DR. INFO at the Point of Care: A Prospective Pilot Study of an Agentic AI Clinical Assistant

Rogerio Corga Da Silva[1,2,3], Miguel Romano[1,2,3], Tiago Mendes[1], Marta Isidoro[1,4], Sandhanakrishnan Ravichandran[1], Shivesh Kumar[1], Michiel van der Heijden[1], Olivier Fail[1], Valentine Emmanuel Gnanapragasam[1*]

[1]Synduct GmbH, [2]University of Minho, School of Medicine,

[3]Unidade Local de Saúde (ULS) do Alto Minho, Intensive Care Medicine,

[4]ULS Almada-Seixal

Abstract

Background: Clinical documentation and information retrieval consume over half of physicians' working hours, contributing to cognitive overload and burnout. While artificial intelligence offers a potential solution, concerns over hallucinations and source reliability have limited adoption at the point of care.

Objective: To evaluate clinician-reported time savings, decision-making support, and satisfaction with DR. INFO, an agentic AI clinical assistant, in routine clinical practice.

Methods: In this prospective, single-arm pilot study, 29 clinicians across multiple specialties in Portuguese healthcare institutions used DR. INFO v1.0 over five working days within a two-week period. Outcomes were assessed via daily Likert-scale evaluations and a final Net Promoter Score. Non-parametric methods were used throughout.

Results: Clinicians reported high perceived time saving (mean 4.27/5; 95% CI: 3.97–4.57) and decision support (4.16/5; 95% CI: 3.86–4.45), with ratings stable across all study days and no evidence of attrition bias. The NPS was 81.2, with no detractors.

Conclusions: Clinicians across specialties and career stages reported sustained satisfaction with DR. INFO for both time efficiency and clinical decision support. Validation in larger, controlled studies with objective outcome measures is warranted.

*Correspondence: valentine.emmanuel@synduct.com

1 Introduction

The volume of biomedical literature and clinical data generated daily far exceeds what any individual clinician can assimilate, widening the gap between available evidence and bedside practice [1]. According to the American Medical Association, physicians work an average of 57.8 hours per week, of which only 27.2 are spent in direct patient care [2]. The balance is consumed by electronic health record documentation and administrative tasks, contributing to the 43.2% burnout prevalence reported nationally [2, 3].

Artificial intelligence has been proposed as a means of reducing this burden [4]. Recent studies have demonstrated that large language models can match or exceed physician performance on standardised medical examinations and patient queries [5–7], and a randomised clinical trial showed that access to a large language model influenced physician diagnostic reasoning [8]. Early usability assessments have also explored potential roles for conversational AI across the clinical workflow [9]. However, the clinical adoption of general-purpose large language models has been limited by well-documented safety concerns, most notably hallucinations, defined as the generation of plausible but factually incorrect outputs [10]. These evaluations have primarily tested standalone models rather than integrated clinical assistants deployed in routine practice. DR. INFO is an agentic AI clinical assistant designed to mitigate these risks by coupling large language model capabilities with retrieval from curated medical knowledge bases and peer-reviewed literature.

Between 2024 and 2025, Portugal ranked first in the European Union for access to digital health data [11], providing a favourable environment for the deployment of DR. INFO across several Local Health Units and private hospitals. This deployment coincides with the enforcement of the EU AI Act, which requires transparency and human oversight for high-risk AI systems. These principles guided the design of DR. INFO [12].

Prior to the study, DR. INFO was developed through a user-centred co-design process in which a multi-specialty group of clinicians, known as the "Innovator Circle," participated in iterative cycles of testing and feedback to ensure suitability for high-acuity clinical settings.

1.1 Technical Benchmarking: HealthBench

Separately from the present usability study, the technical performance of DR. INFO was evaluated using OpenAI's HealthBench, a benchmark comprising 5,000 real-world clinical conversations scored across five dimensions: accuracy, completeness, context awareness, communication quality, and instruction following [13]. The methodology and results are reported in a companion paper [14]. It should be noted that these comparisons reflect fundamental differences in system architecture: DR. INFO uses an agentic architecture that retrieves from curated medical knowledge bases and peer-reviewed literature before

generating a response, whereas comparator models were evaluated as standalone large language models without clinical retrieval infrastructure. The scores therefore reflect the combined value of architecture and retrieval, not model quality alone.

On the 1,000-case "Hard" subset, DR. INFO achieved a score of 0.68, compared with GPT-5 (0.40–0.46), Grok 3 (0.23), and Gemini 2.5 Pro (0.19). Against similar clinical AI assistants, DR. INFO scored 0.72 versus OpenEvidence (0.49) and DoxGPT (0.48). Full comparator details are provided in the companion study [14].

The present pilot study complements that technical evaluation by assessing physician-perceived value in routine clinical practice.

2 Methods

2.1 Pilot Study Design

This prospective, single-arm pilot study [15] was conducted over a two-week period in October 2025 using DR. INFO v1.0. The sample size of 29 reflects all available clinicians within the participating institutions during the study window; no formal power calculation was performed, consistent with the feasibility-oriented objectives of a pilot study [15]. Each participant integrated DR. INFO into their daily clinical workflow on five working days of their choosing within this window, not necessarily consecutive.

Data were collected through daily electronic Case Report Forms (eCRFs) and a final satisfaction questionnaire. Responses were recorded on Likert scales [16], and a Net Promoter Score (NPS) [17] was obtained at study completion.

Twenty-nine clinicians were enrolled on a voluntary basis, comprising 25 physicians and 4 medical students. Of these, 26 completed the baseline eCRF, 28 submitted at least one diary entry (90 named entries total), and 16 completed the final evaluation. Three participants completed the final evaluation but did not submit the baseline form, and 13 did not complete the final evaluation.

2.2 Instruments and Protocol

Three electronic Case Report Forms were administered in sequence. The baseline form (eCRF 1) captured years of practice as an ordinal variable (<1, <3, <10, >10 years), specialty, frequency of clinical decision support tool usage ("How frequently do you use clinical decision support tools in your practice?"), and technology comfort rated on a 5-point Likert scale ("How comfortable are you with technology, e.g., smartphones, AI, new software?").

