Perspective: Listening to Users when Auditing Medical AI Scribes

Allison Koenecke, PhD KOENECKE@CORNELL.EDU
Cornell Tech, USA

John-Jose Nunez, MD, MSc JOHNJOSE.NUNEZ@UBC.CA
University of British Columbia, Canada

Anaïs Rameau, MD, MPhil ANR2783@MED.CORNELL.EDU
Weill Cornell Medicine, USA

Irene Y. Chen, PhD IYCHEN@BERKELEY.EDU
UC Berkeley, USA

Abstract

Medical AI scribes are rapidly being adopted to reduce documentation burdens on clinicians, with systems already deployed across millions of patient visits. While these tools offer substantial efficiency benefits and reduced clinician burnout, they pose serious risks through transcription errors and hallucinations. These risks are disproportionately placed on certain demographics of speakers, from patients with speech disorders to psychiatric illnesses. We argue for more principled audits to be conducted on medical AI scribes, analogous to post-marketing surveillance for medical devices. Our framework for doing so involves: (1) collecting diverse, medically-relevant speech datasets representative of real patient and provider populations, (2) developing metric suites that go beyond the singular gold standard of Word Error Rates, and (3) conducting human-centered design research to align functionality with the needs of both medical providers and patients.

Keywords: AI scribe, speech-to-text transcription, automated speech recognition, algorithmic fairness, dialect bias, speech diversity, hallucinations, audits

1. Introduction

A patient stated their medical history verbally: "*I became ill with a fairly serious strain of viral something.*" However, OpenAI's speech transcription tool Whisper (Radford et al., 2023) appended fabricated phrases to the transcription, instead generating: "*I became ill with a fairly serious strain of viral something, but I didn't take any medication, I took Hyperactivated Antibiotics and sometimes I would think that was worse*" (Koenecke et al., 2024). This phenomenon—known as "hallucination"—occurs in AI-based speech transcription systems that are now being deployed at scale as medical scribes, automatically generating transcriptions and patient notes from clinical encounters. One such system is Nabla, which is built on Whisper and has transcribed over 7 million patient visits from over 30,000 clinicians and 40 health systems (Burke and Schellmann, 2024). Another AI scribe product owned by Microsoft has already been used by over half a million doctors in the United States (Nuance, 2023).

The rapid adoption of AI scribes responds to a genuine crisis in clinical practice. Physicians spend over half of their workday documenting in the EHR, compared to only a quarter of their workday spent interacting with patients (Arndt et al., 2017, 2024; Sinsky et al., 2016). The amount of time spent documenting has steadily increased (Holmgren et al., 2023), and is associated with burnout, reduction in work effort, and turnover (Gardner et al., 2019; Melnick et al., 2021; Doan-Wiggins et al., 1995). The appeal of these technologies is clear: clinical notes are written faster, more patients can be seen, and medical providers can provide more attentive care to patients (Mohr et al., 2003; Suominen et al., 2015; van Buchem et al., 2024; Shin et al., 2025). According to one study, the percentage of clinicians reporting burnout decreased significantly from 51.9% to 38.8% after 30 days with an ambient AI scribe (Olson et al., 2025). These benefits are real and substantial.

However, these promising results (often focused on efficiency) come with currently poorly understood risks (often compromising on accuracy or safety). Whisper has been found to yield hallucinations 1% of the time (Koenecke et al., 2024). Journalists found that not only was there no awareness of the possibility of hallucinations in Nabla output, but it was "also impossible to compare Nabla's AI-generated transcript to the original recording because Nabla's tool erases the original audio for *data safety reasons*" (Burke and Schellmann, 2024). This means that fabricated content, once embedded in a patient's medical record, could propagate through subsequent care decisions—with no mechanism for detection or correction. The implications of these errors extend far beyond simple transcription mistakes. Errors in patients' electronic charts can lead to harm (Singh et al., 2012), with the onus for catching these errors frequently on the patients; serious errors are often reported by patients to their medical providers, due to clinical notes access from the 21st Century Cures Act (Bell et al., 2020).

These risks are not distributed equally. Transcribed hallucinations have been found to be disproportionately frequent for patients with communication disorders such as aphasia (Koenecke et al., 2024) and stutters (Sridhar and Wu, 2025), which points to a further concern: AI scribes could yield worse performance on patients (or in interactions with medical providers) who have diverse speech patterns. This builds on a robust line of auditing work showing that—prior to the generative AI age wherein hallucinations were a leading concern—even the overall *accuracy* of automated speech transcriptions has varied substantially by speaker type. For example, speech transcription audits have quantified accuracy rates that reveal worse performance (relative to a control group) for people with dysphonia (Hidalgo Lopez et al., 2023), dysarthria (Hasegawa-Johnson et al., 2024; Zheng et al., 2025a), stuttering (Lea et al., 2023; Mujtaba et al., 2024; Teleki et al., 2024b), the d/Deaf and hard of hearing (Zhao et al., 2025), and people with other non-"standard," race- and ethnicity-based accents (Koenecke et al., 2020; Wassink et al., 2022; Dubois et al., 2024). This underperformance compounds intersectionally, leading to disproportionately worse transcriptions for already-disadvantaged intersectional identities: for example, by gender and geography (Tatman, 2017), or by race and speech impairment (Mei et al., 2025).

Extrapolating further, there is good reason to believe that AI scribes would exhibit differential performance in non-"standard" settings—not just for patients or medical providers who may have diverse speech patterns, but also in clinical specialties wherein patient speech patterns are less likely to be well-represented in training data (from neurology and speech language pathology to psychiatry and pediatrics). Such concerns are compounded when considering patients and medical providers who may speak in different dialects, code switch, or even require multilingual services.

