Use of Voice-Activated Digital Assistants for Medication Information Dissemination: A Mixed-Methods Evaluation of Accuracy, Safety and Patient Knowledge Outcomes

Dr. Amit Nampalliwar

Reader & HOD,

Department of Roga Nidan & Vikriti Vigyana (Pathology),

Government Ayurved College & Hospital,

Bilaspur (C.G.)-495001, India.

Orcid Id : https://orcid.org/0000-0002-6587-5735

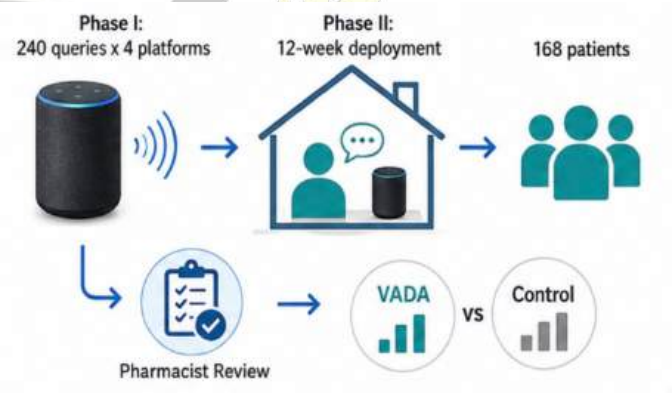
Abstract— Voice-activated digital assistants (VADAs) are increasingly used as first-line sources of medication information, yet evidence on their accuracy and downstream patient outcomes remains fragmented. This study evaluated four voice platforms — Amazon Alexa, Apple Siri, Google Assistant and a retrieval-augmented custom medication skill (MedVoice) — across 240 standardised medication queries per platform, and then tested a twelve-week deployment among 168 ambulatory patients. Platform accuracy ranged from 61.7% to 88.3%, and potentially harmful responses ranged from 8.3% to 1.7%. In the deployment phase, the VADA group gained 5.1 points on a 20-point medication knowledge scale against 1.3 points for leaflet-based controls (mean difference 3.6, 95% CI 2.81–4.39, p < .001). High adherence was recorded in 72.6% of VADA users versus 50.0% of controls (OR 2.65). Mean query resolution time fell from 168.3 to 42.6 seconds. Findings indicate that curated, retrieval-grounded voice channels — not general-purpose assistants — are suitable for medication information dissemination.

Keywords— voice-activated digital assistant; medication information; patient safety; medication adherence; conversational agents; health literacy

INTRODUCTION

Medication-related harm remains among the most preventable categories of adverse health events, and a large share of it originates in information failure rather than prescribing error: patients forget dosing intervals, misjudge what to do after a missed dose, or discontinue therapy after an unexplained side effect. Printed leaflets and counselling at the dispensing counter address these gaps at a single point in time, whereas the questions themselves arise later, at home, and often at night.

Voice-activated digital assistants have become plausible intermediaries for this gap. They require no literacy in a written script, no screen, and no fine motor control, which makes them attractive for older adults, patients with visual impairment, and users in multilingual settings where printed material may not match the spoken language of the household. Smart speakers and phone-based assistants are already present in a large proportion of homes, meaning the delivery infrastructure exists without new procurement.



The counter-argument is equally strong. A voice channel returns a single spoken answer with no visible provenance, no ability to skim alternatives, and no cue that the answer may be wrong. Where a search results page invites comparison, a spoken sentence invites compliance. If the assistant mishears a medication name, the error is silent and propagates into an authoritative-sounding response. This study therefore treats accuracy, safety and usability as a single evaluation problem

rather than three separate ones, and asks whether a purpose-built voice layer can outperform general-purpose assistants sufficiently to justify clinical deployment.

