Regulating Gen-AI enabled Ambient Clinical Documentation Devices

Disclaimer: This content has been generated with the assistance of a large language model. While steps have been taken to ensure the accuracy and relevance of the information presented, there may still be hallucinations and inconsistencies.

Introduction

Ambient clinical documentation (ACD) is a technology that utilizes artificial intelligence (AI) to transcribe and interpret patient-clinician conversations. ACD platforms leverage advanced speech recognition and natural language processing technologies to generate clinical notes for review by clinicians. This technology is commonly implemented via mobile applications on smartphones, utilizing the device's microphone to capture conversations (**Tierney et al., 2024, Misurac et al., 2023,** Albrecht et al., 2024). Ambient clinical documentation is a dominant application of generative AI in healthcare attracting substantial investments in research and development, as well as the rapid implementation of these technologies across diverse healthcare settings. There are various ambient AI platforms and tools designed specifically for this purpose, including AWS HealthScribe, Nuance Dragon Ambient eXperience (DAX), DeepScribe, and 3M M*Modal Fluency Align.

Benefits of Ambient Clinical Documentation:

- **Reduces Documentation Burden:** ACD alleviates the time clinicians spend on documentation, allowing them to focus more on direct patient care and potentially reducing after-hours work (**Tierney et al., 2024, Misurac et al., 2023, Bundy et al., 2024).**
- **Improves Physician Well-being:** By reducing administrative burdens and facilitating more time for patient interaction, ACD can positively impact physician satisfaction and mitigate burnout (**Tierney et al., 2024, Misurac et al., 2023, Bundy et al., 2024,** Albrecht et al., 2024, Nahar et al., 2023).
- **Enhances Patient Care:** ACD enables clinicians to be more present and attentive during patient encounters, potentially leading to improved patient satisfaction and communication (**Tierney et al., 2024, Bundy et al., 2024, Doshi et al., 2024).**
- **Improves Documentation Quality:** ACD can contribute to more complete and accurate clinical notes, as it captures all aspects of the conversation, including nuances that might be missed during manual notetaking (**Tierney et al., 2024, Bundy et al., 2024, Doshi et al., 2024).**

Why Ambient Clinical Documentation Should Be Regulated as a Medical Device by the FDA:

The FDA regulates medical devices to ensure their safety and effectiveness. Given the significant impact ACD can have on patient care and the potential risks involved, it is essential for this technology to be regulated as a medical device. Here is why:

- **Ensuring Accuracy and Reliability:** FDA regulation can mandate rigorous testing and validation of ACD systems to ensure their accuracy in transcribing conversations and generating reliable clinical notes. The FDA can set standards for performance metrics, including word error rate, clinical concept extraction accuracy, and adherence to established medical documentation guidelines.
- **Addressing Potential Biases:** ACD systems rely on AI algorithms trained on large datasets. FDA oversight can help mitigate potential biases embedded in these datasets, ensuring fair and equitable application of the technology across diverse patient populations.
- **Protecting Patient Privacy and Security:** ACD systems manage sensitive patient information. FDA regulation can establish robust security protocols and data privacy measures to prevent unauthorized access, breaches, and misuse of patient data.
- **Managing Risks and Liabilities:** As with any medical device, ACD systems carry potential risks, such as errors in transcription leading to misdiagnosis or inappropriate treatment. FDA regulation can provide a framework for identifying, mitigating, and managing these risks, clarifying liability in case of adverse events.

The integration of ACD into healthcare carries significant potential benefits, but its implementation requires careful consideration of associated risks and ethical implications. FDA regulation can establish necessary safeguards and standards to ensure the safe, effective, and responsible use of this technology, contributing to improved patient care and outcomes.

AI Lifecycle Applied to Ambient Clinical Documentation

The AI Lifecycle concept can be applied to the development and implementation of Ambient Clinical Documentation systems in healthcare.

AI Lifecycle Phases for Ambient Clinical Documentation:

1. **Planning & Design**: Define goals (e.g., reduce documentation burden), requirements (e.g., EHR integration, HIPAA compliance), and Design AI algorithms (speech recognition, NLP).
2. **Data Collection**: Gather training data (patient-clinician conversations), ensure quality (cleaning, de-identification), and establish governance (data privacy).
3. **Model Building**: Develop ML algorithms, extract features, and train models for optimal performance.
4. **Verification & Validation**: Test AI models (accuracy, completeness), validate clinical note quality with clinician feedback.
5. **Deployment**: Integrate with clinical environments, connect to EHR, and train users.
6. **Operation & Monitoring**: Track performance, troubleshoot issues, and update system as needed.
7. **Evaluation**: Assess clinical impact (efficiency, satisfaction), gather feedback for continuous improvement.

Considerations for Premarket Performance Evaluation of GenAI-Enabled Ambient Clinical Documentation Devices

To thoroughly evaluate the safety and efficacy of generative AI-enabled Ambient Clinical Documentation (ACD) devices, the FDA needs access to specific information regarding the device's functionality, particularly considering the evolving nature of foundation models and potential limitations in training data transparency.

