university of groningen

campus fryslân

A Comparative Evaluation of Closed- and Open-Vocabulary ASR Systems for the Recognition of Dutch Healthcare Terms

Shiran Sun

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

A Comparative Evaluation of Closed- and Open-Vocabulary ASR Systems for the Recognition of Dutch Healthcare Terms

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr. Jan-Willem van Leussen, (Gerimedica B.V)
and
Dr. Joshua K. Schäuble, Assistant Professor (Voice Technology, University of Groningen)
with the second reader being
Supervisor 2's title and name (Voice Technology, University of Groningen)

Shiran Sun (S5878594)

June 11, 2025

Acknowledgements

I would like to extend my deep gratitude to my supervisor at Gerimedica, Dr. Jan-Willem van Leussen, for his invaluable technical support and access to essential data resources. His insightful suggestions and innovative ideas throughout the writing process have been instrumental in shaping the direction and quality of this thesis.

I am also sincerely grateful to Dr. Matt Coler for his early contributions. His support during the initial stages laid a strong foundation for this research.

My thanks also go to all the teachers in the Voice Technology program, such as Dr. Phat Do, Dr. Vass Verkhodanova, Dr. Joshua K. Schäuble, and Dr. Shekhar Nayak, thanks for their advice, encouragement, and guidance over the past year.

I would like to thank Andrew Ng for his inspiring and well-structured online machine learning courses, which made machine learning accessible and encouraging for beginners like me.

I am deeply thankful to my family and friends for their unconditional support and encouragement. Their belief in me and their continuous emotional support have been a source of strength and motivation throughout this academic journey.

Finally, I would like to thank myself, for making one brave decision after another. From the moment I considered studying abroad to completing the Voice Technology program, I have experienced countless struggles and uncertainties. I feared I wouldn't be able to keep up with technical or programming demands, and I often questioned whether stepping away from my original career path was the right choice. Looking back, I made it through, better than I had expected. This journey has taught me that perhaps I do have more potential than I once believed. Rather than questioning the meaning of every choice, it may be more important to stay present and keep moving forward, because perseverance itself holds meaning. And perhaps, in the end, it matters more to follow my heart than to follow society's timeline.

Abstract

Automatic Speech Recognition (ASR) technology is becoming more prevalent in clinical settings, but the performance of closed- and open-vocabulary ASR models on domain-specific speech in healthcare is not well studied. In this paper, we present a comparative evaluation of an ASR systems, operating in closed- and open-vocabulary settings, for the recognition of Dutch clinical terminology. We consider a closed vocabulary Kaldi TDNN model and an open vocabulary Pruned RNN-Transducer (K2-RNN-T), both trained on more than 1000 hours of Dutch speech, consisting of 12 hours domain-specific training data. We evaluate both systems on a professionally transcribed Dutch medical consultation corpus containing over 8000 utterances, using both standard evaluation metrics (WER, CER), domain-specific evaluation metrics (Medical WER and CER), and term-level evaluation (precision, recall, F1 score).

We find that in general the closed vocabulary model obtains better recognition results for structured medical terms, such as diseases and drug names: the precision and F1 score is higher while the Medical WER and CER is lower. The open vocabulary model, on the other hand, has better recall and general transcription accuracy and seems more flexible in handling morphologically varied or unknown terms. Evaluation is performed through SNOMED CT and spaCy-based Named Entity Recognition (NER) to extract clinical and contextual entities from the transcription.

This study also uncovers notable error types such as phonetic substitutions, semantic approximations, and truncations, each with distinct clinical implications. Results highlight a trade-off between lexical accuracy and adaptability: while the closed-vocabulary model ensures stability for structured content, the open-vocabulary model captures a broader lexical range, including personal and brand names often missed by fixed lexicons.

This work shows the strengths of ASR approaches in both closed- and open-vocabulary settings and motivates task-specific optimisation in medical speech applications. The evaluation framework presented here can be adapted for other low-resource languages and specialised domains.

Contents

1 Introduction

Automatic Speech Recognition (ASR) has gone through significant developments in the recent decades, advancing from statistical to deep learning-based models. Previous systems utilized fixed, closed-vocabularies, which hindered their capability in dealing with novel or rare words. The open-vocabulary methods came into picture as an alternative, especially open-vocabulary modeling methods such as byte-pair encoding (BPE) and wordpiece segmentation, which build words from sub-units to handle out-of-vocabulary words more effectively (Sennrich, Haddow, & Birch, 2016). These methods are widely adopted in modern ASR toolkits and are often integrated into powerful sequence modeling frameworks, including attention-based encoder-decoder models (Chan, Jaitly, Le, & Vinyals, 2015) and Transformer architectures (Vaswani et al., 2017). These advances have significantly improved ASR performance in general speech domains. However, general domain pre-training model does not always transfer adequately to the clinical domain due to its highly specialized language (Laparra, Mascio, Velupillai, & Miller, 2021). These issues highlight the need for models that are better adapted to the linguistic and contextual complexity of clinical speech.

Using the correct medical words is very important for ASR systems in healthcare. It helps keep care safe, fast, and focused on the patient. For example, if the system hears “no known allergies” but writes “known allergies,” the patient might get the wrong treatment or feel worried. Even small mistakes, like writing the wrong drug name, dose, or test result, can cause problems in communication between doctors, patients, and hospital staff. Medical reports are important for sharing information. If there are mistakes in them, it can reduce trust and make doctors look unprofessional. This might also cause harm to the patient.(Poder, Fisette, & Déry, 2018).

Given its critical role in ensuring safe and effective healthcare delivery, it is essential to understand how current ASR technologies perform in medical contexts. Most research today tries to make ASR more accurate using general tests, but it often ignores the special problems of understanding medical speech. Although open-vocabulary models have shown promise in handling out-of-vocabulary (OOV) terms in broader contexts (Prabhavalkar, Hori, Sainath, Schlüter, & Watanabe, 2023), there is limited comparative analysis between closed- and open-vocabulary approaches in medical ASR tasks. This gap is especially important given the high stakes of transcription accuracy in healthcare. Studies report that Word Error Rates (WER) in medical ASR remain significantly higher than in other domains, ranging from 7.4% to 38.72% for general medical terms and 5.21% to 9% for specialized vocabulary (Blackley, Huynh, Wang, Korach, & Zhou, 2019). To address the gap in domain-specific ASR evaluation and the lack of comparative studies on vocabulary approaches in medical settings, this study investigates Dutch clinical speech data from the healthcare domain. It systematically compares the performance of closed- and open-vocabulary ASR models across different categories of medical terminology, combining quantitative metrics with qualitative error analysis. The aim is to support the development of ASR systems that are better aligned with the practical needs of real-world healthcare environments.

Now that the motivation for this research has been presented, the structure of this thesis is as follows:

- Section 1.1 presents the research questions and hypotheses.

- Section 2 reviews literature on ASR systems in healthcare, Dutch-specific speech challenges, clinical terminology and NER, and evaluation metrics, highlighting gaps in Dutch medical ASR research.
- Section 3 describes the methodological approach and technical framework, detailing the ASR models (Kaldi-TDNN and K2-RNN-T), the use of SNOMED CT and spaCy for term extraction, and the design of the term-matching and evaluation method.
- Section 4.3 presents and analyzes the results across both general and medical-specific metrics, with breakdowns by term category and error type.
- Section 5 discusses the implications of the findings in relation to the research questions and hypothesis, emphasizing the trade-offs between model architectures and evaluating the limitations of the current study.
- Section 6 summarizes the thesis and synthesizes the main contributions, followed by a discussion of future research directions.

1.1 Research Questions and Hypotheses

In light of the preceding discussion, this research addresses the following question:

> **To what extent do closed- and open-vocabulary ASR models differ in recognizing clinical terminology and named entities in Dutch medical speech, as evaluated through multiple quantitative metrics and qualitative error analysis?**

This main question can be broken down into the following sub-questions:

- How can clinical terms and named entities be effectively extracted from medical transcripts for evaluation, and which methods are appropriate for identifying different term categories?
- How do the two ASR models compare in recognizing these terms, based on evaluation metrics and error patterns, and do they show clear strengths or limitations in specific categories?

Our hypothesis is that while open-vocabulary models generally achieve lower Word Error Rates (WER) in overall transcription tasks, closed-vocabulary models demonstrate more stable and controllable performance in domain-specific keyword recognition, particularly in terms of a more reliable precision-recall tradeoff and overall error control in structured terminology.

This hypothesis is supported by results from (Chiu et al., 2017), where a closed-vocabulary CTC model achieved 92% precision and 86% recall on medical phrase recognition, although it had a higher overall WER of 20.1%. In contrast, the open-vocabulary LAS model achieved a lower overall WER of 18.3% and a recall of 98.2% on drug name recognition (precision was not reported). In

addition, Chiu et al. also observed that the CTC model performed better on speech from doctors, which usually contains more structured and term-heavy language. These results suggest that while open-vocabulary systems may be more flexible, closed-vocabulary models provide higher reliability in structured medical contexts.

2 Literature Review

The increasing integration of ASR technologies into clinical workflows has prompted a growing body of research investigating their effectiveness in domain-specific contexts. This literature review situates the current study within four key strands of prior work: (1) the comparative development of closed- and open-vocabulary ASR systems, with a focus on their applicability to healthcare speech; (2) the unique linguistic and acoustic challenges posed by the Dutch language, including compounding, code-switching, and low-resource constraints; (3) the role of clinical terminology systems and named entity recognition (NER) in medical transcription accuracy; and (4) the evolution of evaluation metrics tailored to the needs of domain-sensitive ASR, such as Medical WER and entity-level F1 scores. Together, these dimensions provide a conceptual and methodological foundation for analyzing how different ASR paradigms perform in capturing critical clinical content in Dutch medical speech.