Currently, relatively little is known about the performance of various medical AI scribes apart from what performance metrics the companies release about themselves (and occasionally, their competitors) in the form of a self-audit (Oberst et al., 2024; Radford et al., 2023). However, there are serious limitations of these audits:

1. They are performed on a relatively standard set of benchmark datasets which may not be representative of diverse speakers in medical environments.

2. They rely on a small subset of metrics, which may not map neatly onto optimal performance desired by medical professionals, and often do not capture the potential harms arising from AI hallucinations.

3. The audit results do not lead directly to ideation about the rich, alternative ways in which users may want to experience AI scribes in practice.

In this piece, we propose an auditing framework that considers both quantitative and qualitative research directions as first steps towards addressing each of these concerns. First, we encourage collection of, and benchmarking on, datasets containing greater speech diversity reflective of real-world patient and medical provider populations. Second, we advocate for reporting a wider set of benchmark metrics that can provide medical providers and procurement teams with more signal on the types of errors that different AI scribe products might make, and on which populations. Third, we propose that further human-centered design practices could lead to smoother experiences for medical providers and patients alike. Taken together, we argue for more robust longitudinal audits of medical AI scribes, propose an implementation roadmap, and conclude with open challenges in meeting the moment.

2. Proposed Audit Framework

Performing real-world audits of medical AI scribes is imperative so that both patients and medical providers have a better sense of how well the technology works: patients who understand the limitations of the AI scribes can better advocate for themselves; medical providers may be better informed regarding the types of errors to be aware of when editing transcriptions; and, medical systems on the whole can be better informed when going through the procurement process to determine which AI scribe system to purchase (Corpus et al., 2025). One might imagine the auditing process for AI scribes as analogous to post-marketing surveillance—which encompasses directives and regulation regarding the data collection and monitoring of certain products on the market (such as medical devices or drugs) (Badnjević et al., 2022). In both cases, it is important to monitor performance longitudinally to ensure that patients using the product will not be harmed by the product over time; and, as an added benefit, the collected data can be used to make more informed downstream decisions regarding future product development or usage.

For speech transcription technology, such as medical AI scribes, a standard audit might involve: identifying a benchmark dataset (e.g., speech data with corresponding "ground truth" transcriptions), generating AI scribe-based transcriptions on that benchmark dataset, and calculating performance metrics (such as "accuracy") by comparing the AI scribe-generated transcriptions to ground truth transcriptions. A similar framework to post-marketing surveillance could be applied to such audits, to ensure that the accuracy of medical AI scribes is continuously monitored in the populations they serve. We propose more granular auditing for medical AI scribes in three thrusts below.

2.1. Collect Diverse Datasets

Current benchmark datasets used in audits may not be ecologically valid in the medical context. For example, many existing benchmark datasets (such as the commonly-used Librispeech corpus (Panayotov et al., 2015), explicitly mentioned in the Oberst et al. (2024) audit) are predominantly spoken using "standard" American English, and do not regard medical contexts. Prior work has advocated for diversifying the set of speech datasets used for benchmarking purposes (Papakyriakopoulos et al., 2023; Agnew et al., 2024), though this is a large effort in and of itself, involving participant recruitment, speech recording, and ground truth transcription generation (ideally verified by human experts—yet another layer of costs).

Choices made at the dataset level can highly alter the results of downstream audits. For example, if none of the participants recruited have a clinical speech impairment, it is likely that the audit's performance would overstate the AI scribe's quality on patients with clinical speech impairments. If speech is recorded on lower-quality microphones (Fahed et al., 2025), or includes hospital background noise (Barhoush et al., 2022), performance could similarly differ. Real-world data often include multiple speakers in a back-and-forth conversation, which necessitates technology for "speaker diarization" (i.e., segmenting audio and correctly determining which speaker uttered which phrase), which becomes more technically complicated when multiple parties' speech is overlapping, more still with more speakers in the conversation, and even more with multilingual speakers (as when a medical interpreter is in the room). Training on messy conversational data is important for medical applications, but remains difficult to collect and annotate; diarization error rates remain high for leading AI scribe models (Tran et al., 2023).

Furthermore, the notion of the "ground truth" transcription could be very different depending on who generates it: a patient with a stutter may prefer omitting stutters in a transcription, whereas a speech pathologist might prefer including stutters for diagnostic purposes (Mei et al., 2025). For example, one valid "ground truth" transcription of a speaker could be the verbatim sentence, *"Uh, I'd been saying, n-no, no."* But, another valid "ground truth" transcription could remove filler words and stutters and formalize a contraction, yielding the sentence *"I had been saying no."* The choice of which sentence to use as the ground truth in underlying datasets can have downstream effects, as accuracy metrics are quantified based on similarity of a generated transcription to the ground truth—so, even a one-word change in ground truth can lead to amplified differences in accuracy metrics.

While increasing numbers of medical-specific speech datasets have been open-sourced (Le-Duc et al., 2025; Fareez et al., 2022; MacWhinney, 2019; Suominen et al., 2015; Hasegawa-Johnson et al., 2024), these ironically result in less-meaningful audits of AI scribes, as public speech data sources are likely ingested as training data to improve AI scribes on a rolling basis, thereby rendering them biased as a source of testing data. As such, it is imperative to develop large-scale, diverse, medically-oriented speech datasets that include significant held-out portions to be used for benchmarking purposes. Furthermore, because collecting speech datasets for niche speech diversities is such a difficult task, few auditors go through the extra effort of combining multiple datasets for easy benchmarking across a range of speech impairments, races, genders, etc.—and, encapsulating intersectionality across these features requires even larger combinations of datasets with high quality metadata labeling. So, we propose generating more one-stop shop meta-datasets—both via new data collection, and via aggregation across existing datasets spanning domains—that allow auditors to robustly test performance by different medically-relevant subgroups. This is a space wherein the role of unbiased third parties—such as academics—is crucial in maintaining such benchmark tasks and leaderboards.