- **Transparency regarding foundation models:**
  - While manufacturers might not be able to disclose all training data details due to proprietary concerns, it is crucial to provide the FDA with information regarding the **types of data used for training** these models. This should include the **source of the data**, **data formats**, and any **preprocessing steps**. This transparency enables the FDA to assess the **representativeness and potential biases** within the training data (Biswas et al. 2024, Yim et al. 2023).
  - It is important to provide information on the **model's architecture** and **training methodology** (Tierney et al., 2024, Nuance 2021).
  - To address the evolving nature of foundation models, manufacturers should detail the **mechanisms for updating these models** and explain how these updates will be **validated to ensure the device's continued safety and effectiveness (Tierney et al., 2024, Biswas et al., 2024).**
- **Performance evaluation in realistic clinical settings:**
  - Premarket testing should go beyond simulated environments and involve **evaluation in real-world clinical settings**. This evaluation should include **diverse patient populations and a range of clinical scenarios**, including **edge cases and infrequent medical conditions**, to adequately assess the model's robustness and generalizability (Biswas et al. 2024).
  - The evaluation should focus on **key performance indicators relevant to ACD devices**. These include:
    - **Accuracy of speech recognition**: It is crucial to evaluate the **impact of non-lexical conversational sounds (NLCS)** such as "Mm-hm" or "Uh-huh," as misinterpreting these sounds can lead to inaccuracies in documentation (Tran et al., 2023).
    - **Quality and completeness of generated notes**: The FDA needs evidence demonstrating the **notes' accuracy**, **conciseness**, and **adherence to required clinical documentation standards (Tierney et al., 2024, Misurac et al., 2023)**
    - **Model's ability to manage complex medical terminology and language**: Assess the system's capacity to **accurately interpret and document intricate medical discussions** to ensure accurate and reliable note generation (Biswas et al., 2024).

- **Safety and risk mitigation strategies**:
    - Given the potential for AI models to generate inaccurate information (**"hallucinations"**) or omit critical details, manufacturers must detail the **measures implemented to detect and mitigate these risks** (**Misurac et al., 2023, Yim et al., 2023).**
    - It is vital to provide a clear plan for **ongoing monitoring** of the device's performance in the post-market phase to **identify and address any emerging safety concerns or performance degradation** (Albrecht et al., 2024).
- **Data privacy and security**:
    - The FDA needs assurance that the ACD device **complies with HIPAA regulations** and that patient data is **managed securely**. This includes outlining **data encryption methods**, **access control mechanisms**, and **data storage and retention policies** (Biswas et al., 2024).
    - Transparency regarding **how patient data will be used for model training and improvement** is essential to address ethical concerns related to data privacy (Nahar et al., 2023).
- **Human oversight and validation**:
    - While ACD devices aim to automate documentation, the FDA needs to understand the role of human oversight in this process. Manufacturers should clarify **how clinicians will review and validate the AI-generated notes** to ensure accuracy and completeness before inclusion in the patient's medical record **(Tierney et al., 2024, Biswas et al., 2024).**
- **Explainability and Interpretability**:
    - While complete transparency of the model's inner workings might not be feasible, it is important to provide the FDA with some level of **explainability regarding the model's decision-making process**. This could involve providing **interpretable outputs** or using techniques like **attention mechanisms** to highlight the key aspects of the conversation that influenced the note generation. This level of explainability helps the FDA **build trust** in the model's outputs and assess potential biases (**Tierney et al., 2024).**

These considerations are essential for ensuring the safe and effective deployment of GenAI-enabled ACD devices in clinical practice. By providing the FDA with this information, manufacturers can facilitate a thorough premarket review process, contributing to improved patient care and safety.

Information to Include in a Premarket Submission for GenAI-Enabled Ambient Clinical Documentation Device

Here is a comprehensive overview of the information that should be included as part of a device's description or characterization in the premarket submission for a generative AI-enabled Ambient Clinical Documentation (ACD) device:

- **Intended Use:**
    - Clearly state the intended use of the ACD device, specifying the clinical settings where it is meant to be used, target user groups (e.g., physicians, nurses), and the type of patient encounters it is designed to support.
    - **Human in the Loop:** Explicitly describe the **level of human oversight and interaction** intended with the device. Is the device designed for fully autonomous documentation generation, or does it require clinician review and validation of the AI-generated notes? Explain the rationale behind the chosen level of human involvement and its implications for the device's safety and effectiveness (**Tierney et al., 2024, Biswas et al., 2024).**
    - **Generative vs. Recall:** Clearly articulate **whether the device solely recalls and summarizes information from the patient-clinician interaction or if it also generates new recommendations or interpretations**. This distinction is crucial for assessing the potential risks and benefits associated with the device (Nuance, 2021, **Yim et al., 2023).**
- **Description of Generative AI Components:**
    - **Foundation Models:**
        - **Model Architecture and Training Methodology:** Provide details on the model's architecture, including the type of neural network used (e.g., transformer), the number of layers, and the training methodology employed. Explain the rationale behind selecting this specific model architecture and training approach (**Yim et al., 2023).**
        - **Training Data:** While exhaustive disclosure of training data might not be feasible, provide information on the types of data used to train the foundation models, the source of the data (e.g., publicly available datasets, de-identified patient records), and data formats (e.g., text, audio). Describe any data preprocessing steps performed (Tran et al., 2023, **Yim et al., 2023).**
        - **Model Updates:** Explain how the foundation models will be updated over time and the process for validating these updates to ensure ongoing safety and efficacy. Describe the frequency of anticipated updates and how users will be notified and trained on these changes (**Yim et al., 2023).**

  - **Fine-Tuning and Customization:** Describe any fine-tuning or customization performed on the pretrained models to adapt them for ACD specific tasks. Outline the data used for fine-tuning, the process for evaluating the performance of the fine-tuned models, and any mechanisms for ongoing customization based on user feedback or changing clinical needs (**Yim et al., 2023).**

  - **Generative Output:** Specify the types of output generated by the ACD system. Does it produce full clinical notes, specific sections of notes (e.g., History of Present Illness), or summaries of key clinical findings? Explain how the generative output is presented to the user and the mechanisms for editing or modifying this output (**Yim et al., 2023).**

- **Information Relevant to Evaluating Safety and Effectiveness**

  - **Accuracy and Performance Benchmarks:** Provide comprehensive evidence demonstrating the device's accuracy and performance in realistic clinical settings. This should include:

    - **Speech Recognition:** Report the word error rate of the speech recognition component, paying particular attention to the impact of non-lexical conversational sounds (NLCS). Compare the performance to industry benchmarks or established standards for speech recognition in clinical settings (Tran et al., 2023).