2.1 Search Strategy and Selection Criteria

Keywords by Topic:

- **Topic 1 – Speech Technology:** “automatic speech recognition”, “ASR”
- **Topic 2 – Domain Focus:** “medical terminology”, “healthcare speech”
- **Topic 3 – Vocabulary Modeling:** “closed-vocabulary”, “open-vocabulary”, “lexicon-based”, “subword-based”

Exclusion Criteria:

- Studies without experimental results
- General ASR research unrelated to medical or specialized terminology
- Research focusing on ASR algorithms

2.2 ASR Systems in Healthcare: Closed vs. Open Vocabulary ASR Systems

2.2.1 Closed Vocabulary ASR Systems

Closed-vocabulary ASR systems are traditionally based on a lexicon-driven architecture, where only a predefined set of words in a fixed vocabulary can be recognized. These systems typically follow a hybrid design comprising three key components: an acoustic model (AM), a pronunciation lexicon,

and a statistical language model (LM). The pronunciation lexicon maps each word in the vocabulary to its corresponding phoneme sequence, enabling accurate decoding through a constrained search space.

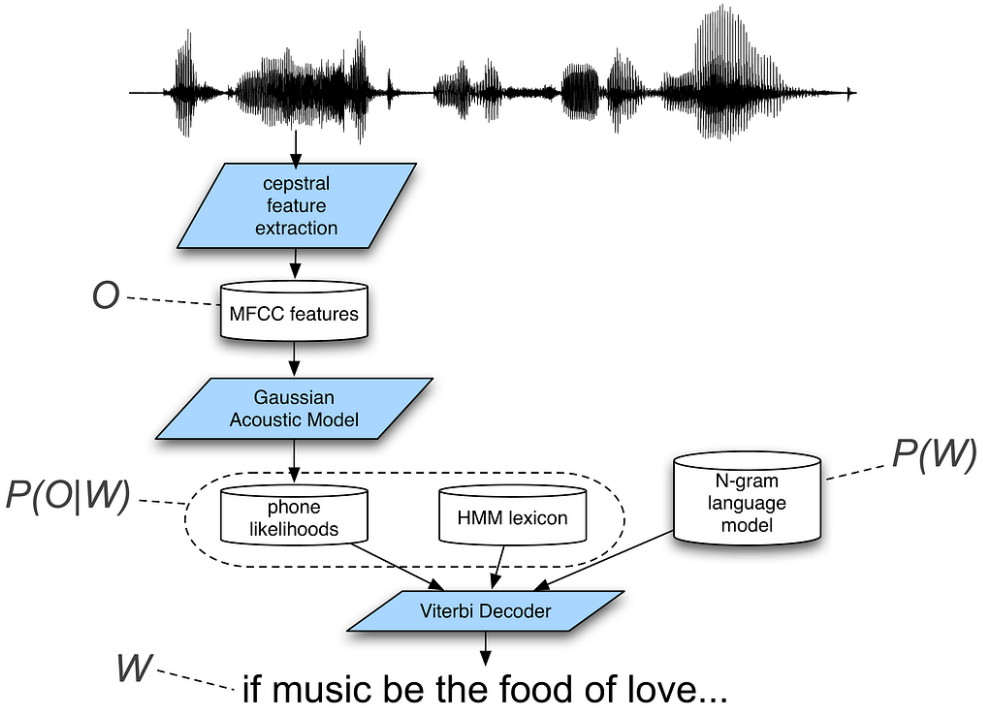

Figure 1: Hybrid ASR architecture[1] showing feature extraction, acoustic model, lexicon, and language model. The AM estimates $P(O \mid W)$: the likelihood of features O given word sequence W; the LM provides $P(W)$, the prior probability of W.

Decoding in these systems is typically performed using a Weighted Finite-State Transducer (WFST)-based architecture, where individual transducers representing different components are composed into a single search graph. The decoding graph, known as HCLG, integrates the Hidden Markov Model (H), context-dependency (C), lexicon (L), and grammar or language model (G), and is constructed as follows:

$$HCLG = \min\left(\det\left(H \circ \det\left(C \circ \det(L \circ G)\right)\right)\right)$$

Here, $\circ$ denotes composition, det denotes determinization, and min represents minimization to optimize the graph size and search efficiency. This precompiled structure allows for highly efficient decoding, especially in domains with limited vocabulary variability (Garg, 2019).

To estimate the acoustic likelihoods used in decoding, modern closed-vocabulary systems often employ Time-Delay Neural Networks (TDNNs) as the acoustic model. TDNNs are particularly ef-

[1] Adapted from: `https://medium.com/intel-student-ambassadors/attention-in-end-to-end-automatic-speech-recognition-9f9e42718d21`

fective at modeling long-range temporal dependencies by capturing contextual information across multiple time frames. Figure 2 illustrates the hierarchical context structure of a typical TDNN architecture used in this setting.

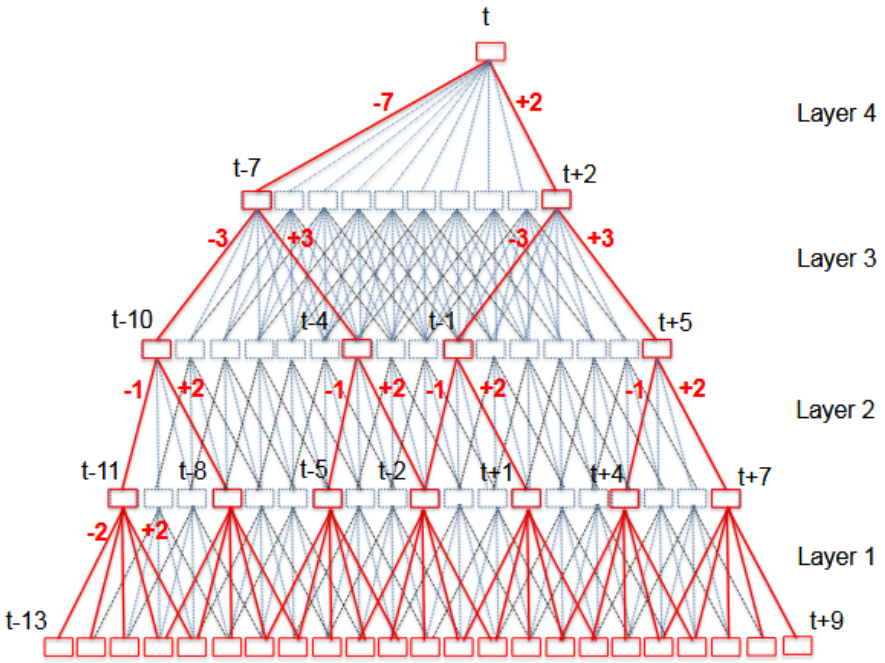

Figure 2: TDNN architecture[2]showing how each layer learns from different temporal contexts, enabling efficient modeling of long-range dependencies

Systems such as the Kaldi TDNN-HMM pipeline and IBM Watson's closed-vocabulary ASR are representative of this architecture. Earlier systems like CMU Sphinx and Google's hybrid ASR frameworks also used WFST-based decoding with fixed-vocabulary lexicons. One of the key advantages of this modular structure is its adaptability using only text data. Since the lexicon and language model are external to the acoustic model, domain-specific terms such as clinical terminology can be integrated without retraining the entire system. This makes closed-vocabulary ASR a good choice for organized settings and limited-resource situations.(Khassanov, 2020).

However, a major drawback of closed-vocabulary ASR is its inability to handle OOV terms. This limitation is particularly problematic in dynamic domains like healthcare, where new drug names, rare diseases, or proper nouns frequently appear. While closed-vocabulary systems provide high recognition accuracy for known terms, they inherently lack the flexibility to adapt to novel or evolving vocabularies.

[2](Peddinti, Povey, & Khudanpur, 2015)

2.2.2 Open Vocabulary ASR Systems

Open-vocabulary ASR systems are designed to recognize words beyond a fixed lexicon, making them well-suited for domains with frequent OOV terms, such as medical terminology or named entities. Unlike classical ASR systems that rely on predefined phoneme dictionaries and modular pipelines, which includes separate acoustic models, lexicons, and language models, open-vocabulary models typically adopt end-to-end (E2E) architectures that map audio directly to subword units (Zhou, Zeineldeen, Zheng, Schlüter, & Ney, 2021). This structure removes the need for hand-crafted pronunciation dictionaries, allowing systems to learn directly from audio-text pairs. As a result, training and deployment are often simpler and more flexible across domains and languages.

In the context of speech recognition, prominent open-vocabulary architectures include Connectionist Temporal Classification (CTC), Attention-based Encoder-Decoder (AED) and Recurrent Neural Network Transducer (RNN-T) models (Figure 3). These architectures are widely adopted in end-to-end ASR systems due to their ability to handle subword-level decoding without fixed vocabularies. These architectures differ in their alignment strategies: CTC and RNN-T use explicit alignment, while AED uses implicit attention-based alignment, but they all support subword decoding without reliance on fixed lexicons. Representative systems include Wav2Vec2 + CTC (Baevski, Zhou, Mohamed, & Auli, 2020), Conformer + RNN-T (Gulati et al., 2020), and transformer-based AED models (Karita et al., 2019). However, it is important to note that CTC's classification as open-vocabulary depends on the decoding strategy: while CTC can function in an open-vocabulary setting with subword or character-based beam search decoding, its traditional use with WFST decoders and lexicon constraints more closely aligns it with closed-vocabulary models (Prabhavalkar et al., 2023).

On the downside, these systems usually need more training data to achieve good accuracy. Also, their outputs can sometimes be less stable, especially in noisy environments or when the input is less structured. This is partly because they don't use decoding graphs, which normally help guide the output in more traditional systems. As a result, they may be more likely to make mistakes with words that sound the same or have tricky spellings.

Furthermore, without fixed lexicons or external language models, open-vocabulary systems can have trouble staying consistent with special terms in certain fields. This can be problematic in settings like healthcare, where precise word choice is critical. These systems may also create words that sound correct but are actually wrong, especially when the speaker has a strong accent, uses rare names, or speaks technical language. To improve robustness, some approaches combine open-vocabulary decoding with domain-adapted language models or post-processing filters.

