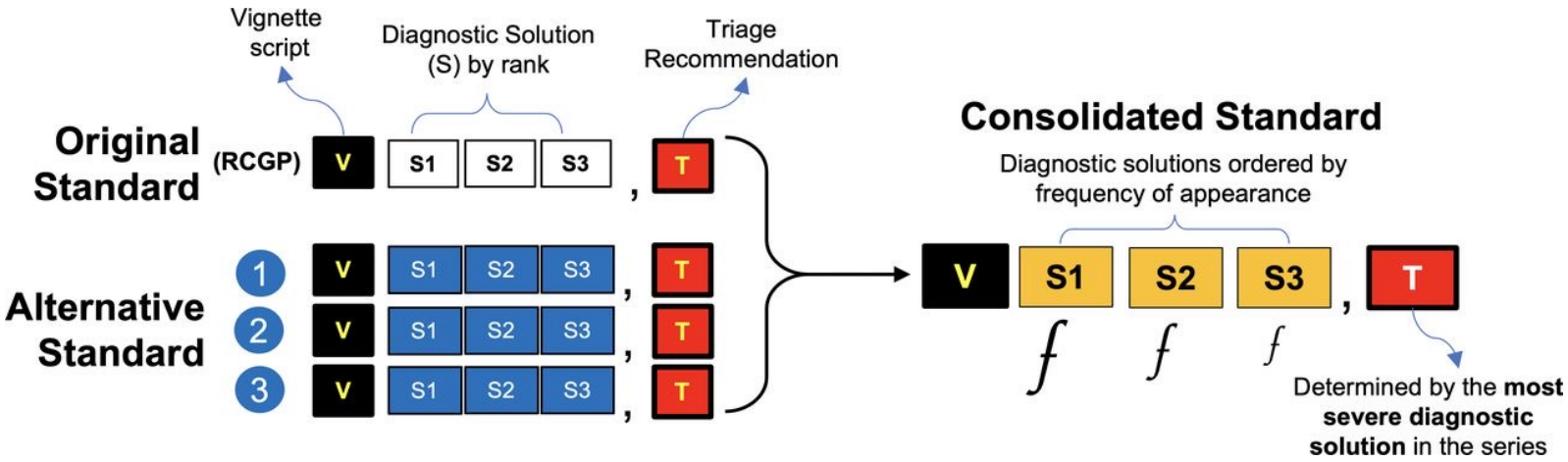
Probleme mit Referenz / Lösung

- Welches Verfahren?
- Finale Diagnose und Empfehlung? Bewertung durch Arzt? Ärzte? Ärztepanel?

Kopka, M., & Feufel, M.A. (2026). How to Evaluate the Accuracy of Symptom Checkers and Diagnostic Decision Support Systems: Symptom Checker Accuracy Reporting Framework (SCARF). *JMIR Human Factors*, *13*. https://doi.org/10.2196/76168

Lösungen für Referenz / Lösung - Ärztepanel

- Mehrere Ärztepanel bewerten Fälle unabhängig voneinander
- Lösungen werden von weiterem Panel aggregiert



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

Lösungen für Referenz / Lösung - Richtige & plausible Lösung

- Festlegung einer **richtigen** Lösung
- Festlegung weiterer **plausibler** Lösungen
- Genauigkeit für richtige und plausible Lösungen

Knitza, J., Hasanaj, R., Beyer, J., Ganzer, F., Slagman, A., Bolanaki, M., Napierala, H., Schmieding, M. L., Al-Zaher, N., Orlemann, T., Muehlensiepen, F., Greenfield, J., Vuillerme, N., Kuhn, S., Schett, G., Achenbach, S., & Dechant, K. (2024). Comparison of Two Symptom Checkers (Ada and Symptoma) in the Emergency Department: Randomized, Crossover, Head-to-Head, Double-Blinded Study. *Journal of Medical Internet Research*, 26, e56514. https://doi.org/10.2196/56514

Probleme mit Bewertung der Antwort

- Reicht Genauigkeit?
- Welche weiteren Metriken sollten berichtet werden?
- Wie wird Empfehlung von System den richtigen Lösungsstufen zugeordnet?