    - **Note Generation:** Present metrics evaluating the quality, completeness, and consistency of the generated notes. This could involve comparing AI-generated notes to clinician-written notes using standardized evaluation tools or expert review (**Yim et al., 2023, Tierney et al., 2024).**

    - **Handling of Complex Medical Language:** Demonstrate the device's ability to accurately understand and document complex medical terminology and nuanced clinical discussions. This could be assessed through expert review of generated notes for encounters involving specialized medical conditions or by evaluating the system's performance on domain-specific language benchmarks (**Yim et al., 2023).**

    - **Benchmarking Against Non-Generative AI:** To highlight the value of generative AI, compare the performance of the ACD device to existing non-generative AI-based clinical documentation tools, such as traditional speech recognition software or rule-based note generation systems (**Yim et al., 2023).**

  - **Safety and Risk Mitigation:**

    - **"Hallucination" Detection and Mitigation:** Describe strategies employed to detect and mitigate the risk of the AI generating inaccurate information or fabricating clinical details ("hallucinations"). This could involve using statistical methods to identify unusual or unlikely text patterns, incorporating domain-specific knowledge constraints into the model, or implementing human review processes to flag potential inaccuracies (**Tierney et al., 2024, Misurac et al., 2023, Bundy et al., 2024).**

    - **Omission Detection:** Outline methods for detecting and preventing omissions of critical clinical information from the generated notes. This might involve cross-referencing the generated notes with key elements of the patient encounter (e.g., chief complaint, vital signs) or implementing prompts for clinicians to review and confirm the completeness of specific note sections (**Misurac et al., 2023).**

    - **Bias Detection and Mitigation:** Detail the steps taken to identify and address potential biases in the AI models or training data. This could include analyzing the model's outputs across different patient demographics or using fairness metrics to assess for disparate performance across subgroups (**Tierney et al., 2024,** Nahar et al., 2023).

- **Data Privacy and Security:**

    - **HIPAA Compliance:** Demonstrate that the device complies with all relevant HIPAA regulations regarding the handling, storage, and transmission of protected health information (PHI). Provide details on data encryption methods, access control mechanisms, and data de-identification procedures (**Yim et al., 2023, Misurac et al., 2023).**

    - **Data Usage Transparency:** Clearly explain how patient data will be used for model training and improvement. Obtain explicit consent from patients for any data usage beyond the immediate clinical documentation purpose. Describe data retention policies and mechanisms for data deletion upon patient request (Nahar et al., 2023).

- **Human Oversight and Validation:** Detail the role of human clinicians in reviewing and validating the AI-generated notes before they are finalized and included in the patient's medical record. Explain how the device facilitates this review process, including mechanisms for highlighting potential errors, inconsistencies, or areas requiring clinician attention. Specify the level of clinician training required to use the device effectively and safely (**Tierney et al., 2024, Misurac et al., 2023, Bundy et al., 2024, Biswas et al., 2024).**

- **Explainability and Interpretability:**

    - **Rationale for Decisions:** While full transparency of the model's decision-making process might not be achievable, strive to provide clinicians with insights into the rationale behind the AI-generated output. This could involve highlighting the key phrases or concepts from the conversation that informed the note generation (**Tierney et al., 2024,** Nahar et al., 2023).

    - **Uncertainty Estimation:** Where applicable, the device should provide an indication of the model's confidence or uncertainty in its generated output. This allows clinicians to focus their review efforts on areas where the AI might be less certain and helps build trust in the system's capabilities (**Yim et al., 2023).**

- **Usability and Workflow Integration:**

    - **Ease of Use:** Demonstrate that the device is user-friendly and integrates seamlessly into existing clinical workflows. Provide evidence from usability studies involving representative users to assess the device's learnability, efficiency, and overall satisfaction (**Tierney et al., 2024, Misurac et al., 2023).**

        - **EHR Integration:** If the device is intended to integrate with electronic health record (EHR) systems, provide details on the integration process and ensure compatibility with commonly used EHR platforms. Address any potential challenges or limitations related to EHR integration (**Tierney et al., 2024, Misurac et al., 2023, Gollaway et al., 2024).**

- **Post-Market Surveillance and Ongoing Monitoring:** Outline a comprehensive plan for post-market surveillance and ongoing monitoring of the ACD device's performance and safety. Describe mechanisms for collecting user feedback, tracking adverse events, and analyzing real-world data to identify any emerging issues or areas for improvement. Specify procedures for implementing corrective actions or updates based on post-market data (**Tierney et al., 2024,** Albrecht et al., 2024).

By providing the FDA with detailed information addressing these points, manufacturers can demonstrate a commitment to the safe, effective, and responsible use of generative AI in clinical documentation, paving the way for a future where AI empowers clinicians to focus on delivering exceptional patient care.