The daily form (eCRF 2) logged two primary outcome measures on 5-point Likert scales (1 = strongly disagree, 5 = strongly agree): time saving ("To what extent do you agree that using DR. INFO today saved you time?") and decision support ("To what

extent do you agree that DR. INFO helped you improve your clinical decisions today, e.g., information seeking, information retrieval, diagnosis, treatment, management?"). Additionally, participants selected applicable clinical use cases from a predefined list and provided optional qualitative examples.

The final form (eCRF 3) collected the Net Promoter Score ("On a scale of 0 to 10, how likely are you to recommend DR. INFO to a colleague?") and two open-ended qualitative questions: "What did you like most about DR. INFO?" and "What did you like least about DR. INFO? What could be improved?"

2.3 Statistical Analysis

Given the exploratory nature of the study, the ordinal measurement scale, and the small sample size, non-parametric methods were used throughout. Attrition analysis compared diary scores between final eCRF completers and non-completers using the Mann-Whitney U test. Ninety-five percent confidence intervals were computed using both percentile bootstrap resampling (10,000 iterations, resampled at the participant level) and t-distribution methods. Longitudinal stability was assessed using the Friedman test (non-parametric repeated measures) applied to the 10 participants with data on all five study days. Spearman rank correlation coefficients were calculated for associations between ordinal variables (years of practice, technology comfort, and tool usage frequency) and the two primary outcome measures (time saving and decision support scores).

Subgroup comparisons across clinical roles were conducted using the Kruskal-Wallis H test. Analysis was performed by role rather than medical specialty, as the majority of specialty subgroups contained fewer than five participants, below the minimum recommended for reliable chi-squared approximation of the H statistic [18]. Because 13 of 29 participants did not complete the final eCRF, NPS was recalculated under several assumptions (e.g., all non-responders scored as Passives, all as Detractors) to test whether the observed result was robust to non-response bias. Open-ended responses were categorized into recurring themes (e.g., speed, source transparency, interface) through content analysis of the 16 responses by a single coder.

All analyses were conducted at the participant level (using each participant's mean across diary entries) to avoid pseudoreplication from repeated measures. No correction for multiple comparisons was applied; all reported associations should be interpreted as exploratory.

2.4 Data Quality and Transparency

During data cleaning, several issues were identified and corrected. Free-text specialty entries were standardized from 24 unique labels to seven categories. Name inconsistencies across forms, such as abbreviated names on later entries, were resolved through manual

matching, reducing the apparent 31 unique diary names to 29 true participants. Four anonymous diary entries were excluded from participant-level analyses.

2.5 Ethical Considerations

This study involved voluntary participation of healthcare professionals and medical students who provided self-reported feedback through electronic questionnaires and case report forms. No patients were involved, and no patient data were collected or processed. Given the non-interventional design and minimal-risk nature of the study, formal ethics committee approval was not required under applicable Portuguese regulations governing quality improvement evaluations (Lei n.º 21/2014) [19]. All participants provided informed consent prior to enrolment. Participation was voluntary, and participants could withdraw at any time without consequence. No incentives were offered.

3 Results

A participant flow diagram (Figure 1) details the progression from enrolment through each data collection phase.

3.1 Participant Characteristics

Of the 29 enrolled participants, 26 provided role information. The sample comprised Attending Physicians (46%), Residents (23%), Medical Students (15%), Consultants (12%), and one Head of Service (4%). The distribution of years of practice is shown in Table 1. Represented specialties included Family Medicine (35%), Intensive Care Medicine (23%), Internal Medicine (12%), Surgery, Cardiology, Anaesthesiology, and Emergency Medicine, with the remaining 12% being medical students.

Table 1: Participant demographics and experience distribution (N = 26 participants with baseline data; three participants did not complete the baseline eCRF)

Years of Practice (Category)	**n**	**Percentage**
<1 year	4	15%
1 to 3 years	5	19%
3 to 10 years	7	27%
>10 years	10	38%
Missing	*3*	–

Baseline data from the first eCRF indicated that participants were already regular users of clinical decision support tools: 36% reported daily use, 60% several times per

week, and 4% once per week. No participant reported using such tools less than once a month.

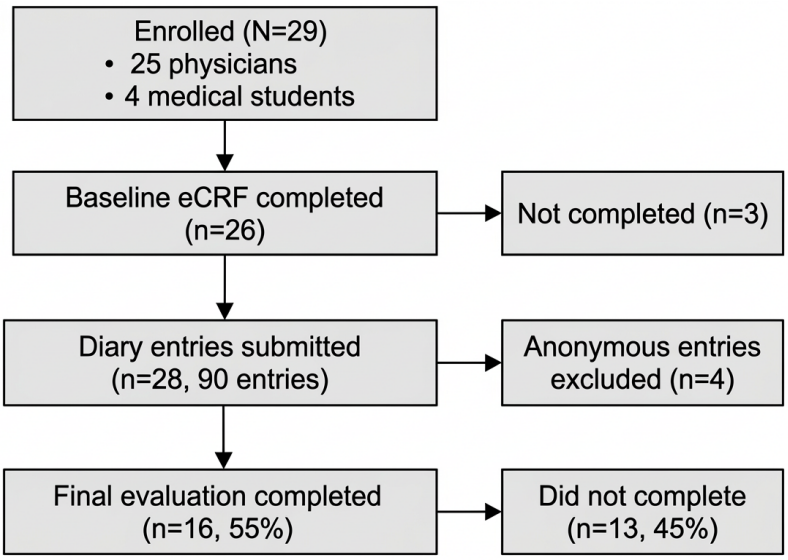

Figure 1: Participant flow diagram. Of the 29 enrolled clinicians (25 physicians and 4 medical students), 26 completed the baseline eCRF, 28 submitted at least one daily diary entry (90 named entries total, with 4 anonymous entries excluded), and 16 (55%) completed the final NPS evaluation. Three participants did not complete the baseline form and 13 did not complete the final evaluation. eCRF, electronic Case Report Form; NPS, Net Promoter Score.