2.2. Generate Metric Suites

Nearly all speech transcription tools are audited with a single metric, the Word Error Rate (WER), which quantifies the normalized edit distance between the ground truth transcription and the AI scribe-generated transcription. Specifically, WER is defined as the number of changes between the ground truth and AI-generated transcription (i.e., the number of word-level substitutions, deletions, or insertions), divided by the total number of words in the ground truth (Jurafsky and Martin, 2009).[1] However, the WER metric does not encapsulate any semantic meaning; metrics more commonly seen in machine translation, such as BLEU or ROUGE scores, provide better performance quantification for more semantically-similar transcriptions. And, the WER is at the word unit level; for languages such as Chinese, quantification at the character unit level (e.g., using CER) could make more sense. Meanwhile, in light of hallucination concerns for AI scribes, reporting the hallucination rate as a separate metric has also been proposed (Koenecke et al., 2024); the hallucination rate cannot be directly proxied from the WER alone (Frieske and Shi, 2024). Finally, for medical applications of speech AI, another metric is of key importance: the medical term recall rate. This metric simply reports recall on some pre-determined dictionary of clinically-relevant vocabulary words, which may differ in relevance by subdomain (Suominen and Ferraro, 2013; Sadeghi et al., 2014; Oberst et al., 2024; Jelassi et al., 2024); there is no singular gold standard currently used as the dictionary of medical terms for generating this metric.

There is clear benefit to reporting a wide range of metrics to quantify medical AI scribe performance: each of the above metrics can give some signal as to cases where the scribe might fail (e.g., a high medical term recall rate, paired with a high hallucination rate, could mean that medical providers would need to pay more attention towards catching and deleting hallucinations, and less attention towards correcting misspellings of medical terms)—something useful for both AI scribe usage and procurement. However, most auditors only report the WER; the primary HuggingFace leaderboard for speech transcription reports only WER and latency (Srivastav et al., 2023)—a metric also solely focused on efficiency, as it refers to the speed of transcription. However, there is a burgeoning line of research advocating for benchmark suites that report a full slate of relevant metrics (Wang et al., 2024; Mei et al., 2025) rather than a single metric in a leaderboard. And, even if considering only a single metric, auditors can still report a range for that metric—for example, by quantifying the set of WERs occurring from different reasonable variants of "ground truth" (Mei et al., 2025).

Furthermore, while a range of speech datasets are included as part of the HuggingFace leaderboard, none are specific to medical speech, nor are subgroup-level breakdowns of accuracy provided. Even when reporting a full suite of metrics, it is still important to report them by subgroup to observe whether disparities in accuracy exist (Mei et al., 2025)—for example, by speaker diagnosis, gender, accent, or other speech diversity. This level of subgroup analysis can only be done if underlying datasets are appropriately labeled with the relevant subgroups, which speaks to the importance of documentation and metadata collection when creating new datasets (Papakyriakopoulos et al., 2023).

Finally, the above metrics are best fit for standard transcription tasks, and may not make sense for medical AI scribes that go one step further and additionally function as speech summarization tools. For such tools, metrics might include: ROUGE or F1 scores (Teleki et al., 2024a) that quantify relevance to a ground truth (which necessitates generation of a summary ground truth), or average ratings gen-

[1] Changing the length of the ground truth (for example, increasing the word count by 1) correspondingly changes the denominator of the WER metric (for example, increasing the length of the ground truth by 1 word would necessarily increase the denominator by 1; the change in the numerator would depend on the comparison to the AI-generated transcription.

erated by human annotators on factors like fluency, consistency, relevance, and coherence (Le-Duc et al., 2024). And, usability features may also be important to quantify, such as EMR integration, time to sign-in and launch the scribe, or adherence to SOAP[2] structure (Ha et al., 2025). Again, reporting a suite of metrics can allow for easier assessment of trade-offs; for example, an AI scribe that is significantly easier and faster for medical practitioners to functionally use may be preferred to an AI scribe that yields a marginally better WER.

At present, little has been done in metric development for summaries best fit to medical contexts, which may require a combination of medical term recall alongside summarization-specific metrics. It will be pertinent to develop new, forward-looking metrics as both speech AI and medical systems advance. And, as we gain a better understanding of clinically-meaningful errors made by AI scribes, we can better pinpoint which (suite of) metrics most explain those errors in particular—allowing for a more clinically-targeted set of metrics to be the focus of future audits.

When considering procurement, it may be natural to seek out prescriptive guidance on how to make trade-offs among the medical AI scribes on the market, and which metrics to prioritize. However, we argue that these are decisions that should be made closely with the medical practitioners and patients in mind. For example, practices with large shares of Mexican or Mexican-American patients may want to focus on AI scribes that perform best in multilingual, Spanglish, or Chicano English transcription. Those AI scribes may differ from the ones that would be preferred by departments working mostly with geriatric populations who speak with more hoarseness. To best align with organizational preferences, large-scale audits could be conducted with an eye towards relevant patient populations, the results of which could be applied in advanced survey methods (such as conjoint analyses, potentially leading to multi-objective optimization) to elicit survey preferences. In Koenecke et al. (2023), stakeholders are surveyed on pairwise preferences chosen along a Pareto frontier generated from audit data; these preferences could be quantified as, e.g., win rates for specific medical AI scribe products. For organizations with more limited resources, a "minimum viable audit" may entail simply ensuring that an AI scribe's WERs are below a certain threshold, with a reporting mechanism for medical providers who run into serious concerns with the scribe.

