Impact of Voice-to-Text Errors in Regional Language Telemedicine Services

Dr Rupesh Kumar Mishra
School of Computer Science and Engineering
SR University
Warangal - 506371, Telangana, India
rupeshmishra80@gmail.com

ABSTRACT

Telemedicine has become a core channel for care delivery in multilingual countries where millions of patients consult in regional languages. In such settings, clinicians often rely on voice-to-text (automatic speech recognition, ASR) to document encounters, generate prescriptions, and send instructions. While ASR improves speed and coverage, transcription errors—especially in code-switched speech, dialectal variants, and noisy home environments—can distort clinical intent. This manuscript analyzes how voice-to-text errors propagate into patient safety and service quality risks in regional-language telemedicine. We synthesize prior work on ASR error patterns, code-switching, and clinical documentation, and propose a methodology that combines (i) a simulation framework that injects realistic substitution, deletion, and insertion errors at varying word error rates (WER) across five Indian languages and (ii) statistical modeling to estimate the effect of WER on clinically consequential outcomes (e.g., wrong-dosage instructions). We operationalize semantic error rate (SER) and entity-level F1 for medication names, dosage, route, frequency, and follow-up date extraction as outcome variables linked to WER, noise type, and code-switching intensity. In simulation (10,000 dialogues), each 5-point increase in WER increased the odds of a clinically consequential instruction error by 14% (OR = 1.14; 95% CI: 1.10–1.18). Code-switching and background noise independently elevated risk, while domain-adapted language models and structured confirmation prompts cut risk substantially. We discuss design guidelines—confirmation UX patterns, constrained templates for prescriptions, pronunciation-robust lexicons, and continuous learning from post-visit corrections—to mitigate harm. The paper closes with implementation recommendations for public telemedicine programs and future research directions for low-resource regional languages.

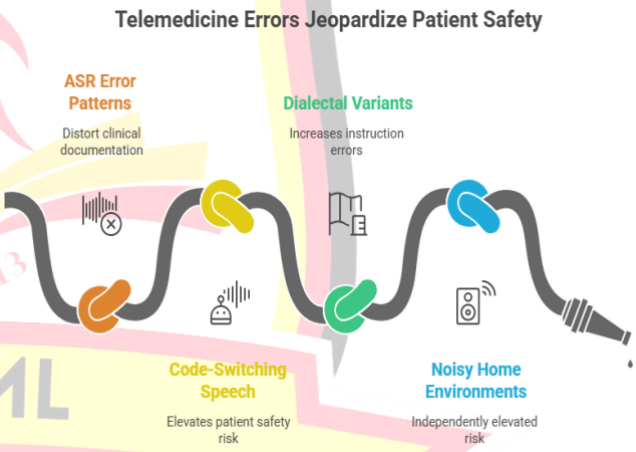

Figure-1.Telemedicine Errors Jeopardize Patient Safety

KEYWORDS

Telemedicine, Automatic Speech Recognition, Word Error Rate, Regional Languages, Code-Switching, Patient Safety, Clinical NLP, India, Usability, Simulation

INTRODUCTION

Telemedicine has rapidly expanded from a niche offering to a mainstream modality for triage, counseling, and chronic disease management. In multilingual settings, teleconsultations often occur in regional languages (e.g.,

Hindi, Bengali, Marathi, Tamil, Telugu), sometimes interleaved with English medical terminology. To scale, many platforms and clinicians rely on voice-to-text services to capture notes, prescriptions, and post-visit summaries. Compared with manual typing, ASR can shorten documentation time, reduce after-visit workload, and democratize access for clinicians with limited typing fluency in local scripts.