3.2 Attrition Analysis

Of the 29 enrolled participants, 16 (55%) completed the final eCRF. To assess whether this attrition introduced systematic bias, we compared diary scores between final form completers (n = 15) and non-completers (n = 13) among the 28 participants who submitted at least one diary entry. As shown in Table 2 and Figure 2, there was no significant difference between the two groups on either metric (Mann-Whitney U; p = 0.672 for time saving, p = 0.886 for decision support).

Table 2: Attrition analysis: comparison of diary scores between final eCRF completers and non-completers (N = 28). SD, standard deviation; U, Mann-Whitney U statistic; p, p-value. Groups compared using the Mann-Whitney U test.

Metric	**Completers (n = 15)**	**Non-completers (n = 13)**	**U**	**p**
Time Saving (mean ± SD)	4.40 ± 0.50	4.12 ± 0.99	107.0	0.672
Decision Support (mean ± SD)	4.19 ± 0.76	4.12 ± 0.79	101.0	0.886

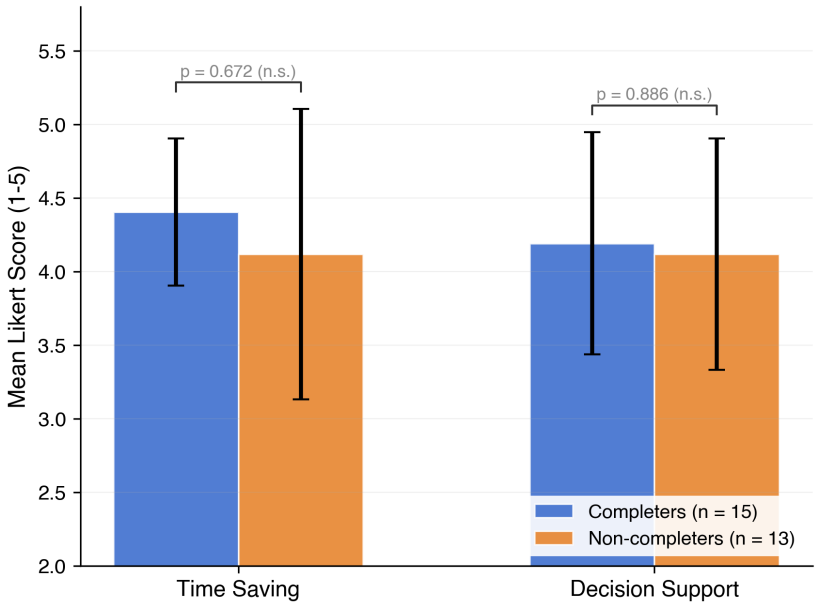

Figure 2: Comparison of mean diary scores between final eCRF completers and non-completers. Error bars represent standard deviation. No statistically significant difference was observed (p = 0.672 and p = 0.886).

3.3 Perceived Time Savings and Clinical Reasoning Support

The participant-level mean score for perceived time saving was 4.27 out of 5 (SD = 0.76; 95% CI: 3.97–4.57), with 87.8% of diary entries indicating agreement or strong agreement.

Perceived clinical decision support showed comparable results, with a mean of 4.16 (SD = 0.76; 95% CI: 3.86–4.45) and 85.6% agreement that the tool supported diagnostic or therapeutic reasoning.

Table 3: Summary of primary outcome metrics with confidence intervals (N = 29). SD, standard deviation; CI, confidence interval. Bootstrap CI computed using percentile resampling (10,000 iterations) at the participant level. Scores measured on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree).

Metric	Mean	SD	95% CI (Bootstrap)	95% CI (t based)
Time Saving	4.27	0.76	3.98–4.53	3.97–4.57
Decision Support	4.16	0.76	3.87–4.43	3.86–4.45

3.4 Longitudinal Stability

No significant temporal trend was detected across the five study days (Figure 3). Among the 10 participants who completed all five days, the Friedman test showed no significant change over time for either time saving (χ^2 = 2.43, p = 0.657) or decision support (χ^2 = 3.60, p = 0.463). These non-significant results are consistent with sustained utility,

although the small sample (n = 10) provides limited power to detect moderate temporal effects.

Table 4: Day by day scores across the study period. n, number of participants with diary entries on that day; SD, standard deviation; SEM, standard error of the mean. Friedman test (n = 10 participants with all five days): time saving $\chi^2 = 2.43$, p = 0.657; decision support $\chi^2 = 3.60$, p = 0.463.

Day	n	Time Mean ± SD	SEM	Decision Mean ± SD	SEM
1	27	4.26 ± 1.02	0.20	4.07 ± 1.17	0.23
2	20	4.10 ± 1.17	0.26	3.85 ± 1.31	0.29
3	17	4.06 ± 1.14	0.28	4.06 ± 1.25	0.30
4	14	4.36 ± 0.50	0.13	4.14 ± 0.53	0.14
5	12	4.33 ± 0.65	0.19	4.25 ± 0.62	0.18

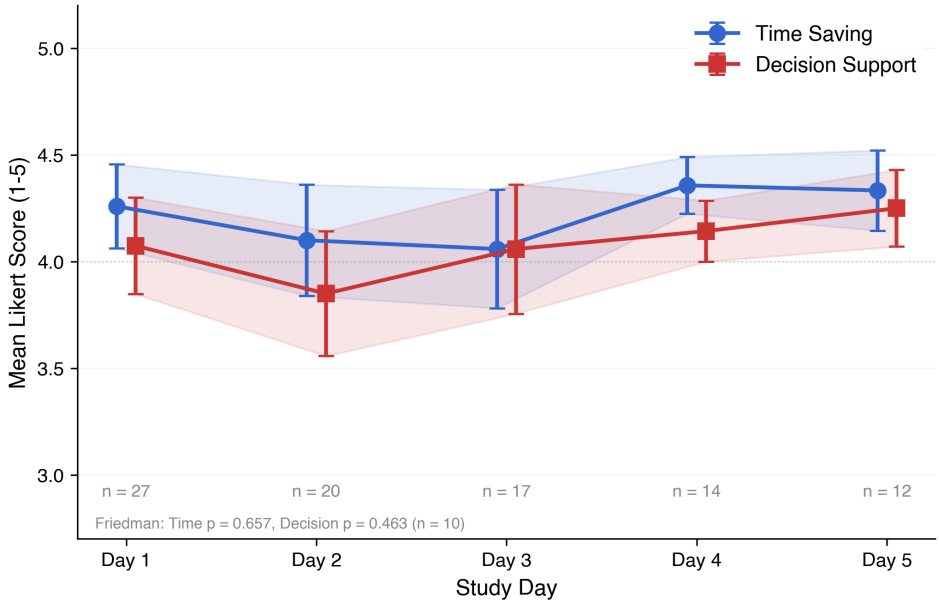

Figure 3: Mean Likert scores for time saving and decision support across the five study days. Error bars represent standard error of the mean. The Friedman test showed no significant change (p = 0.657 for time saving; p = 0.463 for decision support).