2.2.3 Closed-vocabulary and open-vocabulary ASR for Healthcare Speech

In the context of healthcare speech recognition, closed-vocabulary and open-vocabulary ASR systems represent two contrasting approaches with complementary strengths. Closed-vocabulary systems, typically based on TDNNs and implemented via WFST decoders in frameworks like Kaldi,

[3](Li, 2022)

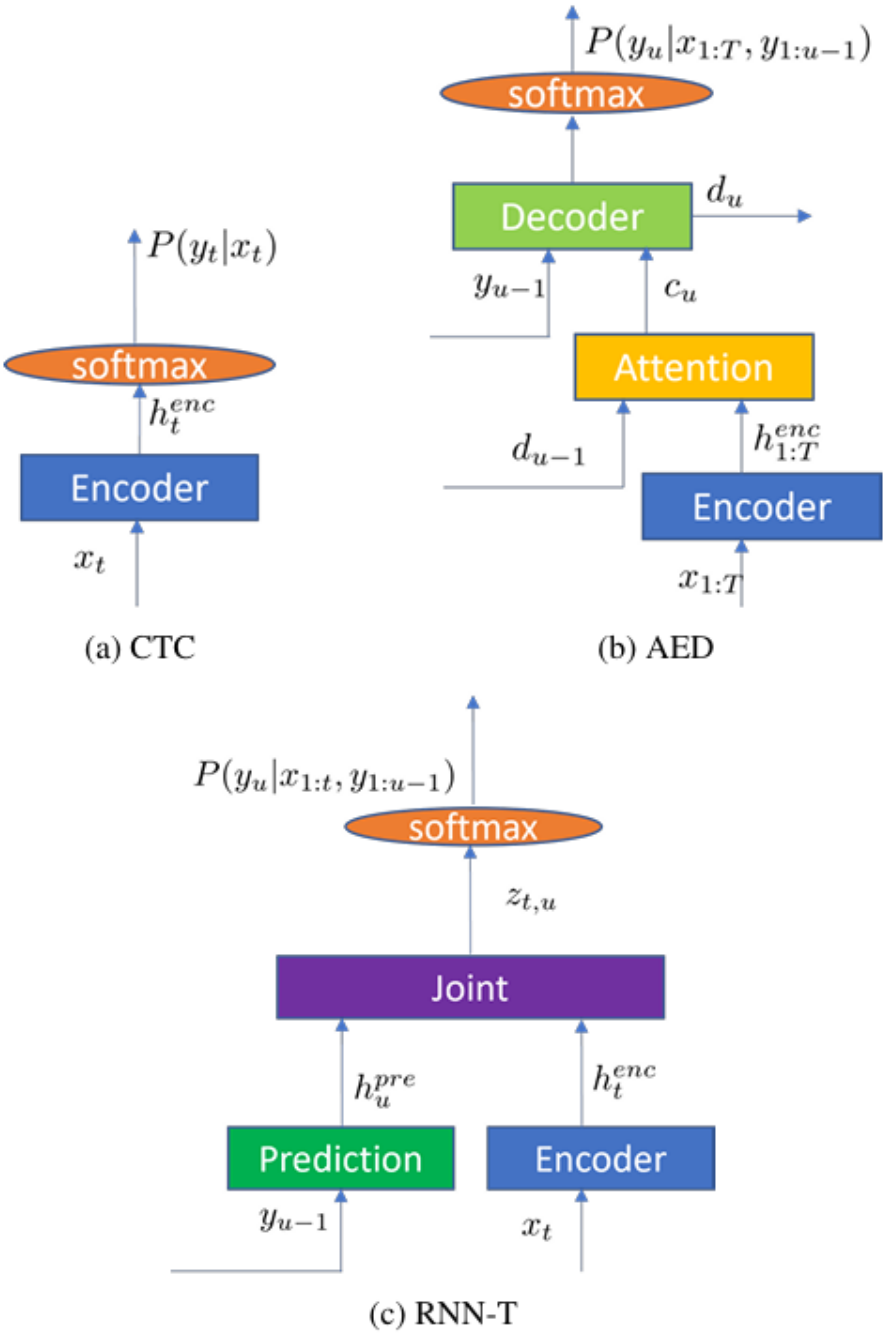

Figure 3: Architectures of CTC, AED, and RNN-T[3]

have proven effective in structured medical tasks (Peddinti et al., 2015). These models rely on predefined phoneme-based lexicons, which ensure consistent pronunciation modeling and help reduce substitution errors for known medical terms (Chu, Chang, & Xiao, 2021). For example, (Popović, Pakoci, & Pekar, 2020) reported that a TDNN-RNN-LSTM system achieved 97% accuracy on structured Serbian medical reports, illustrating the precision and reliability of lexicon-based ASR in predictable clinical documentation scenarios.

However, this controlled precision comes at the cost of adaptability. Closed-vocabulary systems often struggle with novel or evolving terminology and proper names, such as new pharmaceutical brands or emerging disease classifications, due to their reliance on manually curated lexicons. This limits their scalability across different institutions, dialects, or rapidly changing linguistic contexts (Scharenborg et al., 2020).

By contrast, open-vocabulary systems, such as the K2 Pruned RNN-Transducer (K2-RNN-T) (Kuang et al., 2022), employ subword tokenization techniques (e.g., BPE) that allow dynamic word construction and improved handling of OOV terms. This design makes them especially well-suited for spontaneous clinical speech, where domain-specific named entities and foreign terms frequently occur (*Juan*, 2024). Open-vocabulary architectures like RNN-T and Wav2Vec2 have shown strong performance in general-domain ASR, and recent studies have begun to explore their applicability to clinical transcription (Zuluaga-Gomez, 2024). Meanwhile, Open-vocabulary systems have benefitted from self-supervised pretraining and medical-specific fine-tuning strategies. Models like Whisper and Wav2Vec2, when adapted with domain-aware language models, have shown promise in zero-shot transcription for unseen medical contexts (Afonja et al., 2024). These models demonstrate potential for improving transcription quality in noisy, conversational, or low-resource medical speech.

Even with recent progress, there are still not many studies that clearly compare these ASR systems in less-studied languages like Dutch. Dutch brings special challenges for ASR because of its compound morphology, common code-switching with English, and the lack of annotated clinical corpora. These issues make Dutch a good language for testing how closed- and open-vocabulary systems work under both language and domain-specific limits. This study therefore aims to assess the trade-off between adaptability and precision in Dutch medical ASR, with a particular focus on clinical domain related words recognition. The following section explores these Dutch-specific challenges in greater detail and discusses their implications for clinical ASR system design.

2.3 Dutch Language-Specific Challenges in ASR

2.3.1 Dutch Morphological Complexity and Compounding in ASR

One of the key challenges in Dutch ASR lies in its rich morphological structure and extensive use of compounding. Dutch is considered a phrasal compounding language, where compounds are often formed by combining full lexical items into a single word unit (Alexiadou, 2020). Dutch compounds like boeken-kast (“bookcase”) exhibit syntactic structure and semantic transparency, making bound-

ary detection and tokenization more difficult for ASR systems. This complexity is further amplified by the presence of linking elements (e.g., -en-), as shown by (Banga, Hanseen, Neijit, & Schreuder, 2013), which are not merely phonological connectors but carry semantic weight, often interpreted as plural markers by native speakers. In spontaneous speech, these elements may be reduced or omitted, increasing the likelihood of recognition errors.

In language-mixing contexts, Dutch compounds often retain their head-final structure even when combined with foreign elements, such as in vélo-winkel ("bicycle shop", with a French stem and Dutch compound structure). While open-vocabulary systems offer flexibility, their reliance on subword units can make it difficult to accurately segment and interpret complex Dutch compounds. These fixed structures, while systematic, often produce hybrid forms that challenge the model's ability to maintain semantic integrity. On the other hand, these compounds are frequently out-of-vocabulary or non-standard, especially in medical or casual speech, so traditional lexicon-based decoding approaches may also fail to recognize them correctly. For this reason, compounding and morphological ambiguity in Dutch add significantly to the lexical and acoustic variability that ASR models must address, making robust and adaptive modeling strategies particularly important in Dutch-language medical ASR.

2.3.2 Code-Switching in Dutch-English Medical Speech Recognition

An additional complication in Dutch ASR, particularly within healthcare domains, is the frequent occurrence of code-switching between Dutch and English. Due to the international nature of medical terminology, speakers often alternate seamlessly between the two languages, embedding English terms such as MRI scan, chronic fatigue, or CT angiography within otherwise Dutch utterances. These code-switching instances are not only common in informal clinical dialogue but are also present in formal settings like dictations and reports, creating substantial challenges for monolingual ASR systems trained exclusively on Dutch corpora. This is reflected in our clinical dataset, where English phrases such as "leg press" and "five-minute walk test" frequently appear in otherwise Dutch utterances.

Research on multilingual ASR has shown that code-switching significantly increases word error rates, especially when the acoustic or language model lacks robust language identification capabilities (Sitaram, Chandu, Rallabandi, & Black, 2019). For example, in the FAME! speech corpus of Frisian-Dutch code-switching, recognition errors typically cluster around language-switch boundaries and embedded foreign terms (Yılmaz, 2018). While Frisian and Dutch are typologically closer than Dutch and English, similar disruptions have been observed in Dutch-English contexts, especially for specialized terminology. These challenges are further amplified in healthcare, where English terms often serve as domain-specific named entities critical for downstream tasks like entity recognition and concept mapping.

Despite extensive sociolinguistic research on Dutch code-switching (Boumans, 1998), there remains a significant gap in speech recognition research addressing this phenomenon in practical ASR systems. Most Dutch ASR models rely on single-language training pipelines and struggle with code-switched inputs, often misrecognizing or omitting critical medical terms. As such, accounting for

intra-utterance code-switching, particularly at the subword or byte-pair encoding level—may prove essential for improving model robustness in real-world Dutch medical ASR applications.