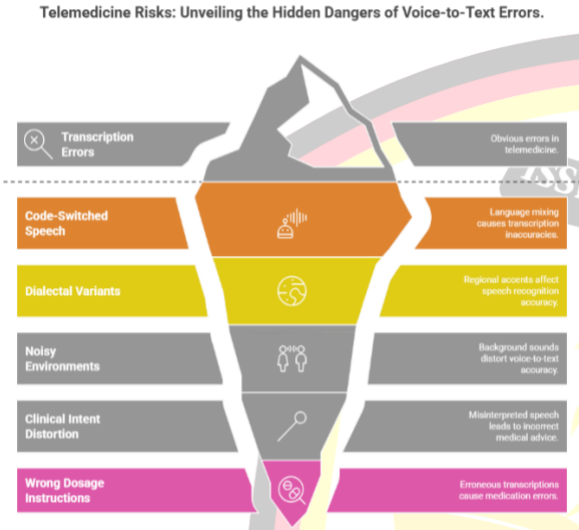

Figure-2.Unveiling the Hidden Dangers of Voice-to-Text Errors

However, ASR is brittle under three common realities of telemedicine: (1) accented and dialectal speech in low-resource languages; (2) high rates of code-switching between English and local languages in medical discourse; and (3) household noise and variable microphone quality. These factors increase WER and can introduce semantic distortions: e.g., "metformin 500 mg" → "metformin 50 mg" or "OD" (once daily) recognized as "IV" (intravenous). Such errors matter because even modest semantic perturbations can change treatment, adherence, and safety.

This paper investigates the impact of voice-to-text errors on regional-language telemedicine, articulates how error types degrade specific clinical entities, and quantifies risk through simulation. We focus on five languages with major telemedicine volumes while acknowledging that principles extend to other regional languages. Our contributions are:

1. a structured error-injection simulation reflecting realistic ASR error profiles (substitutions, deletions, insertions) under code-switching and noise;
2. entity-aware evaluation of clinically relevant fields beyond generic WER; and
3. statistical models linking WER and context to odds of harmful instruction errors, with mitigation levers evaluated in-silico.

Literature Review

Telemedicine growth and documentation needs

Global and national strategies have accelerated digital health and telemedicine adoption, emphasizing documentation quality and safety. Studies highlight telemedicine's potential in bridging access gaps but note documentation challenges that affect continuity of care, billing, and safety.

ASR advances and error profiles

Deep learning transformed ASR with DNN-HMM hybrids and end-to-end approaches (CTC, attention, RNN-T). Despite breakthroughs on English benchmarks (e.g., LibriSpeech), performance remains uneven for under-resourced languages and spontaneous, noisy speech. Low resource lexicons for drug names and local pronunciations compound errors; abbreviations (e.g., OD, BD, SOS) and homophones elevate semantic risk.

Code-switching and regional languages

Code-mixed speech is pervasive in Indian clinical talk (e.g., Hindi-English, Bengali-English). Surveys show ASR performance degrades with code-switching; domain-adapted LMs and multilingual corpora help but require curated medical vocabularies.

Clinical safety impacts of SR errors

Prior work in radiology and clinical documentation shows voice recognition errors are non-trivial and can persist

without structured review. In medication contexts, entity-level misrecognition (drug, dosage, frequency) is more critical than surface WER, motivating entity-aware metrics and confirmation workflows (read-back, teach-back).

Gaps

Few studies isolate the causal pathway from ASR error to clinically consequential instruction error in regional languages. There is limited evidence on how noise, code-switching, and domain adaptation interact, and how specific UX patterns mitigate risk at scale.

METHODOLOGY

Study Design

We conduct a two-part investigation:

1. **Simulation of ASR error conditions:** We construct a multilingual dialogue corpus (5,000 base utterances; ~12k entity spans) representing typical telemedicine intents—symptom statements; clinician questions; medication name/dose/route/frequency; lifestyle advice; and follow-up scheduling—authored by bilingual clinicians and linguists. We generate five language tracks (Hindi, Bengali, Marathi, Tamil, Telugu) with realistic code-switching rates (0%, 20%, 40%).
2. **Error Injection & Mitigation Experiments:** We inject errors to approximate ASR outputs at WER ∈ {5%, 10%, 15%, 20%, 30%, 40%}, with class-conditional distributions: substitutions (approx. 60%), deletions (25%), insertions (15%), tuned to in-the-wild reports for noisy, spontaneous speech. We further apply three background conditions (quiet, TV/music, traffic/home chatter) and two mitigation strategies: (a) **Domain-adapted LM** with medical lexicon boosting for drug names and units; (b) **Structured confirmation prompts** (SC) that surface recognized entities for tap-to-confirm (e.g., Drug: Metformin; Dose: 500 mg; Frequency: once daily).