2.3. Design for Diverse Users Long-term

There is a great need not only for quantitative audits (as described above), but also for qualitative studies to better understand the needs of medical providers when generating patient notes, the desires of patients reading these notes, and the ideal roles of AI scribes as part of a broader workplace pipeline. For example, do users prefer verbatim transcriptions, "cleaned" transcriptions that omit disfluencies (such as filler words—like "uh" and "um"—or stutters), reduction to a set of bulletpoints, or even just outputs of a few keywords? If different medical providers have different preferences, can these be accommodated in a single procured technology? For example, perhaps diagnostic tools could be combined together with AI scribes to support neurologists, speech language pathologists, and otolaryngologists (Bensoussan et al., 2024), whose patients' voices could be an underutilized digital biomarker of health.

User studies have led to design developments in the machine translation field, such as visualizing uncertainty in AI text output (Robertson and Díaz, 2022), and applying medical practitioners' strategies to assist with communication (such as pre-translation and backtranslation of phrases) to technical advancements (Mehandru et al., 2022, 2023). Mei et al. (2025) suggest the idea of community-driven audits (as inspired by the field of participatory design), wherein marginalized speakers themselves are surveyed regarding their preferences for how "ground truth" is recorded. These, and other design and communication choices (Alumäe and Koenecke, 2025; Wu et al., 2025)—such as ensuring that patients fully understand the implications of their verbal consent for using AI scribes—are important to study in the medical domain. It may also be useful to ideate about hardware developments: what types of microphones are unobtrusive, but allow the patient to be aware that their speech is being recorded? What locations of microphones lead to the best recording quality while minimizing nuisance? Could other locations (e.g., microphones embedded in crash carts (Taylor et al., 2019)) be considered?

Another key part of designing for medical use cases involves better understanding the underlying AI scribe technology. For example, modern AI scribes

[2] Acronym for medical note documentation: Subjective, Objective, Assessment, Plan.

apply technology similar to Large Language Models, which—at a high level—predict words likely to come next in a sequence. While this model architecture may work well for "typical" speakers, it may lead to breakdowns in transcriptions—such as disproportionate hallucinations or mistranscriptions—for "atypical" speakers. "Atypical" speakers might not only include speakers with speech impairments, but could also include psychiatric patients whose speech may follow unexpected patterns. Multiple common psychiatric illnesses can affect both the content and production of speech. For example, speech may become rapid and tangential during mania, disorganized or impoverished in psychotic disorders, or monotone, quiet, and stilted in depression (American Psychiatric Association, 2013). In clinical practice, psychiatrists often exercise discretion in documenting sensitive content, summarizing or paraphrasing patient speech to preserve dignity and privacy. For instance, a clinician might note that a patient "endorses a persecutory delusion involving aliens" whereas an AI scribe might transcribe the patient's detailed account verbatim, such as "the patient is currently being pursued by Martians seeking to use his DNA for a galactic cloning project." However, there is limited research evaluating medical AI scribes or related automated techniques in patients with mental illness (Tougas et al., 2022; Gabor-Siatkowska et al., 2023). Speaking to the need for more diverse datasets, existing studies have largely involved individuals who were considered "non-urgent" or "stable." And speaking to designing for consent, obtaining data from patients with mental illness also requires careful consideration, as symptoms can impact their ability to provide informed consent. Psychotic content often reflects contemporary societal themes, particularly those involving technology and surveillance (Higgins et al., 2023). Given that a common paranoid delusion involves being monitored or controlled (American Psychiatric Association, 2013), AI scribes may be especially prone to incorporation into a patient's delusional framework. This may necessitate clinical caution to ensure that the use of such tools does not inadvertently reinforce delusional beliefs or cause harm. Nonetheless, AI scribe providers are already marketing their tools for mental health settings (Heidi Health, 2025), and adoption within clinical practice has begun (Cass County Communication Network, 2025)—though more research must be conducted, from improving model architecture adapting to these users' needs, to more comprehensive consent processes for these users.

3. Implementation Roadmap

Our proposed audit framework involves collection of diverse datasets that include both audio data and (potentially diverse variants of) ground truth; generating transcriptions or patient note summaries from these audio data using a range of medical AI scribes; calculating a suite of metrics regarding the quality and/or usability of the generated outputs; and using the gained knowledge to make procurement decisions and ideate on better design opportunities, all in a community-driven manner. This is a tall ask for anyone, and obfuscates many key implementation details. We attempt to provide an end-to-end roadmap below, but caveat that there is high variance in what audits might look like in practice—across clinical specialties, funding levels, geographies, and so on.

1. **Who** conducts the audit? In our framework, all audits can be conducted by third-party auditors (i.e., it is possible to audit proprietary systems whose internal architectures are unavailable). While we believe it benefits medical organizations to run their own audits, we recognize that there are significant costs to set up audit infrastructure, dataset collection, and continuous monitoring. There are several paths forward: independent auditors in the form of academic or technology bodies could be the de facto auditors. In the longer term, perhaps regulatory bodies could take on this role. There are several examples of collective, recurrent audits that lend themselves to a "leaderboard"—from those based in academic conferences such as the Speech Accessibility Challenge (Zheng et al., 2025b), to public HuggingFace leaderboards (Srivastav et al., 2023), to the speech recognition evaluations run by NIST (the U.S. National Institute of Standards and Technology, which ran their first speech recognition competition in 1996) (Sadjadi et al., 2022). In these settings, medical AI scribe providers themselves could play a role in *submitting* their models to such competitions, but they are not the ones running their own audits (which would be a conflict of interest). We envision that one or multiple teams of third-party auditors could generate medical-specific variants of such "leaderboard" competitions, producing a

suite of relevant metrics that would serve as the primary audit source used by downstream customers (though, they would still need to understand and make trade-offs between the presented metrics).