2.3.3 Resource Scarcity in Dutch Medical ASR

One of the major obstacles to building robust ASR systems for Dutch in the medical domain is the limited availability of high-quality, domain-specific speech corpora. Unlike English, which benefits from large-scale annotated datasets such as MIMIC-CXR or LibriSpeech-med, Dutch remains a low-resource language in the context of both general and clinical ASR (Mustafa et al., 2022). Publicly accessible Dutch medical speech datasets are virtually nonexistent, forcing researchers to rely on general-domain corpora such as CGN (Corpus Gesproken Nederlands), which do not adequately capture the specialized vocabulary, acoustic variability, or interactional structures characteristic of healthcare speech.

This scarcity of in-domain data poses multiple challenges. First, acoustic models trained on non-medical Dutch data often perform poorly when deployed in hospital environments, where speech is affected by background noise, reverberation, and speaker variability, such as doctors using rapid or formal speech. Second, the lack of specialized language modeling resources, such as clinical word embeddings or medical language models in Dutch, limits the ability of open-vocabulary ASR systems to generate accurate transcriptions of rare or technical terms (Wang, 2023). Moreover, since Dutch is often spoken alongside English in medical contexts, datasets that capture spontaneous code-switching phenomena are especially rare, impeding research into multilingual or mixed-lingual recognition systems (Shen, 2022).

As a result, most Dutch ASR systems struggle to generalize beyond structured, scripted speech scenarios. This limitation is particularly concerning for applications such as speech-driven medical documentation, where recognition accuracy of domain-specific terms is critical.

2.4 Clinical Terminology and Named Entity Recognition (NER)

2.4.1 Integration of Clinical Terminology Systems in ASR Research

Among existing standardized medical terminology systems, SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) stands out for its broad conceptual coverage and rich semantic structure. It supports compositional reasoning, post-coordination, and synonym resolution, these features are particularly useful for structuring clinical concepts in ASR and natural language processing (NLP). Rather than manually annotating clinical transcripts, researchers increasingly leverage SNOMED CT in conjunction with NLP tools to automate terminology extraction and normalization.

SNOMED CT's structure includes three key components: concepts, descriptions, and relationships. This makes it both a rich ontology and a semantic graph, supporting advanced capabilities such as postcoordination and compositional grammar. These features enable the construction of new, com-

pound concepts from existing terms, allowing more precise and expressive clinical representations. However, manually encoding free-text clinical notes into SNOMED CT is costly and labor-intensive. Consequently, NLP approaches are frequently employed to automate the mapping process, particularly in large-scale datasets.

In ASR and NLP contexts, SNOMED CT provides the necessary semantic grounding for medical entity normalization, synonym disambiguation, and concept-level inference. Prior studies have demonstrated its utility in evaluating ASR outputs and enabling structured clinical documentation (Melton et al., 2006), for example, used SNOMED's internal relationships to compute semantic distances between ASR hypotheses and ground truth. More recently, (Chang & Sung, 2024) integrated SNOMED CT knowledge into large language models to improve terminology recognition in conversational speech. In this study, we leverage SNOMED CT as the reference terminology system for extracting and evaluating clinical concepts from manually transcribed (gold) ASR transcripts. By aligning ASR hypotheses with SNOMED-defined concepts, we aim to facilitate robust and semantically informed terminology evaluation in Dutch medical speech contexts.

2.4.2 Named Entity Recognition in Clinical ASR Output

Named Entity Recognition (NER) from ASR-generated clinical transcripts presents unique challenges due to the compounded effects of transcription errors and domain-specific linguistic variability. Medical named entities (MNEs), such as drug names, conditions, procedures, and anatomical terms, are especially prone to misrecognition in ASR outputs. (Afonja et al., 2024) evaluated a range of ASR models, including Whisper-large, and found that despite strong overall WER performance, MNE recall remained low: only 42% for medications (MED), 33% for protected health information (PHI), 59% for test and treatment procedures (TTP), and 67% for medical conditions (COND). Even the best-performing category, anatomy (ANA), achieved just 71% recall, likely due to the high frequency and lexical familiarity of terms such as "heart," "brain," and "liver" in web-scale corpora. These findings highlight the insufficiency of surface-level transcription accuracy when it comes to faithfully capturing critical medical terminology in ASR–NER pipelines.

To better quantify semantic recognition performance, Afonja et al. proposed MedTextAlign, a fuzzy alignment tool that accommodates spelling variation and minor phonetic deviations in ASR output; inspired by this, we adopt a similar fuzzy matching approach using the faster and more flexible "RapidFuzz" library in Python to align gold and hypothesis transcripts. Complementing this, (Meripo & Konam, 2022) proposed an entailment-based approach to detecting ASR errors in medical dialogues. By training a multimodal model that jointly encodes audio and text, their system was able to detect inconsistencies between spoken and transcribed clinical content, especially for medical terms that were omitted, substituted, or altered. Importantly, they showed that many transcription errors affecting MNEs went undetected by traditional confidence scoring mechanisms, reinforcing the value of semantically-informed evaluation strategies. Besides, they also evaluated ASR error detection across five semantic entity groups, demonstrating that category-specific errors often escape traditional surface-level scoring. Their use of a semantically-informed entailment model and manual concept alignment with UMLS terminologies showed that both evaluation metrics and align-

ment strategies must be entity-aware to reflect clinical relevance. Their study highlights the value of entity-aware evaluation and preprocessing; while we do not adopt their multimodal entailment model or UMLS-based alignment, we apply similar text cleaning steps and emphasize semantic-level evaluation by aligning ASR outputs with SNOMED-defined entity categories using global alignment.

In conclusion, the present study adopts fuzzy matching techniques to do the alignment work, grounded in SNOMED CT, and applies term-group-based recall reporting to better capture recognition fidelity across heterogeneous medical entity types. In doing so, it addresses the critical gap between surface-level transcription accuracy and domain-specific term fidelity—particularly in low-resource language contexts where pronunciation variation and term novelty further complicate entity recognition.

2.5 Evaluation Metrics and Error Analysis for ASR

Evaluating the performance of ASR systems, especially in clinical contexts, requires a combination of general and domain-specific metrics. Traditionally, Word Error Rate (WER) and Character Error Rate (CER) are the most widely used indicators, capturing substitution, deletion, and insertion errors at the word or character level. While WER provides a high-level approximation of transcription accuracy, it fails to account for the clinical importance of specific terms, which can be an issue particularly critical in healthcare ASR systems.

To address the limitations of general-purpose transcription metrics like WER in clinical contexts, M-WER and M-CER have emerged as task-specific alternatives that evaluate ASR accuracy based on medically salient terms. Early motivations for these metrics were discussed by (Liu, Tur, Hakkani-Tur, & Yu, 2011), who emphasized that clinical misrecognitions often have greater implications than general transcription errors. More recently, (Afonja et al., 2024) formalized M-WER and M-CER to assess model performance on specific medical entities (e.g., medications, conditions), revealing that entity-level recall remained as low as 30–50% even when overall WER was below 10%.

Beyond surface-level transcription metrics, precision, recall, and F1-score are widely adopted in evaluating downstream NLP tasks, especially NER from ASR transcripts. These metrics assess the system's ability to correctly identify and extract clinically relevant entities: precision quantifies the correctness of predictions, recall measures coverage, and F1-score balances the two. Their relevance has grown as clinical NLP systems increasingly rely on ASR outputs, where identifying the correct entity is often more critical than exact word reproduction. Their significance was demonstrated early on by (Wu, Jiang, Xu, Zhi, & Xu, 2018), who applied them to evaluate deep learning–based clinical NER models and highlighted that entity type (e.g., diseases vs. tests) affects model performance asymmetrically, with F1-scores varying notably by category.(Meripo & Konam, 2022) emphasized the importance of entity-level evaluation in ASR by reporting precision, recall, and F1 across clinically relevant term categories, highlighting that critical misrecognitions may persist even with low overall WER. A more comprehensive view is offered by (Navarro et al., 2023), who reviewed over 80 studies and confirmed that more than 90% used these metrics as primary evaluation tools. They further recommended reporting both micro- and macro-averaged F1-scores to better capture per-

formance on low-frequency entities. These findings reaffirm that precision, recall, and F1 remain essential and reliable metrics for quantifying model effectiveness in clinical NLP applications, particularly where accurate and complete identification of medical entities is critical.

In addition, novel task-level metrics such as RadGraph F1 (Huh, Park, Lee, & Ye, 2023) and severity-weighted accuracy scores (Whetten & Kennington, 2023) have emerged, aiming to quantify the downstream clinical utility or risk impact of ASR mistakes.

Building on these insights, our study adopts both traditional and domain-specific evaluation strategies to better capture the clinical relevance of ASR outputs. In particular, we focus on term-level evaluation by matching extracted medical terms in the ASR output against a reference terminology set, allowing more precise measurement of recognition accuracy. This study employs WER, CER, Medical WER, Medical CER, precision, recall, and F1 score as its primary evaluation metrics.

3 Methodology

3.1 Dataset Description

The dataset used in this study was provided by Gerimedica, a Dutch healthtech company that develops electronic health record (EHR) systems and clinical decision support tools specifically for long-term care settings.

The dataset consists of approximately 42.5 hours of Dutch-language healthcare reports transcripts. The data were collected from eight healthcare organizations across multiple Dutch provinces (Noord-Holland, Utrecht, Groningen, Noord-Brabant, Overijssel and Gelderland), which means the transcripts may exhibit regional variants of Dutch. It contains 8,595 transcript segments contributed by a diverse group of healthcare workers, including physicians, psychologists, dietitians, physical therapists, nurses, and speech therapists. The gold transcripts are formatted as structured text files, with each line consisting of a unique utterance ID and the corresponding transcribed. These transcripts represent real-world clinical interactions and are transcribed manually to serve as high-quality references.