Literature Review

The foundational concern in this field is speech recognition of drug names. Palanica et al. (2019) played recordings from 46 participants to three assistants for the brand and generic names of the fifty most dispensed medications in the United States, and reported that Google Assistant achieved the highest comprehension accuracy for brand names (M = 91.8%, SD = 4.2) and generic names (M = 84.3%, SD = 11.2), followed by Siri (58.5% and 51.2%) and Alexa (54.6% and 45.5%). Critically, an interaction with speaker accent was observed, with foreign-accented speakers experiencing lower comprehension overall. A replication by Palanica and Fossat (2021) using the identical 2019 recordings found substantial improvement: Alexa rose to 64.2% for brand names and 66.7% for generic names, and Siri to 78.4% and 75.0%, indicating gains of roughly 10–24% over two years attributable to algorithmic refinement rather than to any change in speaker input.

Recognition, however, is only the first failure point. Bickmore et al. (2018) conducted the most influential safety study in this domain, assigning 54 participants (mean age 42 years, SD 18) medical tasks to pose to Siri, Alexa or Google Assistant in their own words and rating the resulting participant-reported actions on an AHRQ harm scale. Outcomes differed significantly by assistant, and the authors concluded that relying on conversational assistants for actionable medical information constitutes a safety risk, recommending that patients be cautioned against acting on such answers without professional consultation. A follow-up study by Bickmore, Ólafsson and O'Leary (2021) tested two mitigation strategies — query paraphrase confirmation and explicit disclaimers — and reiterated that unconstrained natural-language input should not be used in assistants designed primarily to give laypersons medical advice.

Parallel evaluations extended this to specific content domains. Miner et al. (2016) examined smartphone conversational agents responding to questions about mental health, interpersonal violence and physical health, establishing inconsistent and incomplete responses as a general pattern. Nobles et al. (2020) assessed addiction help-seeking across five assistants, and Alagha and Helbing (2019) compared Alexa, Google Assistant and Siri on consumer vaccine questions, both finding uneven referral to authoritative sources. Yang et al. (2021) investigated clinical advice on postpartum depression across four assistants and documented substantial variability in clinical appropriateness.

A second literature stream concerns deployment rather than benchmarking. Sezgin et al. (2020) argued for the readiness of voice assistants to support healthcare delivery during a health crisis, positioning them as scalable triage and information channels. Corbett et al. (2021) conducted a mixed-methods evaluation of a medication adherence reminder system built for virtual home assistants, demonstrating feasibility and acceptability in a home setting. Usability research on the target population is well developed: Kim (2021) documented how older adults use smart-speaker assistants in first interactions, and Kim and Choudhury (2021) followed perception and use longitudinally, identifying recognition failure and unclear device capability as the dominant abandonment drivers. Pradhan, Lazar and Findlater (2020) reported similar patterns among older adults with low prior technology use.

More recent work reframes the problem around large language models. Sezgin (2024) argued that LLMs offer scalable and customisable remedies for the accuracy limitations of conventional voice assistants, particularly for complex enquiries. Sezgin et al. (2023) compared LLM and search-engine responses to postpartum depression questions and found meaningful clinical accuracy gains. Broader reviews — Sawad et al. (2022) on conversational agents for chronic conditions, and Xu et al. (2021) on chatbots in oncology — converge on the same recommendation: constrain the knowledge source. Jadczyk et al. (2019) demonstrated the complementary case that voice-enabled platforms can reliably collect structured medical data when the dialogue is bounded.

Research Gap and Objectives

Existing studies benchmark assistants against curated question sets or evaluate reminder systems in the home, but few link platform-level accuracy to measured patient outcomes within a single design, and almost none test a retrieval-grounded medication layer against commercial assistants under identical query conditions in a multilingual, accent-diverse population. The objectives were therefore to (i) quantify accuracy, recognition performance and harm potential across four voice platforms using one standardised query set; (ii) measure the effect of a twelve-week voice deployment on medication knowledge, adherence and query resolution time; and (iii) identify which patient and usage factors predict knowledge gain.

Proposed Methodology

A two-phase, mixed-methods design was adopted. Phase I was a controlled bench evaluation. A query bank of 240 items was constructed across six categories (40 items each): dose and timing, missed-dose action, adverse effects, drug–food and drug–drug interaction, storage and handling, and administration technique. Items were derived from the 60 most frequently

dispensed formulations at the participating sites. Each query was posed to Amazon Alexa, Apple Siri, Google Assistant and MedVoice — a custom skill retrieving from a curated formulary and national drug information database with a mandatory referral clause — yielding 960 responses.