2. **What** exactly would the auditors do? Their primary responsibility—in addition to ensuring that aforementioned leaderboard competitions occur regularly—would be to serve as custodians of privately-held data. As public datasets easily become absorbed as training data, the auditors should ideally maintain private validation and test sets of speech data, and update these data sources over the years as speech patterns change (e.g., with new medication names on the market, or different slang being used) and technology advances (e.g., with different microphone recording capabilities). It may be in the public interest to release older collected data for general use, and remove it from newer batches of data used for auditing purposes. In addition to maintaining these data (including collecting new data), the auditors would be in charge of deciding which metrics to use for evaluation (which may include both automated metrics like WER, or ones that require human input—for which the auditors would need to recruit human annotators), and publicly releasing the results of the audit in a timely manner.

3. **When** (or how frequently) would audits occur? With the support of external institutions (whether academic or civic tech coalitions, or government organizations), we expect that audits would occur both regularly and long-term. Towards the former, with API calls, "leaderboards" could even be updated continuously (with new models added as requested). Annual competitions in the style of conference challenges could serve as forcing functions for AI scribe developers to develop better models year over year. Long-term sustainability could be achieved by aligning with existing regulatory or post-market monitoring frameworks.

4. **How** would (or should) audits be regulated? Regulation (let alone enforcement) is nascent for medical AI scribes. While some national agencies (such as the UK Medicines and Healthcare Products Regulatory Agency (MHRA) and NHS England) have released guidance (Shemtob et al., 2025), there do not yet appear to be a comprehensive set of standards for this high-stakes technology. As an analogue, early attempts have been made in the US at regulating speech AI tools used in the employment domain, such as with New York's Local Law 144. While this 2023 law mandated bias audits for "automated employment decisions systems" used in hiring and promotions, the letter of the law with regard to auditing was susceptible to loopholes (allowing for limited audits), and there were minimal monitoring mechanisms to enforce audits—leading to a dearth in compliance (Wright et al., 2024; Gerchick et al., 2025). That said, in the medical domain, incentives are somewhat more aligned: HIPAA covered entities (such as physicians or hospitals) are already liable for any AI bias under the Affordable Care Act (ACA) Section 1557. Paths forward are myriad: while medical AI scribe audits could fall under US FDA regulatory frameworks (though perhaps compromising on speed of audits), they could also focus on regulating *transparency* from AI scribe developers via other (speedier) modalities of public reporting. It remains to be seen how the medical domain will approach this policy problem, and how to enforce such audits.

4. Open Challenges

Taken together, the above auditing directions can allow researchers to advance the field: determining underlying reasons for AI scribe underperformance, fine-tuning models or updating model architecture to ameliorate disparities in performance, and deploying designs for AI scribe interfaces and outputs that better align with medical needs. We hope to center patients and medical providers in this dialogue, by studying whether they perceive or experience compromise on safety and accuracy—in exchange for efficiency—with the use of medical AI scribes.

While we find this line of work promising, we conclude with notes on open challenges. The largest challenge we envision is principled data collection at scale. Obtaining a wide range of diverse voices, in varied audio settings, is difficult at baseline. Generating accurate ground truths to use as comparisons to AI scribes requires nontrivial amounts of labor. Additionally, working with patient speech data in particular leads to complex IRB-related constraints because audio recordings are inherently identifiable and therefore subject to a higher level of data protections.

Another concern regards the speed with which the AI landscape is updating. Audits must be conducted regularly to account for advancements in the AI scribes themselves, but also to account for changes in the devices with which they are used, and the ever-changing context of modern medicine. However, benchmark datasets released for testing should be assumed to be ingested as training data in later models, thereby reducing their validity as testing datasets longitudinally. As such, audits by lower-resourced organizations relying on such public data should account for likely overperformance when assessing downstream metrics.

In conclusion, more research needs to be done to audit AI scribes for accuracy and safety, not just for efficiency. Currently, the onus is primarily on the medical provider to understand limitations of their office-sanctioned AI scribes and edit ensuing transcription errors. While AI scribes can help with administrative burdens, there remains the possibility that massive errors can slip through the cracks. We maintain that the onus should be on medical AI scribe developers—and not medical providers or patients—to mitigate the potential transcription errors, in order for the community as a whole to better provide patient care downstream. As with drugs and medical devices that undergo post-marketing surveillance, medical AI scribes should be subject to ongoing, longitudinal evaluation of their accuracy in real-world clinical populations.

References

William Agnew, Julia Barnett, Annie Chu, Rachel Hong, Michael Feffer, Robin Netzorg, Harry H Jiang, Ezra Awumey, and Sauvik Das. Sound check: Auditing audio datasets. *arXiv preprint arXiv:2410.13114*, 2024.

Tanel Alumäe and Allison Koenecke. Striving for open-source and equitable speech-to-speech translation, 2025.

American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association, 5th edition, 2013. doi:10.1176/appi.books.9780890425596.

Brian G Arndt, John W Beasley, Michelle D Watkinson, Jonathan L Temte, Wen-Jan Tuan, Christine A Sinsky, and Valerie J Gilchrist. Tethered to the ehr: Primary care physician workload assessment using ehr event log data and time-motion observations. *The Annals of Family Medicine*, 15 (5):419–426, 2017. doi: 10.1370/afm.2121.

Brian G Arndt, Mark A Micek, Adam Rule, Christina M Shafer, Jeffrey J Baltus, and Christine A Sinsky. More tethered to the ehr: Ehr workload trends among academic primary care physicians, 2019-2023. *The Annals of Family Medicine*, 22(1):12–18, 2024. doi: 10.1370/afm.3047.