Evidence for FDA Premarket Evaluation of GenAI-Enabled Ambient Clinical Documentation

To understand if a generative AI-enabled Ambient Clinical Documentation (ACD) device is safe and effective, the FDA should consider specific evidence regarding performance evaluation and the characteristics of the training data during the total product lifecycle. Here is a breakdown of key evidence categories:

Performance Evaluation

- **Speech Recognition Accuracy:**
    - **Word Error Rate (WER):** The FDA should review the WER of the speech recognition component, specifically evaluating its performance in handling **non-lexical conversational sounds (NLCS)** like "Mm-hm" or "Uh-huh", which frequently occur in clinical conversations and convey clinically relevant information (Tran et al., 2023).
    - **Impact of NLCS on Downstream Tasks:** Evaluate how errors in recognizing NLCS might impact downstream NLP tasks, potentially leading to incomplete or inaccurate clinical documentation (Tran et al., 2023).
    - **Comparison to Benchmarks:** The WER should be compared to established benchmarks or standards for speech recognition in clinical settings and to the performance of other commercially available ASR engines tailored for clinical conversations (Tran et al., 2023).
- **Note Generation Quality:**
    - **Accuracy, Completeness, and Consistency:** The FDA should assess the accuracy, completeness, and consistency of the AI-generated notes through:

        - **Comparison to Clinician-Written Notes:** Compare AI-generated notes to clinician-written notes using standardized evaluation tools like the Physician Documentation Quality Instrument (PDQI-9), potentially modified to account for AI-specific considerations like "hallucinations" and bias (**Tierney et al., 2024).**

        - **Expert Review:** Subject a sample of AI-generated notes to expert review by clinicians from relevant specialties to evaluate the quality and clinical relevance of the documentation (**Yim et al., 2023).**

        - **Real-World Data Analysis:** Analyze data from actual patient encounters where the ACD device was used to assess its performance in real-world clinical settings. This includes examining the frequency and types of clinician edits made to AI-generated notes (**Tierney et al., 2024).**

    - **Handling of Complex Medical Language:**

        - **Specialized Medical Conditions:** Assess the device's performance in accurately documenting encounters involving specialized medical conditions, complex terminology, and nuanced discussions (**Biswas et al., 2024).**

        - **Domain-Specific Benchmarks:** Evaluate the system's ability to understand and document complex medical language using domain-specific language benchmarks or datasets (**Yim et al., 2023).**

- **Benchmarking Against Non-Generative AI:** Compare the ACD device's performance to existing non-generative AI-based clinical documentation tools (e.g., traditional speech recognition software) to demonstrate the added value and potential benefits of generative AI (Albrecht et al., 2024).

- **Safety and Risk Mitigation:**

    - **"Hallucination" Detection and Mitigation:** The FDA should evaluate the strategies implemented to detect and mitigate AI "hallucinations," including statistical methods, domain knowledge constraints, and human review processes (**Bundy et al., 2024).**

    - **Omission Detection:** Review methods for detecting and preventing omissions of critical information, such as cross-referencing with key encounter elements and prompts for clinician review (**Bundy et al., 2024, Tierney et al., 2024).**

    - **Bias Detection and Mitigation:** Assess the steps taken to identify and address potential biases in models or training data, such as analyzing outputs across demographics and using fairness metrics (**Biswas et al., 2024,** Nahar et al., 2023).

Characteristics of the Training Data

- **Data Sources and Representativeness:** The FDA should consider the sources of data used to train the foundation models and the representativeness of this data in terms of:

    - **Clinical Settings and Specialties:** The data should encompass a range of clinical settings, specialties, and patient demographics to ensure the ACD device performs reliably across diverse clinical contexts (**Misurac et al., 2023).**

    - **Medical Conditions and Terminology:** The training data should include a variety of medical conditions, terminology, and clinical scenarios to ensure the model can manage complex medical language and accurately document diverse patient encounters (**Biswas et al., 2024).**
    - **Language Variations:** The data should account for variations in spoken language, including accents, dialects, and common speech patterns, to ensure the speech recognition component performs consistently across different patient populations (**Tierney et al., 2024).**
- **Data Quality:**
    - **Accuracy and Completeness:** The FDA should evaluate the accuracy and completeness of the training data, as errors or inconsistencies in the data can propagate to the AI model's performance (**Biswas et al., 2024).**
    - **Data Preprocessing:** Assess the data preprocessing steps undertaken to clean and prepare the data for model training, ensuring these steps are appropriate and do not introduce bias or distort the data (Tran et al., 2023).
- **Data Privacy and Security:**
    - **De-Identification:** Evaluate the de-identification process to ensure that all protected health information (PHI) is removed or adequately protected, complying with HIPAA regulations (**Yim et al., 2023, Misurac et al., 2023).**
    - **Data Usage Transparency and Consent:** Review the mechanisms for obtaining informed consent from patients for any data usage beyond the immediate clinical documentation purpose (Nahar et al., 2023).
- **Model Updates:** The FDA should assess the process for updating the AI models over time, including the frequency of updates and the procedures for validating these updates to ensure ongoing safety and efficacy (Albrecht et al., 2024).

Total Product Lifecycle Considerations

- **Post-Market Surveillance:** The FDA should review the manufacturer's plans for post-market surveillance and ongoing monitoring of the ACD device, including mechanisms for:
    - **Collecting User Feedback:** Gather feedback from clinicians using the device to identify any usability issues, performance concerns, or potential areas for improvement (**Tierney et al., 2024, Misurac et al., 2023).**
    - **Tracking Adverse Events:** Establish procedures for reporting and analyzing any adverse events potentially related to the ACD device, such as inaccurate documentation leading to inappropriate clinical decisions.
    - **Real-World Data Analysis:** Analyze real-world data collected from patient encounters to monitor the device's performance over time, detect emerging trends, and identify areas for model refinement or retraining (Galloway et al., 2024).

By considering this comprehensive evidence throughout the total product lifecycle, the FDA can play a vital role in ensuring that generative AI-enabled ACD devices are developed and deployed responsibly, improving patient care and clinician well-being.

Risks and Transparency in Generative AI for Ambient Clinical Documentation

Generative AI, compared to non-generative AI, introduces new and unique usability risks in Ambient Clinical Documentation (ACD). This necessitates conveying specific information to healthcare professionals, patients, and caregivers to improve transparency and control these risks.