3.2 Core Methods and Models

3.2.1 ASR Models

1. **Closed-vocabulary model: Kaldi TDNN**
   We implement a Time Delay Neural Network (TDNN) (Peddinti et al., 2015) based ASR system using the Kaldi toolkit, following the LibriSpeech[4] tdnn_1d chain recipe. The acoustic model consists of 16 TDNN-F layers with bottleneck projections for efficient long-context modeling, preceded by an LDA dimensionality reduction layer and ReLU-activated batch normalization with scheduled dropout (progressing from 0 to 0.5 during training). The system utilizes lattice-free Maximum Mutual Information (LF-MMI) optimization with cross-entropy regularization, trained with a decaying learning rate from 0.00015 to 0.000015 across four epochs. Input features comprise 40-dimensional MFCCs augmented with 100-dimensional iVectors for robust speaker adaptation. It is worth mentioning that in this study the Kaldi TDNN model also uses an external N-gram language model trained on a roughly 6 GB text corpus (78.5 million lines) from reports of 8 Dutch healthcare organizations which were mentioned in the previous section.

2. **Open-vocabulary model: Pruned RNN-Transducer (K2-RNN-T)**
   In this study, the Pruned RNN-Transducer (RNN-T) was adopted (Kuang et al., 2022), specifically the K2[5] Stateless variant, as the representative open-vocabulary ASR model. This model

[4] https://github.com/kaldi-asr/kaldi/tree/master/egs/librispeech/s5

[5] https://github.com/k2-fsa/icefall/tree/master/egs/librispeech/ASR/pruned_transducer_stateless7

is implemented with the Zipformer encoder, which is a hierarchical and efficient variant of the Conformer model, designed to model long-range acoustic dependencies. The encoder consists of 12 stacked ZipformerEncoder layers with 80-dimensional fbank input features, attention dimension of 256, and feedforward dimensions up to 2048. The model performs 2D subsampling for frame rate reduction and applies dropout, layerdrop, and activation balancing to enhance training stability and robustness.

Same with the Kaldi TDNN model, this model was trained on 1046 hours of Dutch transcribed speech supplemented with 12 hours of domain-specific healthcare audio. Training used the Adam optimizer with learning rate scheduling and mixed-precision acceleration. No external language model was used. BPE was applied to generate subword vocabulary units. This method enables the model to form flexible combinations of characters and morphemes, making it capable of recognizing previously unseen or rare words. enabling open-vocabulary recognition and better generalization to rare or unseen terms.

3.2.2 Terminology and Named Entity Processing Methods

To evaluate the performance of ASR systems in recognizing domain-specific language, this study focuses on two major types of lexical content: clinical terminology and named entities with healthcare relevance. The goal is to identify key terms and entities in the gold standard transcripts that can serve as targets for evaluating ASR accuracy in healthcare contexts.

1. **Clinical Terminology Extraction from SNOMED CT** Medical terms, specifically those related to drugs and clinical findings,,were identified using the SNOMED CT terminology system. In this study, we used the March 2025 Netherlands Edition (v1.0) of SNOMED CT[6], published on March 31, 2025. Clinical terms were extracted from the Dutch-language concept and description files located in the "Snapshot/Terminology/" directory. It contains 535,412 distinct medical concepts, each associated with one or more descriptions in Dutch and/or English. These concepts served as the basis for identifying relevant clinical terms in the transcripts.

2. **Named Entity Recognition (NER)** NER was employed in this study to identify contextually important entities that fall outside standard clinical terminologies. These entities, including person names, location names, and product brands, are not strictly medical terms but frequently appear in spoken medical consultations. They provide critical contextual information, such as individuals involved in care, healthcare facilities, or specific product usage. However, their inherent variability, local specificity, and absence from predefined vocabularies or training corpora make them particularly challenging for ASR systems to recognize accurately. This study aims to improve the capture of such information, which is highly susceptible to ASR misrecognition. For this purpose, we employed the spaCy "nl_core_news_lg" model, a pre-trained Dutch-language pipeline that includes part-of-speech tagging, syntactic parsing, and NER. The model is trained on the "Lassy Large Corpus", a richly annotated Dutch treebank, and supports recognition of a wide range of general-domain named entities in Dutch.

[6] https://mlds.ihtsdotools.org/#/viewReleases/viewRelease/128785

3.3 Technical Framework

This section describes the technical framework developed to support the analysis and evaluation of ASR outputs within a Dutch healthcare context. Although the gold transcripts and the ASR outputs were pre-generated and provided by Gerimedica, a structured pipeline was designed to extract clinically relevant terms, match them across the transcripts, and prepare the data for evaluation. The framework (Figure 4) consists of five major components: input processing, terminology extraction, matching strategy, structured output generation, and evaluation preparation.

3.3.1 Input Data Preprocessing

The input data comprised these following files: Gold transcript: the reference transcription created by human annotators. Hypothesis transcripts: one set of transcriptions generated by a Kaldi TDNN ASR model and one by Pruned RNN-Transducer ASR model. All transcripts were structured line-by-line, with each line containing a UUID and the corresponding utterance. As preprocessing, text normalization was applied to remove punctuation, unify lowercase formatting, and tokenize utterances consistently using the "nl_core_news_lg" model from spaCy.

3.3.2 Extraction of Clinical Terms and Named Entities

1. **Method Overview**
   Following the methods outlined in Section 3.2.2, two sources of terms were considered in this study:

   Clinical Terms: Extracted from a subset of the Dutch SNOMED CT terminology, focusing on clinically relevant concepts.

   Named Entity: Identified using a Dutch-language NER model (spaCy: nl_core_news_lg), used to detect contextually important terms outside formal medical vocabularies.

   Term extraction was performed on the gold transcripts using rule-based concept matching and automatic entity recognition. Each detected term was annotated with its source (SNOMED or NER) for subsequent comparison with ASR outputs.

2. **Clinical Terms Selection and Extraction**
   To analyze domain-specific terminology in the gold transcripts, a structured extraction process was applied using the SNOMED CT Netherlands Edition. Three major categories of clinical terms were selected: diseases, drugs, and clinical findings. These categories were chosen based on both their empirical frequency in the transcripts and their practical relevance in clinical and healthcare context. Additionally, they represent high-level semantic groups within the SNOMED ontology and account for a substantial proportion of concepts in the overall vocabulary.

   For each category, high-level SNOMED parent classes or subtypes were used to define the extraction scope. The drug category was derived from two main parent classes: Pharmaceutical

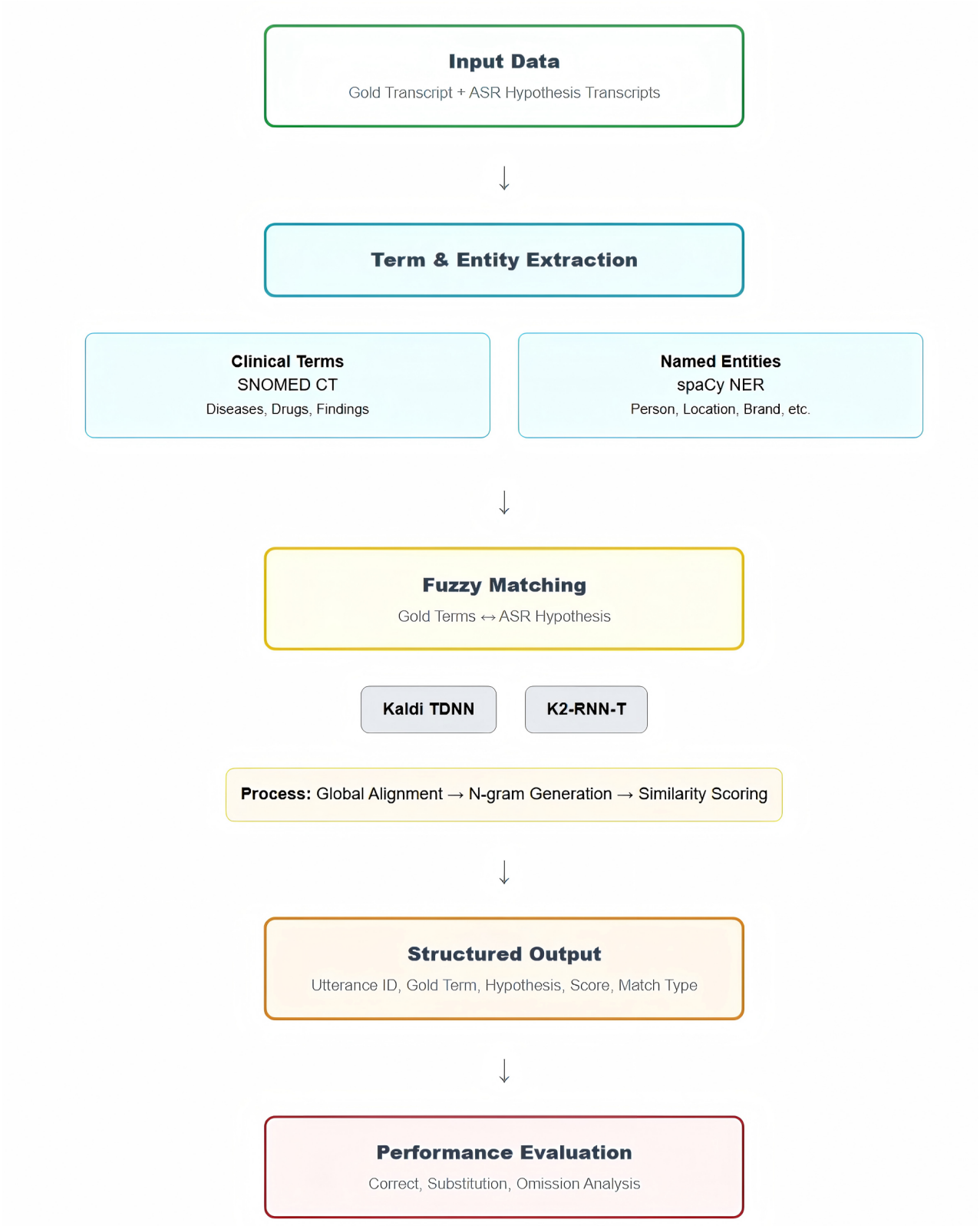

Figure 4: Diagram of the Technical Framework

/ Biologic Product (373873005) and Substance (105590001), This grouping reflects common practices in Dutch clinical language, where chemical substances are often used as proxies for drug names (e.g., "paracetamol" instead of a brand-specific formulation). The disease category included selected subtypes such as Infectious Diseases (40733004), Cardiovascular Disorders (49601007), and Mental Disorders (74732009). These classes represent common and clinically significant diagnostic themes relevant to long-term care. For clinical findings, frequently observed but non-diagnostic expressions were included using classes like Functional Finding (118228005) and Mental State or Behavioral Finding (384821006).