Almir Badnjević, Lejla Gurbeta Pokvić, Amar Deumić, and Lemana Spahić Bećirović. Post-market surveillance of medical devices: A review. *Technology and Health Care*, 30(6):1315–1329, 2022.

Mahdi Barhoush, Ahmed Hallawa, Arne Peine, Lukas Martin, and Anke Schmeink. Localization-driven speech enhancement in noisy multi-speaker hospital environments using deep learning and meta learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:670–683, 2022.

Sigall K Bell, Tom Delbanco, Joann G Elmore, Patricia S Fitzgerald, Alan Fossa, Kendall Harcourt, Suzanne G Leveille, Thomas H Payne, Rebecca A Stametz, Jan Walker, et al. Frequency and types of patient-reported errors in electronic health record ambulatory care notes. *JAMA network open*, 3(6): e205867–e205867, 2020.

Yaël Bensoussan, Olivier Elemento, and Anaïs Rameau. Voice as an ai biomarker of health—introducing audiomics. *JAMA Otolaryngology–Head & Neck Surgery*, 150(4): 283–284, 2024.

Garance Burke and Hilke Schellmann. Researchers say an ai-powered transcription tool used in hospitals invents things no one ever said. *The Associated Press*, 2024. URL https://apnews.com/article/ai-artificial-intelligence-health-business-90020cdf5fa16c79ca2e5b6c4c9bbb14.

Cass County Communication Network. 4c health partners with heidi health to enhance clinician support with ai technology. July 2025. URL https://www.casscountyonline.com/2025/07/4c-health-partners-with-heidi-health-to-enhan ce-clinician-support-with-ai-technology/. Accessed: 2025-10-10.

Isabel Corpus, Eric Giannella, Allison Koenecke, and Don Moynihan. As government outsources more it, highly skilled in-house technologists are more essential. *Commun. ACM*, 68(7):37–40, June 2025. ISSN 0001-0782. doi: 10.1145/3727635. URL https://doi.org/10.1145/3727635.

L Doan-Wiggins, L Zun, MA Cooper, DL Meyers, and EH Chen. Practice satisfaction, occupational stress, and attrition of emergency physicians. wellness task force, illinois college of emergency physicians. *Academic Emergency Medicine*, 2(6):556–563, 1995. doi: 10.1111/j.1553-2712.1995.tb03261.x.

Daniel J Dubois, Nicole Holliday, Kaveh Waddell, and David Choffnes. Fair or fare? understanding automated transcription error bias in social media and videoconferencing platforms. *Proceedings of the International AAAI Conference on Web and Social Media*, 18:367–380, May 2024. ISSN 2162-3449. doi: 10.1609/icwsm.v18i1.31320. URL http://dx.doi.org/10.1609/icwsm.v18i1.31320.

Vitória S Fahed, Emer P Doheny, Monica Busse, Jennifer Hoblyn, and Madeleine M Lowery. Comparison of acoustic voice features derived from mobile devices and studio microphone recordings. *Journal of Voice*, 39(2):559–e1, 2025.

Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, et al. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9(1): 313, 2022.

Rita Frieske and Bertram E Shi. Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models. *arXiv preprint arXiv:2401.01572*, 2024.

Karolina Gabor-Siatkowska, Marcin Sowański, Rafał Rzatkiewicz, I. Stefaniak, Marek Kozłowski, and Artur Janicki. Ai to train ai: Using chatgpt to improve the accuracy of a therapeutic dialogue system. *Electronics*, 2023. doi: 10.3390/electronics12224694.

Rebekah L Gardner, Emily Cooper, Jacqueline Haskell, Daniel A Harris, Sara Poplau, Philip J Kroth, and Mark Linzer. Physician stress and burnout: The impact of health information technology. *Journal of the American Medical Informatics Association*, 26(2):106–114, 2019. doi: 10.1093/jamia/ocy145.

Marissa Kumar Gerchick, Ro Encarnación, Cole Tanigawa-Lau, Lena Armstrong, Ana Gutiérrez, and Danaé Metaxa. Auditing the audits: Lessons for algorithmic accountability from local law 144's bias audits. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 29–44, 2025.

Emily Ha, Isabelle Choon-Kon-Yune, LaShawn Murray, Siying Luan, Enid Montague, Onil Bhattacharyya, and Payal Agarwal. Evaluating the usability, technical performance, and accuracy of artificial intelligence scribes for primary care: Competitive analysis. *JMIR Human Factors*, 12(1): e71434, 2025.

Mark Hasegawa-Johnson, Xiuwen Zheng, Heejin Kim, Clarion Mendes, Meg Dickinson, Erik Hege, Chris Zwilling, Marie Moore Channell, Laura Mattie, Heather Hodges, et al. Community-supported shared infrastructure in support of speech accessibility. *Journal of Speech, Language, and Hearing Research*, 67(11):4162–4175, 2024.

Heidi Health. Ai scribe for mental health. https://www.heidihealth.com/en-ca/solutions/mental-health, 2025. Accessed: 2025-10-10.

Julio C Hidalgo Lopez, Shelly Sandeep, MaKayla Wright, Grace M Wandell, and Anthony B Law. Quantifying and improving the performance of speech recognition systems on dysphonic speech. *Otolaryngology–Head and Neck Surgery*, 168(5): 1130–1138, 2023.

Oliver Higgins, Brooke L. Short, Stephan K. Chalup, and Rhonda L. Wilson. Interpretations of innovation: The role of technology in explanation seeking related to psychosis. *Perspectives in Psychiatric Care*, 2023(1):4464934, 2023. doi: https://doi.org/10.1155/2023/4464934. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2023/4464934.

A Jay Holmgren, Robert Thombley, Christine A Sinsky, and Julia Adler-Milstein. Changes in physician electronic health record use with the expansion of telemedicine. *JAMA Internal Medicine*, 183(12): 1357–1365, 2023. doi: 10.1001/jamainternmed.2023.5738.