New and Unique Usability Risks

- **Hallucinations and Factual Errors:** Generative AI, particularly large language models like ChatGPT-4, tends to generate "hallucinations," meaning it can create medical notes with fabricated information, omission errors, and inaccurate details (**Misurac et al., 2023,** Biswas et al., 2024). This is a unique risk compared to traditional, rule-based systems which are less prone to inventing information, although they might struggle with accurately capturing all the nuances of a conversation.
- **Verbose and Irrelevant Information:** Generative AI can sometimes struggle to distinguish between important clinical information and irrelevant conversation, leading to lengthy notes cluttered with unnecessary details (Bundy et al., 2024). This can make it challenging for clinicians to efficiently review and extract key information from the generated notes. Non-generative systems often rely on structured input or templates, limiting the scope for irrelevant information to be included.
- **Bias and Fairness:** Generative AI models are trained on vast amounts of data, and if this data contains biases, the model can perpetuate and even amplify these biases in the generated clinical notes (**Yim et al., 2023, Misurac et al., 2023)**. This can lead to disparities in healthcare delivery and decision-making. Non-generative systems, with their reliance on more explicit rules and predefined logic, are less susceptible to inheriting biases from training data in the same way.
- **Lack of Transparency and Explainability:** The decision-making process of generative AI can be opaque, making it difficult to understand why the model generated specific outputs (**Biswas et al., 2024,** Nahar et al., 2023). This lack of transparency can make it challenging for clinicians to trust the AI's output and identify potential errors or biases. Non-generative systems often operate with more transparent logic, making it easier to understand how a specific output was generated.

Transparency and Control Measures

- **Inform Healthcare Professionals:**
    - **Explain Hallucinations and Errors:** Clearly communicate the potential for hallucinations and factual errors in AI-generated notes, emphasizing the need for careful review and validation (**Tierney et al., 2024)**.

  - **Training on System Limitations:** Provide training to healthcare professionals on the capabilities and limitations of generative AI in clinical documentation (**Tierney et al., 2024)**. This includes instruction on how to effectively use the system, recognize potential errors, and edit the generated notes for accuracy.

- **Inform Patients and Caregivers:**
  - **Explain AI's Role in Documentation:** Provide clear and concise information to patients about the use of AI in generating clinical notes. Explain that the AI assists with documentation, but a human clinician reviews and validates the information (**Tierney et al., 2024)**.
  - **Obtain Informed Consent:** Establish transparent procedures for obtaining informed consent from patients regarding the use of AI in their care and the handling of their data (**Tierney et al., 2024)**. This should include information about the potential benefits and risks associated with AI-driven documentation.
  - **Address Privacy Concerns:** Openly address patient concerns about data privacy and security related to the use of AI in healthcare (Nahar et al., 2023). Explain how data is de-identified, protected, and used responsibly to build trust and ensure patient comfort with the technology.

- **Promote System-Level Transparency:**
  - **Regular Performance Reports:** Healthcare systems should provide regular reports on the performance of generative AI systems used in clinical documentation. These reports should include metrics related to accuracy, completeness, and bias, allowing for ongoing monitoring and evaluation (**Tierney et al., 2024)**.
  - **Mechanisms for Feedback and Error Reporting:** Establish user-friendly channels for clinicians to provide feedback on the AI's performance and report potential errors or biases (**Biswas et al., 2024)**. This feedback loop is crucial for continuous improvement of the system and addressing user concerns.

- **Enhance System Design:**
  - **Human Oversight and Validation:** Emphasize the importance of human oversight in the clinical documentation process. Generative AI should be positioned as a tool to assist clinicians, not replace them, with a clear understanding that human review and validation are essential (**Biswas et al., 2024)**.
  - **Uncertainty Estimation:** Incorporate features that provide uncertainty estimates alongside AI-generated outputs. This can help clinicians prioritize their review efforts and focus on areas where the AI might be less confident in its output (**Tierney et al., 2024)**.
  - **Explainable AI Features:** Integrate explainable AI (XAI) features that provide insights into the model's decision-making process. This can increase clinician trust in the system by allowing them to understand the rationale behind specific outputs (Nahar et al., 2023).

By implementing these transparency and control measures, healthcare organizations can mitigate the unique risks associated with generative AI in Ambient Clinical Documentation and foster trust among healthcare professionals, patients, and caregivers.

Performance Metrics for Ambient Clinical Documentation

Given the complexity of generative AI for ambient clinical documentation (ACD), choosing suitable performance metrics is crucial.

These metrics should go beyond simply evaluating the quality of the generated text and assess how well the system supports the intended clinical workflow.

Here are some prospective performance metrics that are particularly informative for evaluating generative AI in ACD:

1. Accuracy and Completeness of Clinical Information

- **MedCon F1 Score:** Measures the accuracy and consistency of clinical concepts captured in the generated notes by calculating the F1-score for matching UMLS concept sets between the AI's output and reference clinical notes. This is particularly relevant for ACD as it directly assesses the AI's ability to extract and represent meaningful medical information (**Yim et al., 2023)**

- **Modified Physician Documentation Quality Instrument (PDQI-9):** The PDQI-9, originally designed to evaluate human-generated notes, can be adapted to assess the quality of AI-generated notes. This adaptation involves removing the "up to date" domain and adding domains for assessing "freedom from hallucinations" and "bias" to account for the specific challenges posed by generative AI (**Tierney et al., 2024)**.

- **Error Analysis by Type:** A detailed analysis of the types of errors made by the generative AI system can provide valuable insights for system improvement. This includes classifying errors as hallucinations, omissions, factual inaccuracies, and formatting inconsistencies. Categorizing errors helps pinpoint specific areas where the model needs refinement, allowing for targeted interventions and improved training data (**Yim et al., 2023, Misurac et al., 2023, Bundy et al., 2024)**.