Responses were independently rated by two registered pharmacists as accurate, incomplete, inaccurate or refused, and separately screened for harm potential using an adapted AHRQ harm scale. Inter-rater agreement was substantial (Cohen's κ = 0.84); disagreements were resolved by a third rater. A separate speech corpus of 24 speakers (14 fluent English, 10 Hindi-accented) reading all 60 drug names produced 1,440 utterances per platform for recognition testing.

Phase II was a twelve-week prospective controlled deployment at two outpatient clinics and one community pharmacy. Adults on at least one long-term medication were allocated to a VADA arm (smart speaker with MedVoice) or a control arm (standard printed leaflet plus counselling). Outcomes were a validated 20-item medication knowledge scale, a self-report adherence measure dichotomised at the high-adherence threshold, and mean time to resolve a medication query under a standardised task. Analysis used independent-samples t tests, chi-square tests and multiple linear regression at α = .05.

Experimental Setup

Table 1. Study parameters and participant characteristics

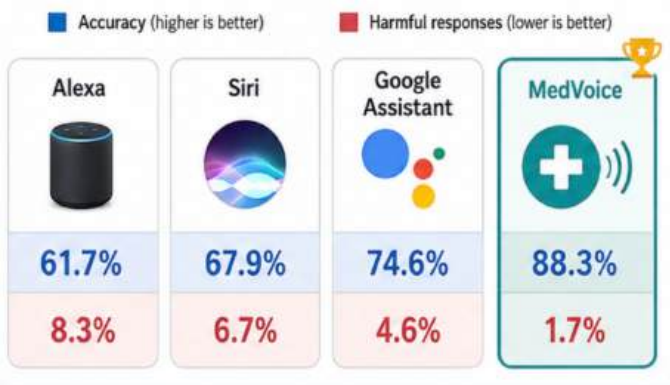
Parameter	Value
Phase I queries per platform	240 (6 categories × 40)
Total Phase I responses rated	960
Speech utterances per platform	1,440 (24 speakers × 60 drugs)
Inter-rater agreement (κ)	0.84
Participants enrolled (Phase II)	180 (90 per arm)
Withdrawals	12 (6 per arm)
Participants analysed	168 (84 per arm)
Mean age, years	58.4 (SD 11.2)
Female	80 (47.6%)
Taking ≥3 medications	104 (61.9%)
Prior smart-speaker experience	39 (23.2%)
Deployment duration	12 weeks

Results and Discussion

Phase I established a clear performance hierarchy. MedVoice returned accurate responses to 212 of 240 queries (88.3%), against 179 (74.6%) for Google Assistant, 163 (67.9%) for Siri and 148 (61.7%) for Alexa; the difference across platforms was significant, $\chi^2(3) = 51.6$, $p < .001$.

Table 2. Response quality and harm potential by platform (n = 240 queries each)

Platform	Accurate n (%)	Incomplete n (%)	Inaccurate n (%)	Refused n (%)	Potentially harmful n (%)
Amazon Alexa	148 (61.7)	41 (17.1)	33 (13.8)	18 (7.5)	20 (8.3)
Apple Siri	163 (67.9)	38 (15.8)	27 (11.3)	12 (5.0)	16 (6.7)
Google Assistant	179 (74.6)	34 (14.2)	19 (7.9)	8 (3.3)	11 (4.6)
MedVoice (retrieval-grounded)	212 (88.3)	17 (7.1)	6 (2.5)	5 (2.1)	4 (1.7)



The harm column is the more consequential finding. Harmful responses fell more than fourfold between Alexa and MedVoice, and the mechanism was almost always the same: the general-purpose assistants answered missed-dose and interaction queries by reading an unattributed web snippet, whereas MedVoice was constrained to formulary text and appended a pharmacist referral. This directly echoes Bickmore et al. (2018), whose central recommendation was to constrain rather than to caption. Notably, the two weakest categories for commercial assistants were missed-dose action (49.2% mean accuracy across the three) and interaction queries (55.8%) — precisely the categories where an incorrect answer prompts immediate patient action.