Mariem Jelassi, Oumaima Jemai, and Jacques Demongeot. Revolutionizing radiological analysis: The future of french language automatic speech recognition in healthcare. *Diagnostics*, 14(9):895, 2024.

Dan Jurafsky and James H. Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, 2009.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14):7684–7689, 2020.

Allison Koenecke, Eric Giannella, Robb Willer, and Sharad Goel. Popular support for balancing equity and efficiency in resource allocation: A case study in online advertising to increase welfare program awareness. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 494–506, 2023.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X Mei, Hilke Schellmann, and Mona Sloane. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1672–1681, 2024.

Khai Le-Duc, Khai-Nguyen Nguyen, Long Vo-Dang, and Truong-Son Hy. Real-time speech summarization for medical conversations. In *Interspeech 2024*, Interspeech 2024, page 1960–1964. ISCA, September 2024. doi: 10.21437/interspeech.2024-2250.

Khai Le-Duc, Phuc Phan, Tan-Hanh Pham, Bach Phan Tat, Minh-Huong Ngo, Thanh Nguyen-Tang, and Truong-Son Hy. MultiMed: Multilingual medical speech recognition via attention encoder decoder. In Georg Rehm and Yunyao Li, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1113–1150, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-288-6. doi: 10.18653/v1/2025.acl-industry.79. URL https://aclanthology.org/2025.acl-industry.79/.

Colin Lea, Zifang Huang, Jaya Narain, Lauren Tooley, Dianna Yee, Dung Tien Tran, Panayiotis Georgiou, Jeffrey P Bigham, and Leah Findlater. From user perceptions to technical improvement: Enabling people who stutter to better use speech recognition. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–16, 2023.

Brian MacWhinney. Understanding spoken language through talkbank. *Behavior research methods*, 51 (4):1919–1927, 2019.

Nikita Mehandru, Samantha Robertson, and Niloufar Salehi. Reliable and safe use of machine translation in medical settings. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2016–2025, 2022.

Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and backtranslation identifies critical errors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647, 2023.

Katelyn Xiaoying Mei, Anna Seo Gyeong Choi, Hilke Schellmann, Mona Sloane, and Allison Koenecke. Addressing pitfalls in auditing practices of automatic speech recognition technologies: A case study of people with aphasia. *arXiv preprint arXiv:2506.08846*, 2025.

Edward R Melnick, Allan Fong, Bidisha Nath, Brian Williams, Raj M Ratwani, Richard Goldstein, Ryan T O'Connell, Christine A Sinsky, Daniel Marchalik, and Mihriye Mete. Analysis of electronic health record use and clinical productivity and their association with physician turnover. *JAMA Network Open*, 4(10):e2128790, 2021. doi: 10.1001/jamanetworkopen.2021.28790.

David N Mohr, David W Turner, Gregory R Pond, Joseph S Kamath, Cathy B De Vos, and Paul C Carpenter. Speech recognition as a transcription aid: a randomized comparison with standard transcription. *Journal of the American Medical Informatics Association*, 10(1):85–93, 2003.

Dena Mujtaba, Nihar Mahapatra, Megan Arney, J Yaruss, Hope Gerlach-Houck, Caryn Herring, and Jia Bin. Lost in transcription: Identifying and quantifying the accuracy biases of automatic speech recognition systems against disfluent speech. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4795–4809, 2024.

Nuance. Nuance announces the general availability of dragon ambient experience copilot to further improve healthcare experiences, outcomes, and efficiency. *Cision PR Newswire*, 2023.

Michael Oberst, Davis Liang, and Zachary C. Lipton. Pioneering the science of ai evaluation. *Abridge*, 2024. URL https://www.abridge.com/ai/science-ai-evaluation#automatic-speech-recognition.

Kristine D Olson, Daniella Meeker, Matt Troup, Timothy D Barker, Vinh H Nguyen, Jennifer B Manders, Cheryl D Stults, Veena G Jones, Sachin D Shah, Tina Shah, et al. Use of ambient ai scribes to reduce administrative burden and professional burnout. *JAMA Network Open*, 8(10):e2534976–e2534976, 2025.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

Orestis Papakyriakopoulos, Anna Seo Gyeong Choi, William Thong, Dora Zhao, Jerone Andrews, Rebecca Bourke, Alice Xiang, and Allison Koenecke. Augmented datasheets for speech datasets and ethical decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 881–904, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

Samantha Robertson and Mark Díaz. Understanding and being understood: User strategies for identifying and recovering from mistranslations in machine translation-mediated chat. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 2223–2238, 2022.

Fateme Sadeghi, Marjan Ghazisaeedi, Reza Safdari, and Abdoljalil Kalantarhormozi. Developing a standardized medical speech recognition database for reconstructive hand surgery. *International Journal of Hospital Research*, 3(3):149–154, 2014.

Seyed Omid Sadjadi, Craig Greenberg, Elliot Singer, Lisa Mason, and Douglas Reynolds. The 2021 nist speaker recognition evaluation. *arXiv preprint arXiv:2204.10242*, 2022.

Lara Shemtob, Azeem Majeed, and Thomas Beaney. Regulation of ai scribes in clinical practice. *BMJ*, 389:r1248, June 2025. ISSN 1756-1833. doi: 10.1136/bmj.r1248. URL http://dx.doi.org/10.1136/bmj.r1248.

H Stella Shin, Herb Williams, Nikolay Braykov, Afrin Jahan, Jeremy Meller, and Evan W Orenstein. The influence of artificial intelligence scribes on clinician experience and efficiency among pediatric subspecialists: A rapid, randomized quality improvement trial. *Applied Clinical Informatics*, 16(04):1041–1052, 2025.