Figure 4 presents individual daily Likert scores for all 28 participants. The predominance of high scores across the heatmap is consistent with the aggregate findings. Grey cells denote days on which no diary entry was submitted.

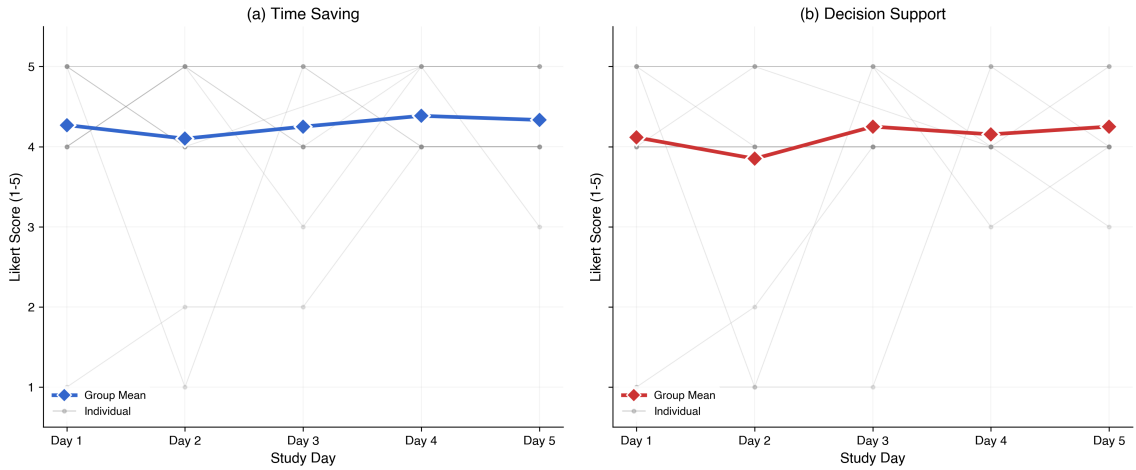

Figure 4: Heatmap of individual daily Likert scores for time saving (left) and decision support (right). Rows represent participants (sorted by mean score, highest at top); columns represent study days. Green indicates agreement (4–5), yellow neutral (3), and red disagreement (1–2). Grey cells denote missing diary entries.

3.5 Variable Analysis and Correlations

Associations between participant characteristics and perceived utility were examined using Spearman's rank correlation coefficient. An inverse correlation between years of practice and perceived decision support did not reach statistical significance ($\rho = -0.381$, $p = 0.055$, $n = 26$); however, the moderate effect size suggests a trend warranting investigation in larger samples. The corresponding correlation with time saving was weaker ($\rho = -0.182$, $p = 0.375$). Junior physicians and students reported higher perceived benefit, while senior clinicians used the tool primarily for drug dosing verification and differential diagnosis confirmation. These associations are illustrated in Figure 5.

No association was found between self-reported technology comfort and perceived time saving ($\rho = 0.060$, $p = 0.771$, $n = 26$), suggesting that the interface was accessible regardless of prior technology proficiency.

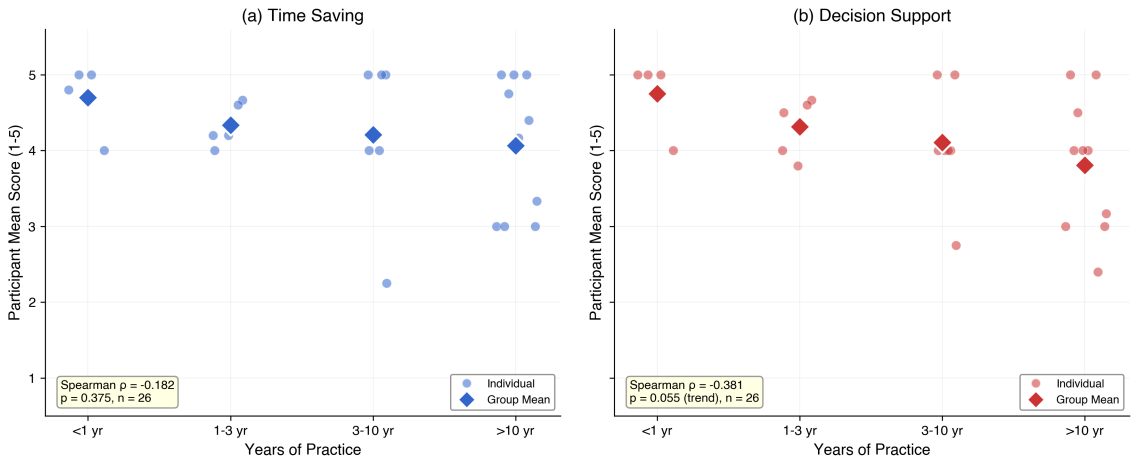

Figure 5: Scatter plots showing the relationship between years of practice and perceived value. Spearman correlation coefficients are annotated. The association with decision support did not reach significance ($\rho = -0.381$, p = 0.055).

3.6 Net Promoter Score and General Satisfaction

Among the 16 participants who completed the final evaluation, DR. INFO achieved an NPS of 81.2 (95% CI: 62.5–100.0), with 81% classified as Promoters (scores 9 to 10), 19% as Passives (scores 7 to 8), and no Detractors (scores 0 to 6) (Figure 6). For context, industry benchmarks place the average NPS for healthcare and technology sectors between 30 and 50 [20].