To extract terms from each SNOMED category, we used the official RF2 release files. Descendant concept IDs were retrieved from the relationship file (typeId = 116680003) using a recursive "is-a" traversal. These concept IDs were then matched to active Dutch-language terms in the description file. The resulting terms were labeled by category (e.g., drug, disease), filtered for duplicates, and saved as structured lexicons for downstream transcript matching. Additionally, to avoid over-representing overly generic expressions, we removed common high-frequency terms (e.g., sleep, meat) that may appear in daily conversations but are not clinically informative. This process resulted in 141,100 disease terms, 20,703 drug terms, and 44,203 clinical finding terms. These lexicons were used for downstream transcript matching.

Although the term categories were initially chosen based on domain knowledge, coverage analysis confirmed their relevance. Drug and clinical finding terms revealed high lexical variability and frequency, while disease terms provided strong semantic cues for evaluating diagnostic accuracy in ASR systems.

These findings support the selection of all three term types for ASR evaluation, reflecting a balance between clinical relevance, lexical diversity, and observable frequency in spontaneous speech. Detailed SNOMED category mappings and full coverage statistics are provided in Appendix A and B.

3. **Named Entities Selection and Extraction**
To complement clinical terminology, we extracted additional named entities from the gold transcripts, including both standard entity types and other medically relevant expressions such as abbreviations. The initial recognition was performed using the Dutch "nl_core_news_lg" model in spaCy. To enhance domain alignment, we applied several post-processing steps, including pattern-based relabeling, rule-based extraction (e.g., for abbreviations, units, brand names), and dictionary-based lookups for known domain-specific terms.

From the resulting annotations, six categories were defined: PERSON (person name), LOC (short for location), BRAND (brand name), ABBR (short for abbreviation), CARDINAL_UNIT, and OTHER. A coverage review indicated that PERSON and ABBR terms were not only frequent in transcripts but also carried significant semantic importance in clinical communication. LOC and BRAND categories, while slightly more context-dependent, were also retained for their relevance to healthcare documentation and product mentions. In contrast, CARDINAL_UNIT terms (e.g., "2 ml", "drie keer") mostly captured dosage and time expressions and lacked semantic depth, while OTHER included miscellaneous or ambiguous content (e.g., languages, dates) with limited clinical interpretability. As such, CARDINAL_UNIT and OTHER were excluded from subsequent ASR evaluation to ensure focus on meaningful named-entity content. See Appendix C and D for entity definitions and coverage statistics.

These refinements enabled a more targeted and domain-informed selection of entities for evaluating ASR performance in real-world Dutch healthcare conversations.

3.3.3 Transcript Matching Strategy

1. **Method Overview**
   To evaluate ASR performance at the term level, we developed a two-step matching strategy. First, each utterance in the gold transcript was aligned with its corresponding ASR output using unique identifiers to ensure a one-to-one mapping. Then, for each aligned utterance pair, predicted terms were extracted from the ASR hypothesis using a fuzzy matching approach, allowing for approximate matches with gold-standard terms. This strategy was applied consistently across all term categories, including clinical concepts and named entities, and provided structured outputs for downstream evaluation.

2. **Global Alignment**
   To establish a consistent comparison basis, we aligned the gold transcript and ASR outputs (from both Kaldi TDNN and K2-RNN-T models) using unique utterance IDs. Each transcript was formatted line-by-line, where each line contained a UUID and its corresponding sentence. This ensured a one-to-one correspondence between gold and hypothesis utterances and avoided issues caused by duplicates or missing entries. The aligned utterance pairs were merged into a unified dataset, forming the foundation for subsequent term-level comparison.

3. **Fuzzy Matching and N-gram Extraction**
   For each aligned hypothesis utterance, we tokenized the text and generated all possible word n-grams (up to length 3) to account for fragmented or rephrased expressions commonly found in ASR outputs. Each gold term was matched against candidate n-grams using the RapidFuzz (Seiffert, 2021) library, which computes Levenshtein-based similarity scores. These scores reflect how closely each ASR output segment resembles the reference term, accounting for surface variations. The resulting gold-hypothesis pairs and similarity scores formed the basis for subsequent evaluation. Details on thresholding and classification are described in Section 3.4.1.

   This process was applied to all target term types (e.g., disease, drug, clinical finding, named entity). The resulting outputs, each consisting of a gold term, its best-matching phrase, and similarity score, served as the input for the evaluation described in the next chapter.

3.3.4 Output for Evaluation

Following the fuzzy matching process, each ASR model's hypothesis transcript was compared against the gold transcript. For each model, a structured output file was generated for each term category. Each file contained instance-level records including the utterance ID, gold term, matched hypothesis phrase, match score, and match type.

This structured output enables a fine-grained evaluation of ASR performance in recognizing clinically relevant terms, supporting the identification of error types such as substitutions, omissions, etc. The structured output generated through this framework forms the basis for both quantitative evaluation and qualitative analysis, as described in the following chapter.

3.4 Evaluation Methodology

To assess ASR model performance in the context of clinical speech transcription, this study employed a combination of quantitative metrics and qualitative error analysis, with particular emphasis on the accurate recognition of medically and contextually relevant terms.

3.4.1 Quantitative Metrics

Several metrics were used to evaluate transcription accuracy at both the overall and terminology-specific levels. These were computed separately for different term types, providing fine-grained insight into model strengths and weaknesses across terminology categories.

1. **Word Error Rate (WER)**: The standard metric for ASR evaluation is the Word Error Rate (WER), which is calculated as:

$$\text{WER} = \frac{S + D + I}{N}$$

where S = substitutions, D = deletions, I = insertions, and N = total number of words in the reference transcript. WER was computed over the entire transcript to assess general model performance. A lower WER indicates better performance.

2. **Medical WER(M-WER)**: M-WER measures the word-level error rate for medically relevant terms, particularly the clinical terms and named entities mentioned in previous sections. This provides a more focused assessment of ASR performance on medically significant content. A lower M-WER indicates better performance.

3. **Character Error Rate (CER)**: A fine-grained metric for ASR evaluation, calculated at the character level as:

$$\text{CER} = \frac{S + D + I}{N}$$

where S = substitutions, D = deletions, I = insertions, and N = total number of characters in the reference transcript. It complements WER by providing more fine-grained insight, especially useful for short or morphologically complex words. A lower CER indicates better performance.

4. **Medical CER (M-CER)**: A domain-specific variant of CER, focusing on character-level accuracy within medically relevant terms, particularly the clinical terms and named entities mentioned in previous sections. It offers a more targeted evaluation of ASR performance in clinical contexts. A lower M-CER indicates better performance.

5. **Precision**: Precision is defined as the proportion of correctly predicted positive instances among all instances predicted as positive. In the context of terminology recognition, it reflects the accuracy of the model's predictions, i.e., the likelihood that a predicted term is indeed correct. It is formally defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

where TP denotes true positives and FP denotes false positives. A higher precision indicates better performance.

6. **Recall**: Recall measures the proportion of correctly predicted positive instances among all actual positive instances. It reflects the model's ability to capture relevant terms from the reference. It is defined as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

where FN denotes false negatives. A higher recall indicates better performance.

7. **F1 Score**: F1 Score is the harmonic mean of precision and recall, providing a single metric that balances both. It is particularly useful in evaluating systems where both false positives and false negatives carry significant consequences, as is often the case in clinical information extraction. The formula is:

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

A higher F1 score indicates better performance.

In this study, to compute term-level evaluation metrics such as Precision, Recall, and F1-score, we first categorized each fuzzy match based on its similarity score:

- **Correct:** similarity = 100 (exact match)
- **Substitution:** $75 \leq$ similarity < 100 (near match)
- **Missing:** similarity < 75 (no sufficient match)

These match types were then mapped to standard evaluation categories as follows:

- **True Positive (TP)** = correct
- **False Positive (FP)** = substitution
- **False Negative (FN)** = missing

3.4.2 Qualitative Error Analysis

While quantitative metrics such as WER and F1 Score provide aggregate performance measures, they do not capture the nuanced ways in which recognition errors affect the semantic integrity of medical content. In this way, a qualitative error analysis was conducted to characterize common patterns of ASR misrecognition in clinically relevant terms.

Errors were manually reviewed and categorized into three types, based on surface similarity and semantic distortion:

1. **Substitution errors**: e.g., incorrect terms or semantically misleading replacements.
2. **Approximate similarity errors**: e.g., phonetically similar but incorrect words.
3. **Truncation or omission errors**: e.g., partially recognized or dropped terms.

Representative examples were selected from the structured matching output described in Section 3.3.4.

3.5 Ethics and Research Integrity

3.5.1 Data Ethics and Privacy

This study uses sensitive healthcare data provided by Gerimedica. To ensure ethical and legal compliance, the researcher has signed a confidentiality agreement and obtained a certificate of conduct, as required for accessing and processing sensitive patient-related information. The study will not involve any direct interaction with patients or healthcare workers, no personally identifiable information is stored or shared. In line with data protection regulations such as the GDPR, all data used in this study is processed in a secure, access-restricted environment. The research laptop is encrypted using BitLocker, and the data is strictly confined to local use.