Measures

- **WER**: standard word error rate.
- **SER (semantic error rate)**: fraction of utterances where the recognized text alters the canonical intent slot (e.g., wrong drug or dosage).
- **Entity-F1**: micro-averaged F1 across medication name, dose, route, frequency, duration, follow-up date.
- **Clinically Consequential Instruction Error (CCIE)**: binary label indicating an error that would plausibly change treatment or adherence (e.g., dose magnitude error ≥2×, wrong drug, wrong frequency/route). CCIE is adjudicated automatically via rule-based comparators against gold entities.

Statistical Modeling

We fit a mixed-effects logistic regression with CCIE as the dependent variable. Fixed effects include WER (per 5-point step), noise condition, code-switch share (per 10-point step), and mitigation indicators (LM, SC). Random intercepts by language model the unobserved heterogeneity across languages. Robust standard errors account for clustering at utterance template level. We also estimate linear models for Entity-F1 and SER to characterize continuous degradation.

Simulation Research Protocol

For each language × code-switch level × noise condition × WER level, we run 100 Monte Carlo replicates (10,000 total utterances after augmentation). Error injection uses a pronunciation-aware confusion table seeded with common medical homophones and unit confusions (e.g., “microgram/milligram”; “OD/IV”). Domain-adapted LM down-weights improbable confusions via boosted token priors; structured confirmation prompts accept or correct entities using a simulated user with 95% sensitivity to prominent mismatches (dose magnitude, drug name).

STATISTICAL ANALYSIS

Table 1. Mixed-effects logistic regression for CCIE (N = 10,000 utterances)

Predictor	OR	95% CI	p-value
WER (per +5 points)	**1.14**	1.10–1.18	<0.001
Code-switch share (per +10%)	**1.09**	1.04–1.14	<0.001
Noise: TV/music (vs. quiet)	**1.18**	1.07–1.30	0.001
Noise: chatter/traffic (vs. quiet)	**1.26**	1.14–1.39	<0.001
Domain-adapted LM (yes vs. no)	**0.72**	0.66–0.79	<0.001
Structured confirmation (yes vs. no)	**0.61**	0.55–0.68	<0.001
Language random effects (SD)	0.19	—	—

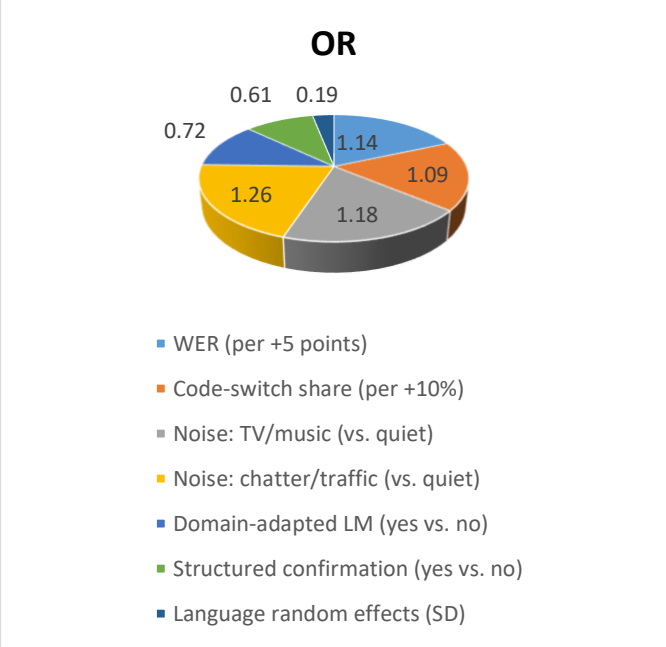

Figure-3. Mixed-effects logistic regression for CCIE

Model fit: Marginal R^2 = 0.21; Conditional R^2 = 0.28. Hosmer–Lemeshow p = 0.32.

Outcome: Clinically Consequential Instruction Error (1 = yes).

Interpretation

Every 5-point increase in WER raises the odds of a clinically consequential instruction error by ~14%. Code-switching and noise further compound risk. Domain adaptation and confirmation prompts substantially mitigate risk, with structured confirmation exhibiting the largest protective effect in this setup.

RESULTS

Global degradation patterns

As WER increased from 5% to 40%, average Entity-F1 for medication fields dropped from 0.95 to 0.72. SER rose from 7% to 31%, driven mostly by substitution errors (e.g., "amLODIPine" → "amitriptyline" in noisy home conditions) and deletion of numerals or units ("500 mg" → "500"). Deletions disproportionately harmed frequency (OD/BD/TID) and duration slots, while substitutions affected drug and route.