Hardeep Singh, Traber Davis Giardina, Samuel N Forjuoh, Michael D Reis, Steven Kosmach, Myrna M Khan, and Eric J Thomas. Electronic health record-based surveillance of diagnostic errors in primary care. *BMJ quality & safety*, 21(2):93–100, 2012.

Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Annals of Internal Medicine*, 165(11):753–760, 2016. doi: 10.7326/M16-0961.

Charan Sridhar and Shaomei Wu. J-j-j-just Stutter: Benchmarking Whisper's Performance Disparities on Different Stuttering Patterns. In *Interspeech 2025*, pages 3753–3757, 2025. doi: 10.21437/Interspeech.2025-2700.

Vaibhav Srivastav, Somshubra Majumdar, Nithin Koluguri, Adel Moumen, Sanchit Gandhi, et al. Open automatic speech recognition leaderboard. https://huggingface.co/spaces/hf-audio/open_asr_leaderboard, 2023.

Hanna Suominen and Gabriela Ferraro. Noise in speech-to-text voice: Analysis of errors and feasibility of phonetic similarity for their correction. In Sarvnaz Karimi and Karin Verspoor, editors, *Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013)*, pages 34–42, Brisbane, Australia, December 2013. URL https://aclanthology.org/U13-1006/.

Hanna Suominen, Liyuan Zhou, Leif Hanlen, and Gabriela Ferraro. Benchmarking clinical speech recognition and information extraction: new data, methods, and evaluations. *JMIR medical informatics*, 3(2):e4321, 2015.

Rachael Tatman. Gender and dialect bias in YouTube's automatic captions. In Dirk Hovy, Shannon Spruit, Margaret Mitchell, Emily M. Bender, Michael Strube, and Hanna Wallach, editors, *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1606. URL https://aclanthology.org/W17-1606/.

Angelique Taylor, Hee Rin Lee, Alyssa Kubota, and Laurel D Riek. Coordinating clinical teams: Using robots to empower nurses to stop the line. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–30, 2019.

Maria Teleki, Xiangjue Dong, and James Caverlee. Quantifying the impact of disfluency on spoken content summarization. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13419–13428, Torino, Italia, May 2024a. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1175/.

Maria Teleki, Xiangjue Dong, Soohwan Kim, and James Caverlee. Comparing asr systems in the context of speech disfluencies. *Interspeech 2024*, pages 4548–4552, 2024b.

Hailee Tougas, S. Chan, Tara Shahrvini, Alvaro D Gonzalez, Ruth Chun Reyes, Michelle Burke Parish, and Peter M Yellowlees. The use of automated machine translation to translate figurative language in a clinical setting: Analysis of a convenience sample of patients drawn from a randomized controlled trial. *JMIR Mental Health*, 9, 2022. doi: 10.2196/39556.

Brian D Tran, Ramya Mangu, Ming Tai-Seale, Jennifer Elston Lafata, and Kai Zheng. Automatic speech recognition performance for digital scribes: a performance comparison between general-purpose and specialized models tuned for patient-clinician conversations. In *AMIA Annual Symposium Proceedings*, volume 2022, page 1072, 2023.

Marieke Meija van Buchem, Ilse MJ Kant, Liza King, Jacqueline Kazmaier, Ewout W Steyerberg, and Martijn P Bauer. Impact of a digital scribe system on clinical documentation time and quality: usability study. *JMIR AI*, 3(1):e60020, 2024.

Angelina Wang, Aaron Hertzmann, and Olga Russakovsky. Benchmark suites instead of leaderboards for evaluating ai fairness. *Patterns*, 5(11), 2024.

Alicia Beckford Wassink, Cady Gansen, and Isabel Bartholomew. Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication*, 140:50–70, 2022.

Lucas Wright, Roxana Mika Muenster, Briana Vecchione, Tianyao Qu, Pika Cai, Alan Smith, Comm 2450 Student Investigators, Jacob Metcalf, J Nathan Matias, et al. Null compliance: Nyc local law 144 and the challenges of algorithm accountability. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1701–1713, 2024.

Shaomei Wu, Kimi Wenzel, Jingjin Li, Qisheng Li, Alisha Pradhan, Raja Kushalnagar, Colin Lea, Allison Koenecke, Christian Vogler, Mark Hasegawa-Johnson, et al. Speech ai for all: Promoting accessibility, fairness, inclusivity, and equity. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2025.

Robin Zhao, Anna SG Choi, Allison Koenecke, and Anaïs Rameau. Quantification of automatic speech recognition system performance on d/deaf and hard of hearing speech. *The Laryngoscope*, 135(1): 191–197, 2025.

Xiuwen Zheng, Bornali Phukon, Jonghwan Na, Ed Cutrell, Kyu Han, Mark Hasegawa-Johnson, Pan-Pan Jiang, Aadhrik Kuila, Colin Lea, Bob MacDonald, et al. The interspeech 2025 speech accessibility project challenge. *arXiv preprint arXiv:2507.22047*, 2025a.

Xiuwen Zheng, Bornali Phukon, Jonghwan Na, Ed Cutrell, Kyu J. Han, Mark Hasegawa-Johnson, Pan-Pan Jiang, Aadhrik Kuila, Colin Lea, Bob MacDonald, Gautam Mantena, Venkatesh Ravichandran, Leda Sari, Katrin Tomanek, Chang D. Yoo, and Chris Zwilling. The interspeech 2025 speech accessibility project challenge. In *Interspeech 2025*, interspeech2025, page 3269–3273. ISCA, August 2025b. doi: 10.21437/interspeech.2025-566. URL http://dx.doi.org/10.21437/Interspeech.2025-566.