Kopka, M., & Feufel, M.A. (2026). How to Evaluate the Accuracy of Symptom Checkers and Diagnostic Decision Support Systems: Symptom Checker Accuracy Reporting Framework (SCARF). *JMIR Human Factors*, *13*. https://doi.org/10.2196/76168

Lösungen für Bewertung der Antwort - §360b GeDIG

- Genauigkeit
- Sensitivität (-> Positive Predictive Value?)
- Spezifität (-> Negative Predictive Value?)
- Weitere Metriken (Laienverständlichkeit, Informationssicherheit, …)

Lösungen für Bewertung der Antwort - symptomcheckR

- Systematische Übersicht zu berichteten Metriken
- Genauigkeit, Genauigkeit pro Level, Sicherheit, Vollständigkeit, Übertriage, Untertriage, Capability Comparison Score
- R-Package zur einfachen Bestimmung der Metriken

Kopka, M., & Feufel, M. A. (2024). Software symptomcheckR: An R package for analyzing and visualizing symptom checker triage performance. *BMC Digital Health*, *2*(1), 43. https://doi.org/10.1186/s44247-024-00096-7

Lösungen für Bewertung der Antwort - symptomcheckR

Kopka, M., & Feufel, M. A. (2024). Software symptomcheckR: An R package for analyzing and visualizing symptom checker triage performance. *BMC Digital Health*, *2*(1), 43. https://doi.org/10.1186/s44247-024-00096-7

Checklist – Symptom Checker Accuracy Reporting (SCARF)

How to Evaluate the Accuracy of Symptom Checkers and Diagnostic Decision Support Systems: Symptom Checker Accuracy Reporting Framework (SCARF)

Marvin Kopka[1]; Markus A Feufel[1]

- 19 Themen: Vignetten, Referenzstandard, Eingabe, Metriken, Analyse, …
- Standardisierung von Benchmarkingstudien
- Transparenz

Checklist – Symptom Checker Accuracy Reporting (SCARF)

Topic	Item Number	Item Description	Page Number
Title & Abstract			
Title	1	Title should indicate that the study evaluates a symptom checker or diagnostic decision support system	
Abstract	2	Summary of evaluation objective, methods, results, and conclusions	
Introduction			
Background and Objectives	3a	State the intended use case of the symptom checker (e.g., self-triage, emergency care triage)	
	3b	Define the target population to which findings are intended to generalize	
Methods			
Case Vignettes	4a	Describe the source of vignettes (e.g., medical education textbooks, patient records, patients' descriptions, case studies, fictitious)	
	4b	Report the sampling frame and rationale (e.g., which conditions, prevalence data, population-level statistics informed vignette selection or creation)	
	4c	Report how statistical representativeness of vignettes was ensured (i.e., representative of prevalence within the sampling frame)	
	4d	Report how content representativeness of vignettes was ensured (e.g., real cases, cases derived from patient records)	
	4e	Report whether atypical cases were included or excluded	
	4f	Specify the number of vignettes and provide a rationale (e.g., power analysis, feasibility)	
	4g	Describe refinement and selection procedures (e.g., test-theoretical metrics such as item-total correlations or item difficulty indices)	
	4h	State whether vignette content was lay-friendly or phrased for clinicians	
Gold Standard Assignment	5a	Describe how the reference standard was established	
	5b	Report the number and background of experts involved	
	5c	Explain how symptom checker outputs were mapped to triage categories	

Offene Fragen

- Wie viele Fälle sind sinnvoll und machbar?
- Qualität von LLM Patientensimulatoren?
- Genauigkeit von Referenzstandard?
- **Sinnvolle** Metriken?

Kontakt

Linkedin.com/in/marvin-kopka

@kopka_sci

mail@marvinkopka.com