However, given that only 55% of participants completed the final form, we conducted a sensitivity analysis to assess the robustness of this finding. The 13 non-respondents with diary data were imputed under three increasingly conservative scenarios (Table 5).

Table 5: NPS sensitivity analysis under different assumptions for non-respondents (n = 13 with diary data). NPS, Net Promoter Score. Observed NPS based on 16 respondents who completed the final eCRF.

Scenario	Assumed Non-responder Score	Resulting NPS
Observed (n = 16 only)	–	81.2
All non-responders Passive	7 (Passive range)	44.8
All non-responders Detractor	0–6 (any Detractor score)*	0.0

*NPS classifies scores 0–6 as Detractors. Because NPS is calculated as %Promoters − %Detractors (without weighting by score magnitude), all Detractor-range imputations (whether 0 or 5) yield the same NPS.

The attrition analysis (see Section 3.2) mitigates the concern that non-respondents were systematically less satisfied, supporting the interpretation that the observed NPS reflects genuine positive sentiment. Under the conservative assumption that all non-respondents were Passives, the NPS would remain positive at 44.8.

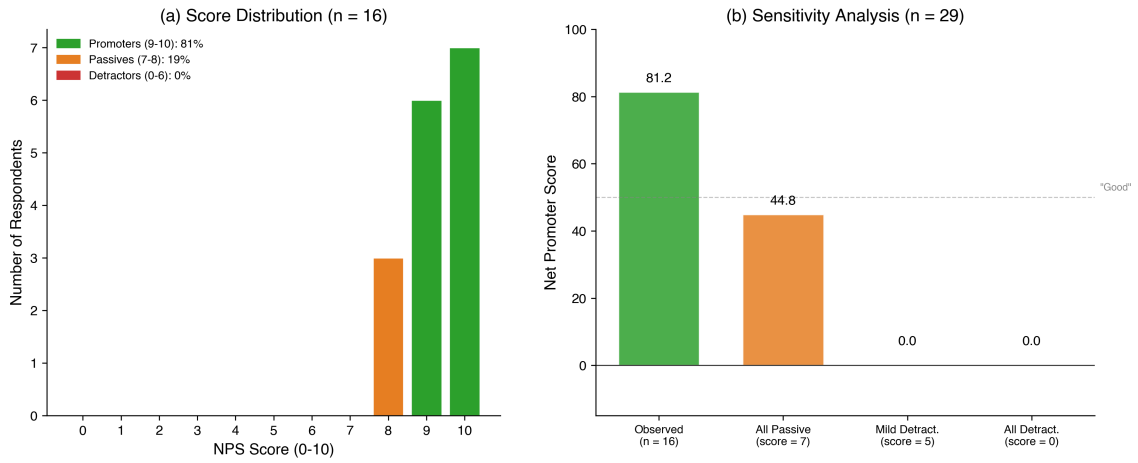

Figure 6: (a) Distribution of NPS scores among the 16 respondents. No detractors were identified. (b) Sensitivity analysis showing NPS under different imputation scenarios for the 13 non-respondents.

3.7 Subgroup Analysis by Clinical Role

Evaluations were consistently positive across career stages, with higher mean scores among students and residents (Table 6, Figure 7). The Kruskal-Wallis test was not significant for either time saving ($H = 1.79$, $p = 0.774$) or decision support ($H = 7.32$, $p = 0.120$), likely reflecting limited statistical power given the small subgroup sizes.

The single Head of Service participant rated decision support lower (2.40/5) than all other groups while rating time saving comparably (4.40/5). This pattern may reflect the different needs of senior departmental leaders, for whom general clinical reasoning adds less incremental value over established expertise, whereas time efficiency in information retrieval remains relevant. As this represents a single participant, no generalisable inference can be drawn.

Table 6: Subgroup analysis of perceived value by clinical role ($N = 26$). Values reported as mean ± SD. Time saving and decision support scored on 5-point Likert scales (1 = strongly disagree, 5 = strongly agree). Kruskal-Wallis H test: time saving $H = 1.79$, $p = 0.774$; decision support $H = 7.32$, $p = 0.120$.

Role	n	Time Saving	Decision Support	Primary Use Cases
Medical Student	4	4.70 ± 0.48	4.75 ± 0.50	Individual study, case preparation
Resident	6	4.44 ± 0.37	4.43 ± 0.45	Guidelines, drug dosing, differentials
Attending Physician	12	4.03 ± 0.99	3.94 ± 0.72	Complex cases, multimorbidity
Consultant	3	4.11 ± 0.84	4.06 ± 0.92	Teaching, case review, infographics
Head of Service	1	4.40	2.40	Decision support, bibliography

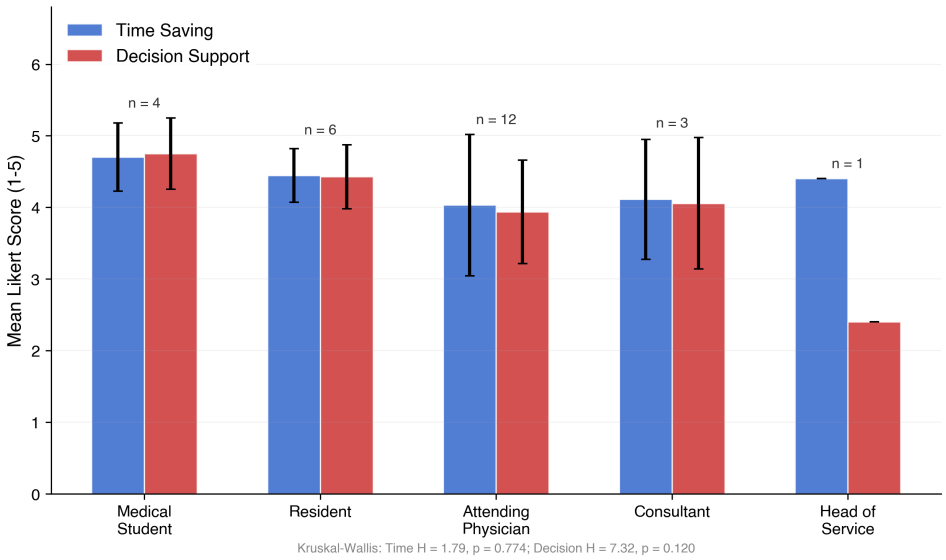

Figure 7: Mean Likert scores by clinical role. Error bars represent standard deviation. Students and residents reported the highest perceived value. Kruskal-Wallis test: time saving H = 1.79, p = 0.774; decision support H = 7.32, p = 0.120.