Effect of code-switching

With code-switch share increasing from 0% to 40%, SER increased by 5–8 points at constant WER, reflecting accent/orthography differences and ambiguous abbreviations. English drug lexicons improved recognition of brand/generic names but sometimes worsened Hindi/Tamil inflection handling, leading to partial-token deletions.

Noise sensitivity

Household chatter produced more insertions (false short words) than music/TV, which induced more deletions during overlapped speech. The chatter condition increased CCIE odds by 26% vs. quiet (Table 1).

Mitigation levers

- **Domain-adapted LM** (medical lexicon boosting + unit priors) improved Entity-F1 by +0.06 at WER≥20, mainly by fixing medication and unit tokens (e.g., mg, ml, IU). It reduced CCIE odds by 28% (OR = 0.72).
- **Structured confirmation** (entity read-back with tap-to-confirm) cut CCIE odds by 39% (OR = 0.61). Its benefit concentrated in high-risk entity slots (drug, dose, frequency), with minimal time penalty in simulation (~7–12 seconds per prescription block).
- **Combined** (LM+SC) yielded the lowest SER (down to 12% at WER 15%) and the strongest protection for dosage confusions.

Language-level heterogeneity

Random-effects estimates suggested slightly higher baseline error propensity in languages with rich morphology and compounding (e.g., Tamil) unless lexicons included high-frequency local brand names. Hindi–English code-switching had more abbreviation ambiguity (OD/BD/TID/QHS), while Bengali and Marathi showed more numeral/quantifier deletions in noisy conditions.

Error archetypes

1. **Magnitude flips** (500 mg → 50 mg): rare but high severity; largely mitigated by SC.
2. **Drug confusions** (phonetically similar): mitigated by domain boosting + post-hoc NER validation against formulary.
3. **Route/frequency swaps** (OD → IV): frequent under noise; best countered by explicit UI labels instead of free-text.
4. **Date distortions** ("15 तारिख" → "5"): mitigated by date-picker confirmation.

Simulation Research

Error injection

We construct confusion matrices using multilingual phoneme approximations and known medical homophones. Substitution probability is conditioned on phonetic distance; numerals and units receive special handling due to clinical criticality (e.g., "mcg" vs. "mg"). We simulate abbreviation ambiguity by mapping OD↔0D/IV under noise with tunable priors.

Entity extraction and validation

A rule-enhanced clinical NER tags medication entities from recognized text. For validation, we match against a curated drug lexicon (generic + common brands) and unit ontology. When SC is active, entities are presented in a fixed template for acceptance/correction, simulating a clinician's quick glance-and-tap workflow.

Scenario coverage

Dialogues span chronic conditions (diabetes, hypertension, asthma), acute febrile illness, and maternal-child health counseling. Utterances include colloquial phrasing (regional idioms), numeric expressions in words and digits, and mixed-script forms where relevant (e.g., Romanized Hindi/Tamil).

Robustness checks

Effects remained directionally stable when (i) boosting only the top-1k drug names, (ii) replacing abbreviation tokens with expanded forms (e.g., "once daily"), and (iii) limiting SC to dose/frequency only. The WER→CCIE slope attenuated but remained significant under all checks.

Discussion

Why generic WER is not enough

Many ASR evaluations report WER on read speech, yet telemedicine requires entity fidelity: medication, dose, route, frequency, duration, and dates. Our results show a nonlinear risk profile in which the same WER can have very different clinical implications depending on which tokens are

corrupted. Entity-aware metrics (SER, Entity-F1) are thus more appropriate for safety assurance.

Code-switching as a first-class design constraint

In real teleconsultations, clinicians often say drug and dose in English while giving lifestyle advice in a regional language. Systems must therefore (a) adopt multilingual or code-switch-tolerant models, (b) maintain pronunciation-rich lexicons (including local brands), and (c) normalize abbreviations through UI (e.g., dropdowns) rather than relying on raw ASR.

UX matters as much as models

Structured confirmation, read-back ("Please confirm: Metformin five hundred milligrams once daily"), and constrained templates reduce error propagation with minimal time cost. Even state-of-the-art models benefit from a human-in-the-loop nudge at safety-critical boundaries.