2. Efficiency and Workflow Integration

- **Time Spent on Documentation:** Measuring the time clinicians spend reviewing, editing, and finalizing AI-generated notes provides insights into the system's efficiency. A significant reduction in documentation time compared to traditional methods suggests the AI effectively supports clinical workflow (**Tierney et al., 2024)**.

- **Time to Note Closure:** The time it takes for a clinical note to be finalized after a patient encounter is another important metric. Faster note closure indicates improved efficiency in the documentation process and can have positive implications for patient care coordination and billing (Balloch et al., 2024, Albrecht et al., 2024).

- **Impact on "Pajama Time":** This metric assesses the amount of time clinicians spend working on documentation outside of regular work hours. A reduction in "pajama time" indicates the AI is helping clinicians achieve a better work-life balance and reduce burnout (**Tierney et al., 2024)**.

- **Adoption Rate and Usage Patterns:** Tracking the adoption rate of the ACD system and analyzing how clinicians use it in their daily practice provides valuable information about the system's usability and acceptance. High adoption and sustained usage suggest that clinicians find the system beneficial and are integrating it into their workflows (**Tierney et al., 2024**).

3. User Experience and Satisfaction

- **Clinician Satisfaction Surveys:** Regular surveys to assess clinician satisfaction with the ACD system can provide feedback on its usability, effectiveness, and impact on their workflow. This feedback is essential for iterative system improvement and addressing user concerns (**Tierney et al., 2024,** Albrecht et al., 2024, **Misurac et al., 2023)**

- **Patient Satisfaction Surveys:** While the primary users of ACD are clinicians, it is important to also assess patient perceptions of the technology. Surveys can gauge patient comfort levels with AI-assisted documentation, perceived impact on the quality of care, and concerns about data privacy. Gathering patient feedback helps address ethical considerations and ensure patient-centric design. (**Tierney et al., 2024, Misurac et al., 2023)**

4. Multimodal Performance Metrics

As ACD technology evolves, it is likely to incorporate multimodal inputs and outputs, such as images, videos, and sensor data, in addition to text and audio. This will require developing new performance metrics that can effectively evaluate the system's ability to integrate and interpret information from multiple sources.

Here are some considerations for multimodal performance metrics:

- **Cross-Modal Consistency:** Metrics should assess the consistency of information extracted and represented across different modalities. For example, does the system accurately link a patient's verbal description of symptoms with corresponding findings in medical images?

- **Multimodal Integration:** Evaluate how well the system combines insights from different modalities to generate a comprehensive and coherent clinical note. Does the AI effectively synthesize information from audio, video, and sensor data to provide a holistic view of the patient's condition?

- **Task-Specific Performance:** Performance metrics should be tailored to the specific clinical tasks that the ACD system is designed to support. For example, if the system is used to assist with diagnosis, metrics should focus on diagnostic accuracy and the AI's ability to identify relevant clinical features from multimodal data.

Risk Management of Generative AI in Ambient Clinical Documentation

Generative AI is creating new opportunities and applications for medical devices, particularly in the realm of ambient clinical documentation. This technology offers a promising solution to the growing burden of documentation for clinicians, but it also introduces new risks that need to be addressed through appropriate controls.

New Opportunities Enabled by Generative AI

- **Reduced Documentation Burden and Increased Clinician Efficiency:** By automating the process of generating clinical notes from patient-clinician conversations, generative AI can significantly reduce the time clinicians spend on documentation, freeing them up to focus more on patient care (Tierney et al., 2024, Bundy et al., 2024, Nahar et al., 2024). This can lead to increased efficiency, improved productivity, and potentially even allow clinicians to see more patients (Doshi et al., 2024).

- **Enhanced Documentation Quality and Completeness:** Generative AI can leverage the vast amounts of data it is trained on to create more complete and accurate clinical notes, potentially exceeding the capabilities of human scribes in capturing relevant details (Doshi et al., 2024, Tierney et al., 2024). This can lead to improved communication among healthcare providers, better-informed clinical decisions, and potentially even better patient outcomes.

- **Improved Patient-Clinician Interactions:** By reducing the need for clinicians to focus on data entry during patient encounters, generative AI can enable more meaningful and engaged interactions between clinicians and patients (Tierney et al., 2024, Bundy et al., 2024). This can lead to increased patient satisfaction, enhanced trust, and a more positive overall care experience.

- **Direct EHR Integration:** Ambient AI scribe tools can be directly integrated into the electronic health record (EHR), further streamlining clinical workflows, and reducing the potential for errors or inconsistencies (Galloway et al., 2024). This integration can facilitate seamless data flow, enhance data accessibility, and support more efficient and effective care delivery.

New Controls Needed to Mitigate Risks

1. Governance:

- **Establish Clear Regulatory Frameworks:** Regulatory bodies need to develop clear guidelines and standards for the development, validation, and deployment of generative AI-enabled medical devices (Biswas et al., 2024). This includes defining clear requirements for data privacy and security, accuracy, and reliability, explainability, and human oversight.

- **Develop Robust Internal Governance Policies:** Healthcare organizations need to establish their own internal governance policies for the use of generative AI in clinical practice. These policies should address data governance, algorithm bias, transparency and explainability, user training, and performance monitoring (Gerke et al., 2020).

- **Ensure Ethical Use:** Establish ethical guidelines for the use of generative AI, focusing on patient safety, privacy, autonomy, and fairness. This includes addressing potential biases in the AI system and ensuring that it is used in a way that aligns with the values and principles of the healthcare profession (Nahar et al., 2024).