3.8 Clinical Use Case Patterns

Reported use cases spanned a broad range of clinical scenarios (Figure 8). The most frequently selected categories were treatment advice (n = 47), disease details and differential diagnosis (n = 46), medication details (n = 36), diagnosis (n = 36), and individual study (n = 31), the latter primarily reported by medical students and residents.

These patterns indicate that perceived value was not confined to a single clinical context but extended across outpatient and acute hospital settings.

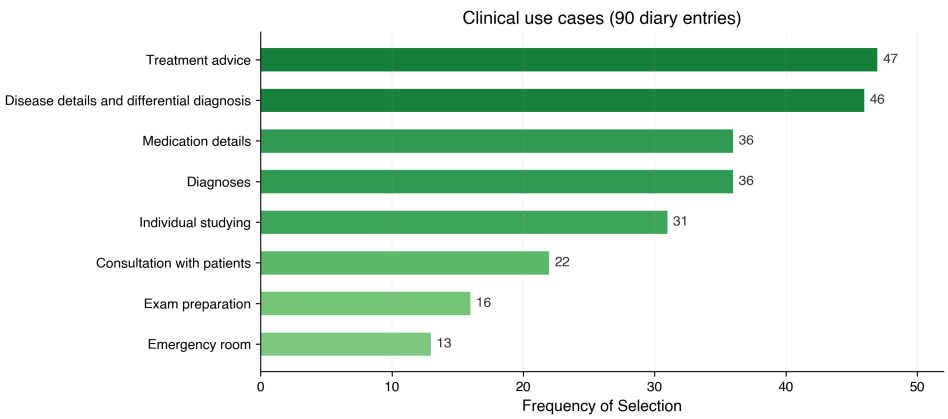

Figure 8: Frequency of clinical use case categories reported by participants across all diary entries.

3.9 Qualitative Feedback Analysis

Content analysis of the 16 open-ended responses regarding areas for improvement identified five recurrent themes (Figure 9). The most frequently cited concern was response latency (50%), followed by insufficient specificity of answers (38%) and limitations in understanding nuanced clinical queries (25%).

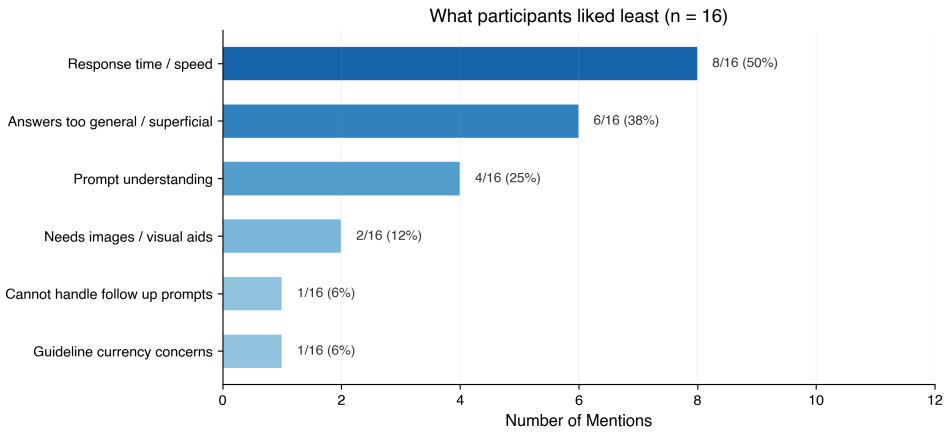

Figure 9: Thematic analysis of participant feedback on areas for improvement. Response time and answer specificity were the most frequently cited concerns.

4 Discussion

Several features of the study population provide important context for interpreting these findings.

First, 96% of the 26 participants with baseline data were already frequent users of established clinical decision support tools (e.g., UpToDate, Amboss, BMJ Best Practice). The high ratings therefore came from clinicians accustomed to such resources, not from users encountering AI-assisted tools for the first time.

Second, the mean time-saving score of 4.27/5 (95% CI: 3.97–4.57) was reported by clinicians who already had established workflows with comparable tools. This study did not directly compare DR. INFO against specific alternatives; the ratings reflect perceived utility rather than measured incremental performance. A recent evaluation of OpenEvidence, another AI clinical decision support platform, reported physician satisfaction scores of 3.60 $\pm$ 0.38 out of 4 across common primary care conditions, with the platform primarily reinforcing rather than modifying clinical plans [21].

Third, no participant reported using clinical decision support tools less than weekly, indicating that the cohort was well positioned to evaluate whether an AI clinical assistant adds practical value to an already digitally supported workflow.

Technology comfort was not associated with perceived time saving (Spearman ρ = 0.060, p = 0.771). The temporal stability of satisfaction scores across the five study

days (Friedman test: time saving $p = 0.657$; decision support $p = 0.463$) suggests that perceived benefit was not limited to initial exposure, although low statistical power ($n = 10$) limits this inference.

Response latency was the most frequently cited concern (50% of respondents), a trade-off inherent to multi-step retrieval and reasoning architectures. Subsequent development iterations have targeted response time reduction.

Perceived value was observed across diverse clinical contexts, from chronic disease management in primary care to acute hospital scenarios. A scoping review of AI-based clinical decision support systems in primary care found promising outcomes for clinical management and workload reduction, though implementation remained influenced by physician perceptions and cultural context [22]. The Kruskal-Wallis test showed no significant difference in scores across clinical roles (time saving $H = 1.79$, $p = 0.774$; decision support $H = 7.32$, $p = 0.120$), and the distribution of participant-level means (IQR 4.0–5.0 for both metrics) confirmed that positive ratings were broadly distributed rather than driven by a small number of outliers. Together with the absence of an association between technology comfort and perceived utility ($\rho = 0.060$, $p = 0.771$), these findings suggest that the tool was accessible and useful across career stages and technical backgrounds.