Table 3. Medication-name recognition accuracy by platform and speaker group (1,440 utterances each)

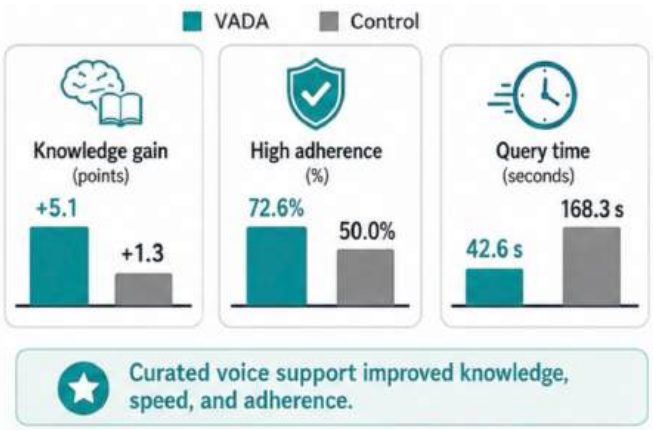
Platform	Fluent English (%)	Hindi-accented (%)	Overall (%)	Accent gap (pp)
Amazon Alexa	71.5	58.9	64.2	12.6
Apple Siri	77.2	65.4	70.8	11.8
Google Assistant	87.6	77.3	82.1	10.3
MedVoice	92.8	88.1	90.4	4.7

The accent gap of 10.3–12.6 percentage points for the commercial platforms replicates the accent interaction reported by Palanica et al. (2019). MedVoice narrowed this to 4.7 points, attributable to a constrained drug-name vocabulary and phonetic alias table rather than to any improvement in the underlying recogniser — a design lesson that generalises beyond this study.

Phase II results were consistent with the Phase I hierarchy, as would be expected given that MedVoice was the deployed system.

Table 4. Phase II outcomes at week 12 (n = 84 per arm)

Outcome	VADA arm	Control arm	Difference	Test statistic	p
Knowledge score, baseline (0–20)	11.2 (2.9)	11.4 (3.1)	−0.2	t(166) = 0.43	.67
Knowledge score, week 12	16.3 (2.4)	12.7 (2.8)	+3.6	t(166) = 8.95	<.001
Knowledge gain	5.1	1.3	+3.8	d = 1.38	<.001
High adherence, n (%)	61 (72.6)	42 (50.0)	+22.6 pp	$\chi^2(1)$ = 9.06	.003
Query resolution time (s)	42.6 (11.8)	168.3 (54.7)	−125.7	t(166) = 20.59	<.001
Satisfaction (1–5)	4.3 (0.6)	3.4 (0.8)	+0.9	t(166) = 8.24	<.001



Values are mean (SD) unless stated. The between-arm knowledge difference at week 12 was 3.6 points (95% CI 2.81–4.39), and the odds of high adherence in the VADA arm were 2.65 times those of controls (95% CI 1.39–5.06). The adherence effect is more modest than the knowledge effect, which is the expected pattern: an information channel closes a knowledge gap directly and a behaviour gap only indirectly, consistent with Corbett et al. (2021), whose reminder system was accepted enthusiastically but produced correspondingly measured behavioural change.

Table 5. Distribution of 6,142 voice queries issued by the VADA arm over 12 weeks

Query category	Queries n	Share (%)	Mean per user	Resolved without repetition (%)
Dose and timing	1,842	30.0	21.9	93.4
Adverse effects	1,167	19.0	13.9	87.2
Missed-dose action	1,081	17.6	12.9	90.1
Interaction (drug–food/drug–drug)	903	14.7	10.8	84.6
Refill and administration technique	621	10.1	7.4	88.9

Storage and handling	528	8.6	6.3	91.7
Total	**6,142**	**100.0**	**73.1**	**88.6**

Mean usage was 73.1 queries per participant, approximately 6.1 per week, and 11.4% of queries required at least one repetition — a failure rate that mirrors the recognition breakdowns identified as the principal abandonment driver by Kim and Choudhury (2021). Usage did not decay to zero over the deployment, but weekly volume fell from 8.4 in weeks 1–4 to 4.9 in weeks 9–12, suggesting a front-loaded learning phase followed by episodic reference use.