2. Training:

- **Comprehensive Clinician Training:** Provide thorough training to clinicians on the capabilities and limitations of the AI system. This includes educating them on how to use the system effectively, how to interpret its outputs, how to identify potential errors or biases, and how to provide feedback for system improvement (Tierney et al., 2024, Liu et al., 2024)

- **Patient Education:** Inform patients about the use of generative AI in their care, addressing their potential concerns about privacy, security, and the role of human oversight. Ensure that patients understand how the AI system works and how it will be used to document their encounters (Tierney et al., 2024).

3. Feedback Mechanisms:

- **Establish Multifaceted Feedback Channels:** Create multiple channels for clinicians and patients to report errors, provide suggestions, and share their experiences with the AI system. This can include in-app feedback forms, dedicated email addresses, user forums, and regular surveys (Biswas et al., 2024).

- **Develop a Systematic Process for Feedback Analysis and Incorporation:** Establish a clear process for analyzing feedback, identifying trends, and incorporating this information into system updates and improvements (Biswas et al., 2024). This ensures that user feedback is actively used to enhance the AI's accuracy, relevance, and reliability over time.

4. Real-World Performance Evaluation:

- **Continuous Monitoring of Key Performance Metrics:** Implement robust systems for continuous monitoring of the AI system's performance in real-world clinical settings, focusing on accuracy, completeness, bias, hallucinations, and impact on clinician workflows and patient safety (Tierney et al., 2024).

- **Regular Audits and Evaluations:** Conduct regular audits and independent evaluations of the AI system to assess its compliance with regulatory requirements, internal policies, and ethical guidelines. This helps ensure that the AI system is used safely, responsibly, and effectively (Biswas et al., 2024).

- **Longitudinal Studies:** Conduct longitudinal studies to assess the long-term impacts of the AI system on patient outcomes, clinician well-being, and healthcare system efficiency. This will provide valuable insights into the effectiveness and sustainability of the technology and inform future development and policy decisions (Albrecht et al., 2024, Misurac et al., 2024).

- **Focus on Adaptive Learning and Drift Detection:** As generative AI systems continuously learn from new data; it is crucial to monitor for drift—the potential for performance decline due to changes in data patterns or clinical practices. Implement mechanisms for early drift detection and develop strategies to mitigate its impact on accuracy and reliability (Biswas et al., 2024).

By implementing these controls, healthcare organizations can effectively mitigate the risks associated with generative AI in ambient clinical documentation, paving the way for the safe and beneficial adoption of this technology. This will allow for the realization of generative AI's potential to improve clinician well-being, enhance patient care, and transform the healthcare landscape.

Critical Aspects of Post-Market Monitoring and Evaluation for Ambient Clinical Documentation

Post-market performance monitoring and evaluation are particularly important for generative AI-enabled ambient clinical documentation devices because:

- They are **non-deterministic** meaning their outputs can vary even with the same input.
- They undergo **continuous adjustment** based on new data, user interactions, and changing conditions after deployment, meaning that their behavior can change over time.

Here are some critical aspects / metrics of post-market real world evaluation that will be crucial for maintaining the safety and effectiveness of the device:

1. Accuracy and Completeness:

- **Word Error Rate (WER):** WER measures the accuracy of the AI's transcription by comparing it to a human-generated reference transcript. A lower WER indicates higher accuracy.
- **Clinical Concept Extraction Accuracy:** This assesses the AI's ability to identify and extract relevant clinical concepts from the conversation, such as symptoms, diagnoses, and treatments. The accuracy can be measured by comparing the AI's extracted concepts to a gold-standard set of concepts identified by human experts.
- **Completeness Score:** This metric evaluates how thoroughly the AI system documents all relevant information discussed during the patient encounter. It can be determined by comparing the content of the AI-generated note to a comprehensive reference note created by a human clinician, considering all essential clinical elements.

2. Relevance and Appropriateness:

- **Clinical Note Quality Assessment:** Use standardized instruments, like a modified version of the Physician Documentation Quality Instrument (PDQI-9), to assess the relevance, accuracy, completeness, and organization of the AI-generated notes (Tierney et al., 2024). This assessment should include an evaluation of whether the note accurately reflects the patient encounter and is appropriate for the intended clinical purpose.
- **Physician Review and Feedback:** Physicians should review the AI-generated notes for accuracy, relevance, and completeness. Feedback from physicians can be used to improve the AI system's performance and to ensure that the notes are clinically appropriate (Tierney et al., 2024).

3. Bias Detection and Mitigation:

- **Demographic Analysis:** Analyze the performance of the AI system across different patient demographics, such as age, gender, race, ethnicity, and socioeconomic status. Look for disparities in accuracy, completeness, or relevance of the generated documentation, which could indicate potential bias.
- **Non-Lexical Conversational Sounds (NLCS) Recognition:** Monitor the AI system's performance in recognizing and interpreting non-lexical conversational sounds (NLCS), such as "Mm-hm" and "Uh-uh." These sounds often convey important information, and

the AI should be able to accurately transcribe and interpret them to avoid bias or missing important information (Tran et al., 2023).

- **Bias Mitigation Techniques:** Implement bias mitigation techniques in the AI development and training process. These can include using algorithms designed for fairness, incorporating fairness constraints into the training process, or employing data augmentation techniques to balance the data set and reduce demographic bias.

4. Hallucination Detection and Prevention:

- **Fact Verification:** Implement mechanisms to verify the accuracy of the information generated by the AI system, especially when it comes to medical facts, diagnoses, or treatment recommendations. This could involve cross-referencing the AI's output with established medical databases or knowledge bases.

- **Confidence Scores:** The AI system could provide confidence scores for the information it generates. Low confidence scores could flag potential hallucinations for further review by a clinician.

- **Human Review:** Maintain a human-in-the-loop approach, where clinicians review and validate the AI's output, especially in cases where the AI system generates information that is not directly supported by the patient-clinician conversation.