The single Head of Service participant rated decision support notably lower (2.40/5) than time saving (4.40/5), a pattern that may reflect the different information needs of senior departmental leaders for whom general clinical reasoning adds less incremental value over established expertise.

Beyond latency, 38% of respondents cited insufficient specificity of answers and 25% noted limitations in understanding nuanced clinical queries. These areas have been prioritised for subsequent development iterations.

4.1 Attrition and Potential Bias

With only 55% of participants completing the final evaluation, selective dropout was a plausible threat to validity. The attrition analysis showed that completers and non-completers did not differ significantly in their diary scores ($p = 0.672$ and $p = 0.886$). Combined with the NPS sensitivity analysis (NPS = 44.8 under the conservative assumption that all non-respondents were Passives), these findings mitigate the concern of survivorship bias, although equivalence cannot be established from a non-significant comparison at this sample size.

4.2 Broader Implications: Time Efficiency and Cognitive Load

Although this study did not directly measure burnout or administrative workload, the high proportion of entries reporting perceived time efficiency (87.8%) may have implications for clinical workflow. Physician burnout has been consistently linked to administrative

burden and cognitive overload [2, 3], and emerging evidence suggests that AI-based clinical assistants may help reduce information-seeking time [23]. Causal relationships cannot be inferred from the present pilot, but sustained perceptions of time efficiency warrant further investigation.

The inverse correlation between seniority and perceived decision support ($\rho = -0.381$, $p = 0.055$), although not statistically significant, is consistent with the hypothesis that the tool may be particularly useful in medical education and residency training, where rapid access to guideline-based information directly supports learning. A systematic review encompassing 60 studies across 39 countries found that over 60% of physicians and medical students reported positive attitudes toward clinical AI, with younger clinicians generally expressing greater openness [24]. For senior physicians, the perceived value appears to centre on verification and rapid cross-referencing against current evidence.

5 Limitations

Several limitations should be acknowledged. First, the study did not include independent verification of the clinical accuracy of DR. INFO's responses or any patient outcome measures. All outcome measures reflect physician perceptions rather than adjudicated correctness. While the companion HealthBench evaluation [14] provides system-level accuracy benchmarking, expert review of actual clinical queries from the pilot would have strengthened the safety assessment. Future studies should incorporate clinician-adjudicated accuracy audits alongside perception-based outcomes.

Second, the use of 5-point Likert scales, combined with high baseline satisfaction, may have produced ceiling effects that limit the ability to distinguish between degrees of positive perception. Future studies should consider wider response scales.

Third, all participants were recruited from Portuguese healthcare institutions, which may limit the generalisability of findings to other healthcare systems and cultural contexts.

6 Conclusion

In this prospective pilot study of 29 clinicians, DR. INFO was rated favourably for both time efficiency (mean 4.27/5; 95% CI: 3.97–4.57) and clinical decision support (mean 4.16/5; 95% CI: 3.86–4.45), with scores stable across five study days and a Net Promoter Score of 81.2 with no detractors. These ratings came from clinicians who were already regular users of established clinical decision support tools, reinforcing the credibility of the findings. Attrition analysis confirmed that non-completers did not differ from completers in their diary scores, and the NPS remained positive under conservative imputation.

Positive ratings were consistent across clinical roles and independent of prior technology proficiency, with a trend toward higher perceived decision support among junior clini-

cians suggesting potential relevance for medical education. Participant feedback identified response latency and answer specificity as the primary areas for improvement. Larger, controlled studies incorporating objective performance metrics and multicentre designs are warranted to determine whether these perceived benefits translate to measurable clinical impact.

7 Acknowledgements

We gratefully acknowledge the technical support provided by Sonu Kumar and Kritika Singh. We also thank all participating clinicians for their time and candid feedback throughout the study.

8 Ethics and Consent

This study was conducted as a prospective quality improvement evaluation of an AI clinical assistant already deployed in routine clinical practice. It involved only voluntary, self-reported feedback from clinicians and did not involve patients, patient data, or any experimental intervention. Formal ethics committee approval was therefore not required under applicable Portuguese regulations governing quality improvement evaluations (Lei n.º 21/2014). All participants provided written informed consent prior to enrolment. Participation was voluntary, and participants could withdraw at any time without consequence. No patient-identifiable data were collected or processed.

9 Conflicts of Interest

All authors (Rogerio Corga Da Silva, Miguel Romano, Tiago Mendes, Marta Isidoro, Sandhanakrishnan Ravichandran, Shivesh Kumar, Michiel van der Heijden, Olivier Fail, and Valentine Emmanuel Gnanapragasam) are employed by Synduct GmbH, the developer of DR. INFO. To mitigate potential bias: (a) all data collection instruments used validated Likert scales and standardised Net Promoter Score methodology; (b) all statistical analyses are reproducible from the provided dataset; and (c) statistical analyses were independently verified for reproducibility. Participating clinicians were volunteers and received no compensation for their involvement.

10 Funding

This work was supported by Synduct GmbH.

References

[1] Siru Liu, Allison B. McCoy, and Adam Wright. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *Journal of the American Medical Informatics Association*, 32(4):605–615, 2025. doi: 10.1093/jamia/ocaf008.

[2] Sara Berg. Doctors work fewer hours, but the EHR still follows them home. American Medical Association, August 2025. URL https://www.ama-assn.org/practice-management/physician-health/doctors-work-fewer-hours-ehr-still-follows-them-home. Accessed: 2026-03-23.

[3] Tait D. Shanafelt, Colin P. West, Christine Sinsky, Mickey Trockel, Michael Tutty, Hanhan Wang, Lindsey E. Carlasare, and Liselotte N. Dyrbye. Changes in burnout and satisfaction with work–life integration in physicians and the general US working population between 2011 and 2023. *Mayo Clinic Proceedings*, 100(7):1142–1158, 2025. doi: 10.1016/j.mayocp.2024.11.031.