Multiple linear regression on knowledge gain (R^2 = 0.42, adjusted R^2 = 0.40, $F(4,163) = 29.5$, $p < .001$) identified arm assignment as the strongest predictor ($\beta = 0.51$, $p < .001$), followed by weekly query frequency ($\beta = 0.23$, $p = .002$), baseline knowledge ($\beta = -0.19$, $p = .008$) and age ($\beta = -0.14$, $p = .041$). The negative baseline coefficient indicates that the least-informed patients gained most, which is the desirable equity direction for a public-health information channel. The age coefficient is small but real, and points to interface support rather than exclusion for older users.

Taken together, the results support a narrow rather than a broad conclusion. General-purpose assistants should not be positioned as medication information sources: at 61.7–74.6% accuracy with harm rates of 4.6–8.3%, they fail the standard any pharmacy service would be held to. A constrained, retrieval-grounded voice layer over an authoritative formulary reaches accuracy consistent with routine counselling and delivers measurable knowledge and adherence benefit at a fraction of the time cost.

Limitations

The deployment was not blinded, and the control arm received no active digital comparator, so a portion of the knowledge effect may reflect attention rather than the voice modality specifically. Adherence was self-reported rather than verified by pill count or refill records, which risks over-estimation in the intervention arm. The twelve-week window cannot address long-term abandonment, and the declining weekly query volume suggests that a longer follow-up might attenuate effects. Phase I used a curated 240-item query bank posed in controlled acoustic conditions; real households introduce noise, overlapping speakers and interruptions that were not simulated. Finally, the speech corpus covered two speaker groups only, and the accent findings should not be generalised to other language backgrounds without replication.

Future Scope

Priority extensions include a randomised trial with objective adherence measurement and a six-to-twelve-month follow-up; evaluation in additional Indian languages with code-switched utterances, which are common in practice and poorly handled by current recognisers; and integration with dispensing records so that the assistant answers about the patient's actual regimen rather than a generic monograph. The harm-mitigation design also warrants isolation testing — specifically, whether the referral clause, the constrained corpus, or the phonetic alias table contributes most to the reduction in harmful responses. Finally, the interaction between conversational fluency and inappropriate trust deserves direct study: a more natural-sounding assistant may increase compliance with wrong answers as readily as with right ones.

Conclusion

Voice-activated assistants are already being consulted about medicines, whether or not health systems endorse them. This study shows that the general-purpose platforms performing that role today answer accurately in 61.7–74.6% of cases and give potentially harmful answers in up to 8.3%, with recognition penalties of over ten percentage points for accented speakers. A retrieval-grounded medication layer raised accuracy to 88.3%, cut harmful responses to 1.7%, narrowed the accent gap to 4.7 points, and in a twelve-week deployment produced a 3.6-point knowledge advantage, 2.65-fold higher odds of high adherence, and a fourfold reduction in query resolution time. The policy implication is that the value lies not in the voice interface itself but in what it is permitted to say.