3.5.2 FAIR Principles Implementation

In accordance with the FAIR principles (Findable, Accessible, Interoperable, Reusable), this study aims to promote transparency and reusability of research materials where possible. While the primary data used in this project, sensitive healthcare consultation transcripts provided by Gerimedica, and the SNOMED Dutch terminology database are subject to confidentiality and licensing restrictions and therefore cannot be shared publicly, all supplementary resources will be made openly available. Specifically, all custom scripts, preprocessing steps, and analysis code developed for this study will be published in a well-documented GitHub repository. The code will be organized in a

clear and structured format, with accompanying documentation and example configurations to facilitate reuse by other researchers. File naming conventions, folder hierarchies, and dependencies will follow standard best practices to ensure interoperability across systems and platforms. By doing so, the project supports the broader FAIR goals to the extent permitted by data privacy and licensing constraints.

3.5.3 Open Science Practices

To support open science, the analysis code and documentation developed for this project will be made publicly available via GitHub [7]. Although the original data cannot be shared due to confidentiality agreements, the repository will include a clear README file explaining the project structure and usage, allowing others to understand, reproduce, or adapt the methodology.

3.5.4 Bias and Fairness

This study acknowledges the potential biases inherent in healthcare language data, such as variations in terminology, spelling, and documentation styles across institutions or practitioners. Efforts will be made to minimize these biases by designing preprocessing steps carefully and by validating entity identification results. Any limitations related to data diversity and potential impacts on the findings will be explicitly discussed.

3.5.5 Environmental Impact

The environmental impact of this study is minimal, as it relies on lightweight computational methods and small-scale experiments. No large model training is involved; only existing language models and standard processing tools are used. All experiments are run on a local machine or limited cloud resources to reduce energy consumption.

3.5.6 Reproducibility and Replicability

To ensure reproducibility, all analysis code, preprocessing scripts, and configuration details will be made publicly available through a GitHub repository. Although the original healthcare data cannot be shared, the methodology will be fully documented, enabling others to replicate the workflow with similar datasets. Key steps and parameters will be clearly described to support consistent replication of the results.

[7] https://github.com/SaraSun01/thesis_closed_and_opened_ASR_comparison

4 Evaluation Results

This section presents a comprehensive evaluation of the two ASR models, focusing on both overall transcription performance and their ability to recognize domain-specific terms relevant to healthcare. The analysis is structured in two parts. The first part provides a quantitative comparison of the overall performance of the two models. The second part delves into a detailed assessment of how each model handles different types of healthcare-related terminology and named entities, as previously categorized. This evaluation aims not only to measure recognition accuracy, but also to reveal the models' respective strengths and weaknesses in handling specialized vocabulary. By combining quantitative metrics with qualitative error analysis, this section seeks to offer deeper insights into the challenges of ASR in the medical domain and inform potential strategies for model improvement.And more detailed data can be found in the Appendix E.

4.1 Overall ASR Performance Comparison

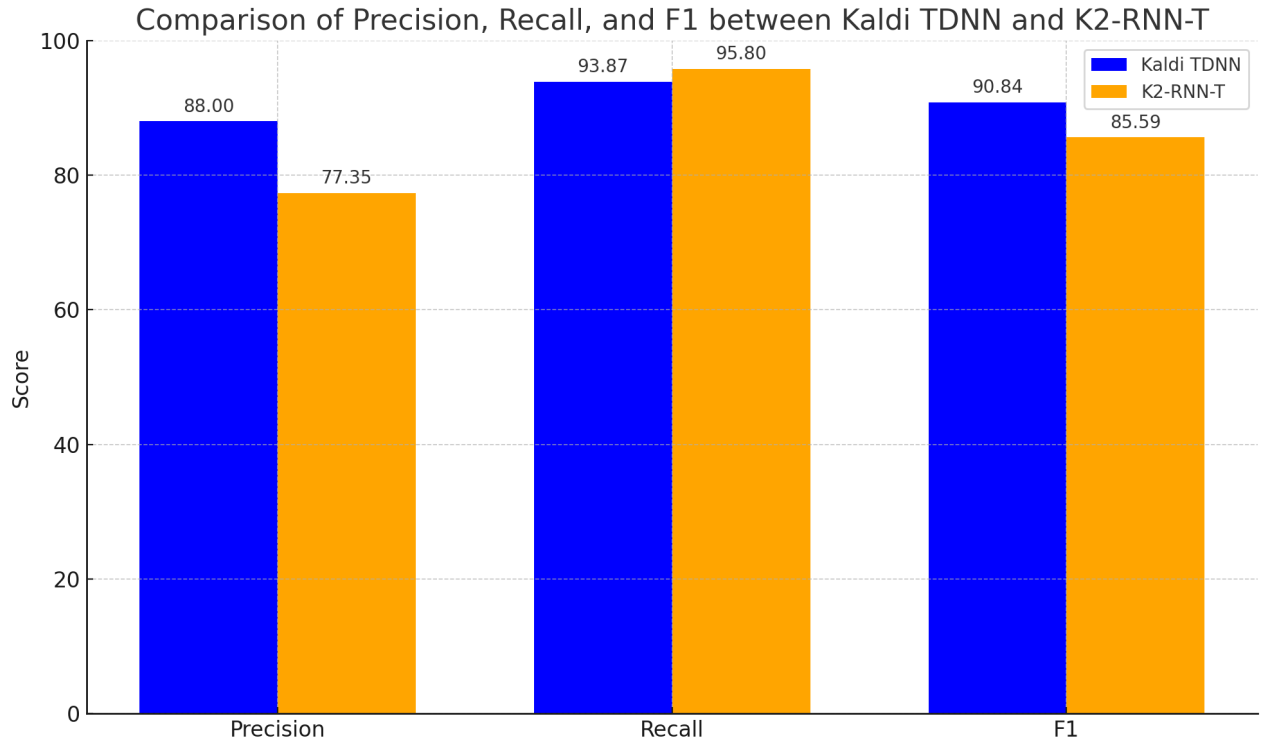

Figure 5

Figure 5 presents a macro comparison of the two models by gathering all recognized terms across categories and calculating overall precision, recall, and F1 score. These metrics reflect the models' general recognition tendencies beyond specific term types. Kaldi TDNN achieves higher precision (88.00%) and F1 score (90.84%) compared to K2-RNN-T (77.35% and 85.59%, respectively), sug-

gesting stronger reliability in producing correct predictions and maintaining a good balance between precision and recall. On the other hand, K2-RNN-T achieves a higher Recall (95.80%) compared to Kaldi (93.87%), indicating that it more frequently identifies relevant terms, even at the risk of including more incorrect ones. This behavior may reflect the model's subword-based decoding, which allows it to accommodate a wider range of lexical variations. Overall, the distribution of Precision, Recall, and F1 scores reflects a trade-off between conservativeness and inclusiveness in term recognition, with each model demonstrating a distinct balance in its classification behavior.

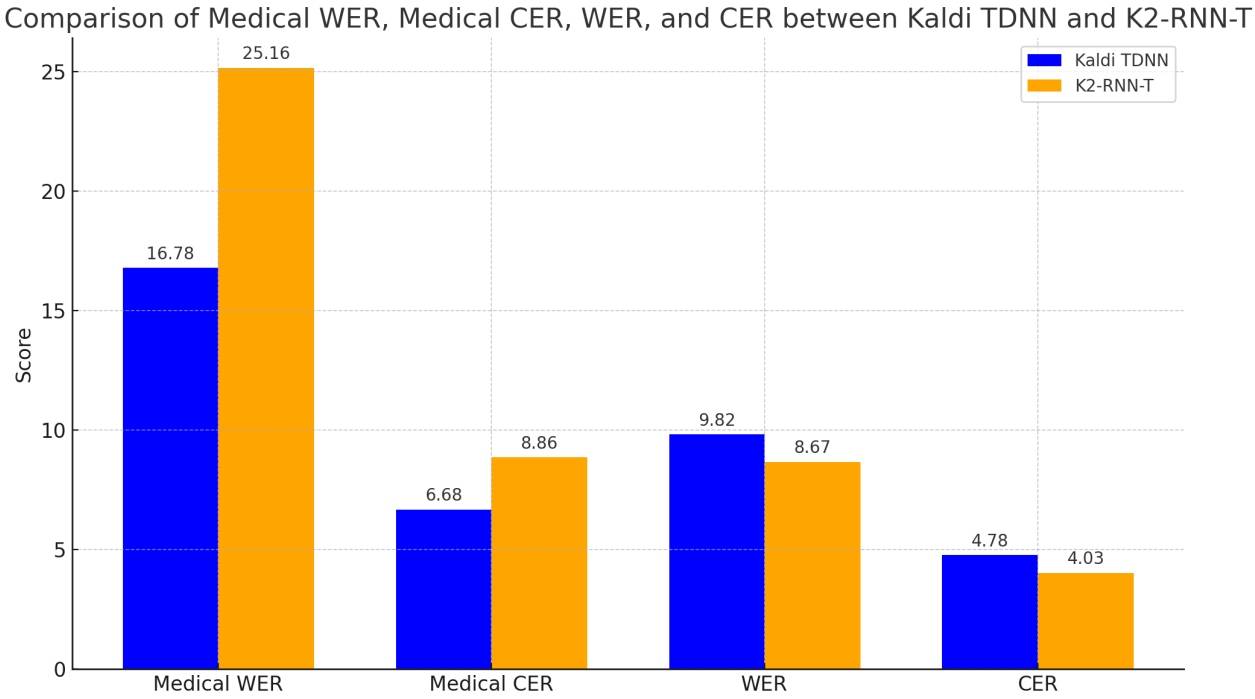

Figure 6

In addition to the precision, recall, and F1 score comparison, Figure 6 presents a comparison of word- and character-level error rates, including both overall and medical-specific metrics. Kaldi TDNN demonstrates significantly lower M-WER (16.78% vs.25.16%) and M-CER (6.68% vs. 8.86%), indicating more consistent and accurate recognition of clinical terms at both the word and character levels. These results align with its earlier advantage in precision and F1 score, particularly in structured medical terminology. For general transcription, however, the performance gap narrows. K2-RNN-T outperforms Kaldi in overall WER (8.67% vs. 9.82%) and CER (4.03% vs. 4.78%), suggesting greater flexibility in transcribing diverse or less standardized input. This reflects the open-vocabulary model's ability to generalize across broader linguistic variation, especially at the subword level.