[4] Stanford HAI. Stanford develops real-world benchmarks for healthcare AI agents, September 2025. URL https://hai.stanford.edu/news/stanford-develops-real-world-benchmarks-for-healthcare-ai-agents. Accessed: 2026-03-23.

[5] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachber, Philip Mansfield, Sushant Prakash, Bradley Green, Avinatan Hassidim, Sara Mahdavi, Greg S. Corrado, Yossi Matias, Katherine Chou, David Fleet, Laurent El Shafey, Shekoofeh Virmani, Vivek Natarajan, Shravya Shetty, Joelle Barral, Jim Soraker, Greg Corrado, Lisa Tsai, Shekoofeh Azizi, Nikolaos Flemotomos, Ankit Singhal, Khaled Ayoub, Quoc V. Le, and Alan Karthikesalingam. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025. doi: 10.1038/s41591-024-03423-7.

[6] John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J. Faix, Aaron M. Goodman, Christopher A. Longhurst, Michael Hogarth, and Davey M. Smith. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6):589–596, 2023. doi: 10.1001/jamainternmed.2023.1838.

[7] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940, 2023. doi: 10.1038/s41591-023-02448-8.

[8] Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Jaden A. Cool, Zahir Kanjee, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, and Jonathan H. Chen. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Network Open*, 7(10):e2440969, 2024. doi: 10.1001/jamanetworkopen.2024.40969.

[9] Arya Rao, Michael Pang, John Kim, Meghna Kamineni, Winston Lie, Anoop K. Prasad, Adam Landman, Keith Dreyer, and Marc D. Succi. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. *Journal of Medical Internet Research*, 25:e48659, 2023. doi: 10.2196/48659.

[10] Mahmud Omar, Reem Agbareia, Benjamin S. Glicksberg, Girish N. Nadkarni, and Eyal Klang. Benchmarking the confidence of large language models in answering clinical questions: cross-sectional evaluation study. *JMIR Medical Informatics*, 13: e66917, 2025. doi: 10.2196/66917.

[11] COMPETE 2030. Portugal lidera saúde digital na união europeia, December 2024. URL https://www.compete2030.gov.pt/comunicacao/portugal-lidera-saude-digital-na-uniao-europeia/. Accessed: 2026-03-23.

[12] Serviços Partilhados do Ministério da Saúde (SPMS), EPE. Inteligência artificial na saúde em Portugal: regulamentação, impactos e perspetivas de futuro. White paper, Ministério da Saúde, Portugal, February 2025. URL https://www.spms.min-saude.pt/wp-content/uploads/2025/03/White-Paper_Inteligencia-Artificial-na-Saude-em-Portugal_-Final2-1.pdf. Accessed: 2026-03-23.

[13] OpenAI. Introducing HealthBench, May 2025. URL https://openai.com/index/healthbench/. Accessed: 2026-03-23.

[14] Sandhanakrishnan Ravichandran, Shivesh Kumar, Rogerio Corga Da Silva, Miguel Romano, Michiel van der Heijden, Olivier Fail, and Valentine Emmanuel Gnanapragasam. OpenAI's HealthBench in action: evaluating an LLM-based medical assistant on realistic clinical queries. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2509.02594. URL https://arxiv.org/abs/2509.02594.

[15] Lehana Thabane, Jinhui Ma, Rong Chu, Ji Cheng, Afisi Ismaila, Lorena P. Rios, Reid Robson, Marroon Thabane, Lora Giangregorio, and Charles H. Goldsmith. A tutorial on pilot studies: the what, why and how. *BMC Medical Research Methodology*, 10:1, 2010. doi: 10.1186/1471-2288-10-1.

[16] Gail M. Sullivan and Anthony R. Artino, Jr. Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education*, 5(4):541–542, 2013. doi: 10.4300/JGME-5-4-18.

[17] Frederick F. Reichheld. The one number you need to grow. *Harvard Business Review*, 81(12):46–54, 2003.

[18] Sidney Siegel and N. John Castellan, Jr. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, 2nd edition, 1988. doi: 10.1086/416341.

[19] Assembleia da República. Lei n.º 21/2014 de 16 de abril: Aprova a lei da investigação clínica, 2014. URL https://diariodarepublica.pt/dr/detalhe/lei/21-2014-25344024. Accessed: 2026-03-23.

[20] Laura Kurasińska. NPS benchmarks for 2025: good net promoter scores by industry. Survicate, December 2025. URL https://survicate.com/nps-benchmarks/. Accessed: 2026-03-23.

[21] Ryan T. Hurt, Curtiss R. Stephenson, Emily A. Gilman, Caroline A. Aakre, Ivana T. Croghan, Manpreet S. Mundi, Kalyani Ghosh, and Jithinraj Edakkanambeth Varayil. The use of an artificial intelligence platform OpenEvidence to augment clinical decision-making for primary care physicians. *Journal of Primary Care & Community Health*, 16:21501319251332215, 2025. doi: 10.1177/21501319251332215.

[22] Cesar A. Gomez-Cabello, Sahar Borna, Sophia Pressman, Syed Ali Haider, Clifton R. Haider, and Antonio J. Forte. Artificial-intelligence-based clinical decision support systems in primary care: a scoping review of current clinical implementations. *European Journal of Investigation in Health, Psychology and Education*, 14(3):685–698, 2024. doi: 10.3390/ejihpe14030045.

[23] Kevin B. O'Reilly. How much can ambient AI scribes help cut doctor burnout? American Medical Association, October 2025. URL https://www.ama-assn.org/practice-management/physician-health/how-much-can-ambient-ai-scribes-help-cut-doctor-burnout. Accessed: 2026-03-23.

[24] Mingyang Chen, Bo Zhang, Ziting Cai, Samuel Seery, Maria J. Gonzalez, Nasra M. Ali, Ran Ren, Youlin Qiao, Peng Xue, and Yu Jiang. Acceptance of clinical artificial intelligence among physicians and medical students: a systematic review with cross-sectional survey. *Frontiers in Medicine*, 9:990604, 2022. doi: 10.3389/fmed.2022.990604.