5. User Experience and Satisfaction:

- **Clinician Surveys and Interviews:** Conduct regular surveys and interviews with clinicians to assess their satisfaction with the AI system, its usability, and its impact on their workflow and documentation burden. Gather feedback on specific issues, such as the accuracy of the generated notes, the ease of editing the notes, and the overall time savings. (**Tierney et al., 2024,** Albrecht et al., 2024, **Misurac et al., 2023)**

- **Patient Surveys and Feedback:** Collect feedback from patients on their experiences with the AI-assisted documentation process. This could include questions about their comfort level with the technology, their perceptions of the clinician's engagement during the visit, and any concerns about privacy or data security (**Tierney et al., 2024, Misurac et al., 2023)**

6. Performance Monitoring of External AI Services:

- **Service Level Agreements (SLAs):** Establish clear service level agreements (SLAs) with the providers of the external consumer-grade AI services. The SLAs should specify the expected performance levels, such as accuracy, latency, and availability, and include penalties for failing to meet those standards.

- **Independent Evaluation:** Conduct independent evaluations of the external AI services to verify their accuracy, reliability, and compliance with relevant standards. This could involve using benchmark datasets or engaging third-party testing organizations.

7. Adaptability and Continuous Improvement:

- **Monitoring for Drift:** Track the performance of the AI system over time to detect any signs of drift, where the accuracy or relevance of the generated documentation declines due to changes in language use, clinical practices, or patient demographics.

- **Model Retraining and Updates:** Implement mechanisms for retraining the AI models or updating the algorithms to adapt to new data, changing conditions, and feedback from clinicians and patients. This ensures that the system remains accurate and relevant over time.
- **Iterative Development:** Foster an iterative development process, where feedback from clinicians, patients, and performance data is continuously used to improve the AI system, address identified issues and enhance its usability and effectiveness.

By implementing these methods and metrics, developers and healthcare organizations can effectively monitor and evaluate the post-market performance of generative AI-enabled ambient clinical documentation devices. Consistent monitoring, data analysis, and a commitment to continuous improvement are essential for ensuring that these systems maintain accuracy, relevance, and reliability while adapting to new data and evolving clinical practices.

References

Misurac, J., Knake, L.A. and Blum, J.M., 2024. Impact of Ambient Artificial Intelligence Notes on Provider Burnout. *medRxiv*, pp.2024-07.

Tierney, A.A., Gayre, G., Hoberman, B., Mattern, B., Ballesca, M., Kipnis, P., Liu, V. and Lee, K., 2024. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst Innovations in Care Delivery*, *5*(3), pp.CAT-23.

Bundy, H., Gerhart, J., Baek, S., Connor, C.D., Isreal, M., Dharod, A., Stephens, C., Liu, T.L., Hetherington, T. and Cleveland, J., 2024. Can AI-Facilitated Clinical Documentation Alleviate the Administrative Loads of Physicians? *Journal of general internal medicine*, pp.1-6.

Nahar, J.K. and Kachnowski, S., 2023. Current and potential applications of ambient artificial intelligence. *Mayo Clinic Proceedings: Digital Health*, *1*(3), pp.241-246.

Biswas, A. and Talukdar, W., 2024. Intelligent Clinical Documentation: Harnessing Generative AI for Patient-Centric Clinical Note Generation. *arXiv preprint arXiv:2405.18346*.

Balloch, J., Sridharan, S., Oldham, G., Wray, J., Gough, P., Robinson, R., Sebire, N.J., Khalil, S., Asgari, E., Tan, C. and Taylor, A., 2024. Use of an ambient artificial intelligence tool to improve quality of clinical documentation. *Future Healthcare Journal*, *11*(3), p.100157.

Doshi, G.K., Jensen, T.L., Graziano, A., Enenmoh, C. and Lindsey, J., 2024. Use of ambient AI scribing: Impact on physician administrative burden and patient care.

Liu, T.L., Hetherington, T.C., Stephens, C., McWilliams, A., Dharod, A., Carroll, T. and Cleveland, J.A., 2024. AI-Powered Clinical Documentation and Clinicians' Electronic Health Record Experience: A Nonrandomized Clinical Trial. *JAMA Network Open*, *7*(9), pp. e2432460-e2432460.

Albrecht, M., Shanks, D., Shah, T., Hudson, T., Thompson, J., Filardi, T., Wright, K., Ator, G. and Smith, T.R., 2024. Enhancing Clinical Documentation Workflow with Ambient Artificial Intelligence: Clinician Perspectives on Work Burden, Burnout, and Job Satisfaction. *medRxiv*, pp.2024-08.

Nuance, Automatically Document Care with the Dragon Ambient EXperience.

Yim, W.W., Fu, Y., Ben Abacha, A., Snider, N., Lin, T. and Yetisgen, M., 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, *10*(1), p.586.

Tran, B.D., Latif, K., Reynolds, T.L., Park, J., Elston Lafata, J., Tai-Seale, M. and Zheng, K., 2023. "Mm-hm," "Uh-uh": are non-lexical conversational sounds deal breakers for the ambient clinical documentation technology? *Journal of the American Medical Informatics Association*, *30*(4), pp.703-711.

Galloway, J.L., Munroe, D., Vohra-Khullar, P.D., Holland, C., Solis, M.A., Moore, M.A. and Dbouk, R.H., 2024. Impact of an Artificial Intelligence-Based Solution on Clinicians' Clinical Documentation Experience: Initial Findings Using Ambient Listening Technology. *Journal of General Internal Medicine*, pp.1-3.

Gerke, S., Yeung, S., and Cohen, I.G., 2020. Ethical and legal aspects of ambient intelligence in hospitals. *Jama*, *323*(7), pp.601-602.