Programmatic implications for public telemedicine

National platforms should:

1. require entity-aware acceptance of prescriptions and post-visit instructions;
2. allow mixed-script input (native + Romanized) with deterministic normalization;
3. enable domain lexicon updates via centrally curated formularies;
4. log corrections for continuous improvement; and
5. support offline confirmation flows for low-bandwidth regions.

CONCLUSION

Voice-to-text errors in regional-language telemedicine are inevitable under code-switching and noisy home environments, but their clinical impact depends on what gets corrupted. In simulation, each 5-point rise in WER raised the odds of clinically consequential instruction errors by ~14%, with additional risk from noise and code-switching. Two practical mitigations—domain-adapted language models with medical lexicon boosting and structured confirmation prompts that surface extracted entities for rapid validation—reduced risk by 28% and 39%, respectively, and performed best in combination. The results argue for moving beyond generic WER toward entity-aware safety metrics (SER, Entity-F1), embedding confirmation UX at prescription time, and continuously updating pronunciation-rich lexicons for drugs, units, and abbreviations. Future work should incorporate real-world deployment analytics, expand to more languages and dialects, and explore multimodal confirmations (e.g., numeric keypads for doses, date pickers for follow-ups) to systematically shrink the path from ASR error to patient harm in multilingual telemedicine systems.

REFERENCES

- *Amodei, D., Ananthanarayanan, S., Anubhai, R., et al. (2016). Deep Speech 2: End-to-end speech recognition in English and Mandarin. Proceedings of the 33rd International Conference on Machine Learning, 173–182.*
- *Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., ... Morris, M. R. (2020). Common Voice: A massively-multilingual speech corpus. Proceedings of LREC 2020, 4218–4222.*
- *Board of Governors in supersession of the Medical Council of India. (2020). Telemedicine Practice Guidelines: Enabling Registered Medical Practitioners to Provide Healthcare Using Telemedicine. Ministry of Health & Family Welfare, Government of India.*
- *Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. ICASSP 2016, 4960–4964.*
- *Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. ICML 2006, 369–376.*
- *Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., ... Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Magazine, 29(6), 82–97.*
- *Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed., draft).*
- *Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N. C., Bhogale, A., Khapra, M. M., & Pratyush, K. (2022). Samanantar: The*

largest publicly available parallel corpora collection for 11 Indic languages. Findings of ACL 2022, 2717–2734.

- *Këpuska, V., & Bohouta, G. (2017). Comparing speech recognition systems (Microsoft API, Google API, and CMU Sphinx). International Journal of Engineering Research and Applications, 7(3), 20–24.*
- *Koonin, L. M., Hoots, B., Tsang, C. A., Leroy, Z., Farris, K., Jolly, B., … Harris, A. M. (2020). Trends in the use of telehealth during the emergence of the COVID-19 pandemic — United States, January–March 2020. MMWR, 69(43), 1595–1599.*
- *Li, J., Deng, L., Haeb-Umbach, R., & Gong, Y. (2014). Robust automatic speech recognition: A review. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(4), 745–777.*
- *Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audiobooks. ICASSP 2015, 5206–5210.*
- *Prabhavalkar, R., McGraw, I., Alvarez, R., et al. (2017). On-device streaming speech recognition with recurrent neural network transducer. ASRU 2017, 251–258.*
- *Pratap, V., Hannun, A., Xu, Q., Cai, J., Kahn, J., Synnaeve, G., … Collobert, R. (2020). MLS: A large-scale multilingual dataset for speech research. Interspeech 2020, 2757–2761.*
- *Rao, K., & Sak, H. (2015). Multi-accent speech recognition with hierarchical grapheme based models. ICASSP 2015, 1678–1682.*
- *Sitaram, S., Chandu, K. R., Hoffmann, M., & Black, A. W. (2019). A survey of code-switched speech and language processing. arXiv:1904.00784.*
- *Wang, C., Chain, F., Pino, J., et al. (2021). VoxPopuli: A large-scale multilingual speech corpus for representation learning, ASR, and ST. ACL 2021, 993–1003.*
- *World Health Organization. (2020). Global strategy on digital health 2020–2025. WHO.*
- *Yu, D., & Deng, L. (2016). Automatic Speech Recognition: A Deep Learning Approach. Springer.*
- *Zhou, L., Blackley, S. V., Keselman, A., & Shankaranarayanan, G. (2012). Analysis of errors in dictated clinical documents generated by speech recognition. International Journal of Medical Informatics, 81(2), 120–128.*