References

1. Palanica A, Thommandram A, Lee A, Li M, Fossat Y. Do you understand the words that are comin outta my mouth? Voice assistant comprehension of medication names. *npj Digital Medicine*. 2019;2:55. doi:10.1038/s41746-019-0133-x
2. Palanica A, Fossat Y. Medication name comprehension of intelligent virtual assistants: a comparison of Amazon Alexa, Google Assistant, and Apple Siri between 2019 and 2021. *Frontiers in Digital Health*. 2021;3:669971. doi:10.3389/fdgth.2021.669971
3. Bickmore TW, Trinh H, Olafsson S, O'Leary TK, Asadi R, Rickles NM, Cruz R. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. *Journal of Medical Internet Research*. 2018;20(9):e11510. doi:10.2196/11510
4. Bickmore TW, Ólafsson S, O'Leary TK. Mitigating patient and consumer safety risks when using conversational assistants for medical information: exploratory mixed methods experiment. *Journal of Medical Internet Research*. 2021;23(11):e30704. doi:10.2196/30704
5. Miner AS, Milstein A, Schueller S, Hegde R, Mangurian C, Linos E. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Internal Medicine*. 2016;176(5):619–625. doi:10.1001/jamainternmed.2016.0400
6. Nobles AL, Leas EC, Caputi TL, Zhu SH, Strathdee SA, Ayers JW. Responses to addiction help-seeking from Alexa, Siri, Google Assistant, Cortana, and Bixby intelligent virtual assistants. *npj Digital Medicine*. 2020;3:11. doi:10.1038/s41746-019-0215-9

7. Alagha EC, Helbing RR. Evaluating the quality of voice assistants' responses to consumer health questions about vaccines: an exploratory comparison of Alexa, Google Assistant and Siri. *BMJ Health & Care Informatics*. 2019;26(1):e100075. doi:10.1136/bmjhci-2019-100075
8. Sezgin E, Huang Y, Ramtekkar U, Lin S. Readiness for voice assistants to support healthcare delivery during a health crisis and pandemic. *npj Digital Medicine*. 2020;3:122. doi:10.1038/s41746-020-00332-0
9. Yang S, Lee J, Sezgin E, Bridge J, Lin S. Clinical advice by voice assistants on postpartum depression: cross-sectional investigation using Apple Siri, Amazon Alexa, Google Assistant, and Microsoft Cortana. *JMIR mHealth and uHealth*. 2021;9(1):e24045. doi:10.2196/24045
10. Corbett CF, Combs EM, Chandarana PS, Stringfellow I, Worthy K, Nguyen T, Wright PJ, O'Kane JM. Medication adherence reminder system for virtual home assistants: mixed methods evaluation study. *JMIR Formative Research*. 2021;5(7):e27327. doi:10.2196/27327
11. Kim S. Exploring how older adults use a smart speaker–based voice assistant in their first interactions: qualitative study. *JMIR mHealth and uHealth*. 2021;9(1):e20427. doi:10.2196/20427
12. Kim S, Choudhury A. Exploring older adults' perception and use of smart speaker-based voice assistants: a longitudinal study. *Computers in Human Behavior*. 2021;124:106914. doi:10.1016/j.chb.2021.106914
13. Pradhan A, Lazar A, Findlater L. Use of intelligent voice assistants by older adults with low technology use. *ACM Transactions on Computer-Human Interaction*. 2020;27(4):1–27. doi:10.1145/3373759
14. Sezgin E. Redefining virtual assistants in health care: the future with large language models. *Journal of Medical Internet Research*. 2024;26:e53225. doi:10.2196/53225
15. Sezgin E, Chekeni F, Lee J, Keim S. Clinical accuracy of large language models and Google search responses to postpartum depression questions: cross-sectional study. *Journal of Medical Internet Research*. 2023;25:e49240. doi:10.2196/49240
16. Sawad AB, Narayan B, Alnefaie A, Maqbool A, Mckie I, Smith J, Yuksel B, Puthal D, Prasad M, Kocaballi AB. A systematic review on healthcare artificial intelligent conversational agents for chronic conditions. *Sensors*. 2022;22(7):2625. doi:10.3390/s22072625
17. Xu L, Sanders L, Li K, Chow JCL. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. *JMIR Cancer*. 2021;7(4):e27850. doi:10.2196/27850
18. Jadczyk T, Kiwic O, Khandwalla RM, Grabowski K, Rudawski S, Magaczewski P, et al. Feasibility of a voice-enabled automated platform for medical data collection: CardioCube. *International Journal of Medical Informatics*. 2019;129:388–393. doi:10.1016/j.ijmedinf.2019.07.001