These observations reflect differences in vocabulary coverage, error tolerance, and decoding strategies between the models, which manifest in their respective error rate profiles across domain-specific and general content.

4.2 Analysis of Clinical Terminology and Entity Recognition

4.2.1 Metric-wise Comparative Analysis

This section analyzes model performance on different categories of clinical terminology and named entities, organized by evaluation metric. By examining precision, recall, F1 score, M-WER, and M-CER separately, we aim to uncover how each model performs across various term types and to highlight distinct patterns or differences in their recognition behavior. The comparison is based on the category-level results presented in the follwing evaluation figures.

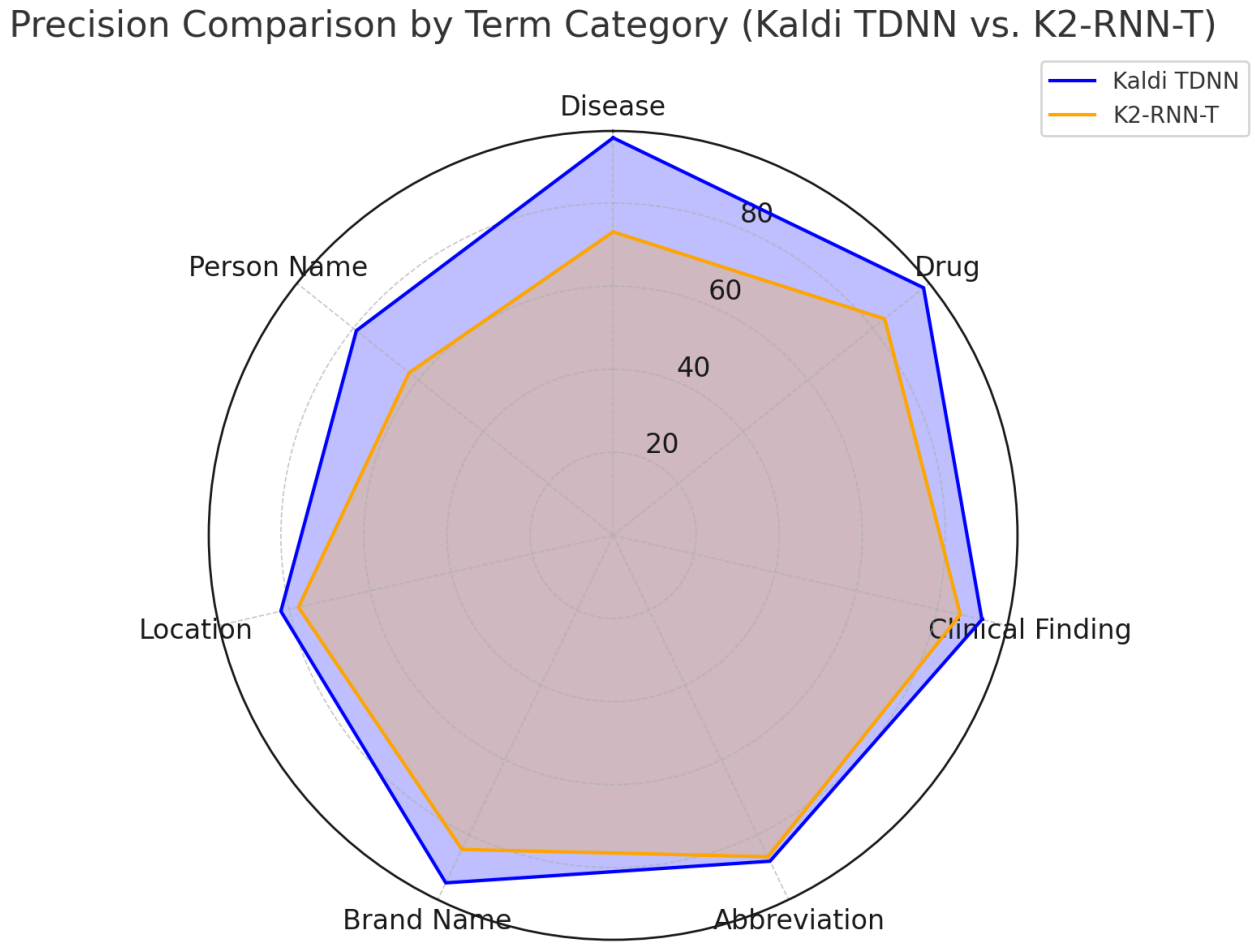

Figure 7

Precision, as an evaluation metric, reflects the proportion of correctly identified terms among all terms predicted by the model. High precision indicates a model's ability to avoid false positives, making it particularly important in clinical contexts where misidentifying irrelevant or incorrect terms can lead to misleading interpretations. As shown in Figure 7, the Kaldi TDNN model consistently outperforms the K2-RNN-T model in precision across all term categories. This suggests that Kaldi adopts a more conservative and stable recognition strategy, favoring accuracy over coverage.

Its consistently higher precision indicates a reduced tendency to overgenerate or mislabel entities, which is especially advantageous for structured medical terminology such as diseases and drugs. These results underscore Kaldi's strength in precise term identification, particularly in domains requiring high fidelity and trust in entity recognition.

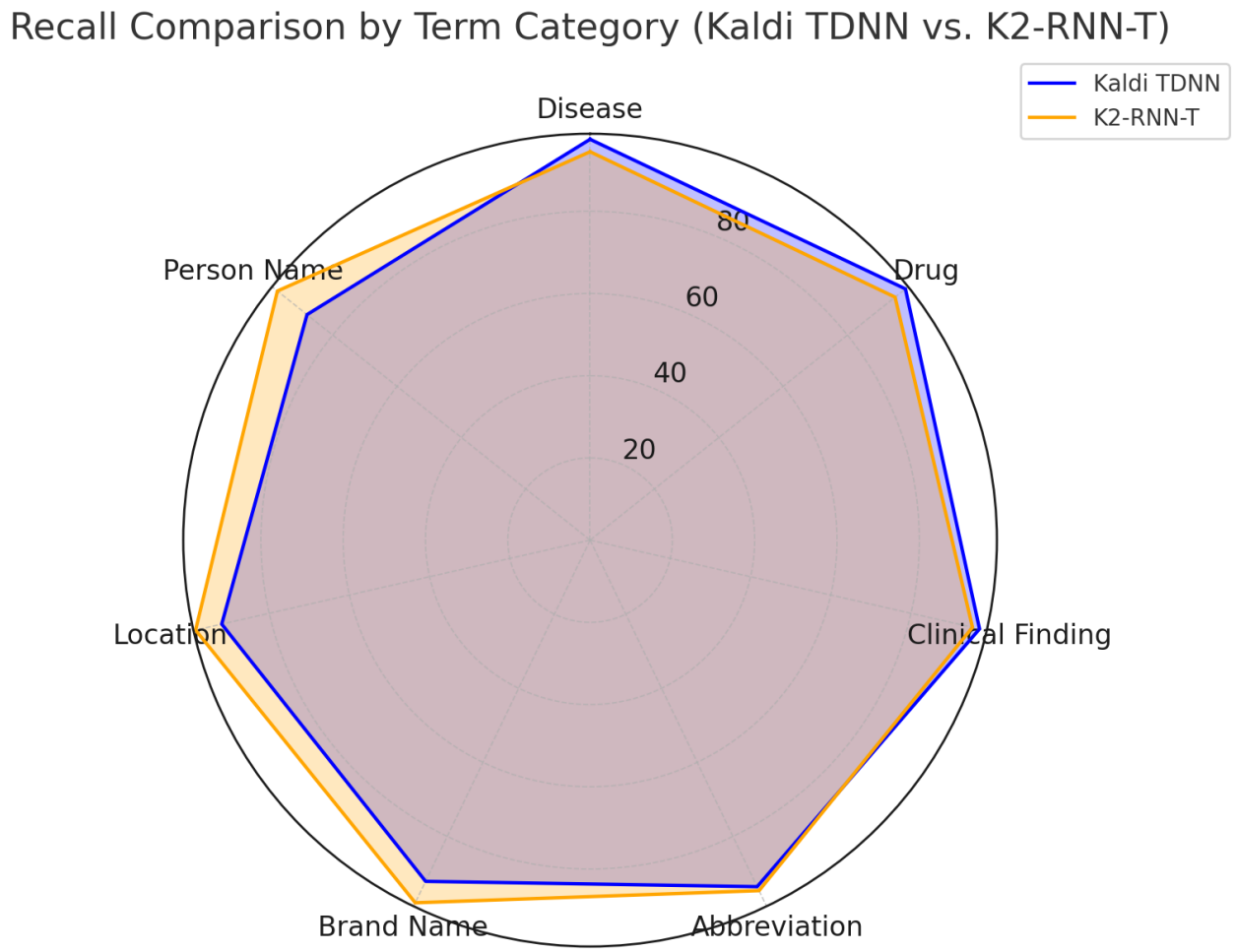

Figure 8

Recall measures a model's ability to correctly identify all relevant instances, reflecting how well it captures the full range of target entities. In the context of clinical terminology recognition, high recall is particularly desirable when the goal is to avoid missing important terms, especially in applications like information extraction or clinical decision support. As shown in Figure8, the K2-RNN-T model demonstrates slightly higher recall than Kaldi TDNN across most term categories, notably for Location (98.34%), Person Name (97.19%), and Brand Name (97.92%). These categories are typically characterized by greater lexical variability and less standardization. The superior recall observed in K2-RNN-T suggests that open-vocabulary models, by design, are more capable of capturing out-of-vocabulary or morphologically diverse terms, aligning with their intended flexibility and generalization strength. Nevertheless, Kaldi TDNN also maintains consistently high recall values (e.g., 97.51% for Disease, and 98.02% for Drug), demonstrating that its strong precision does

not come at the cost of substantial recall reduction.

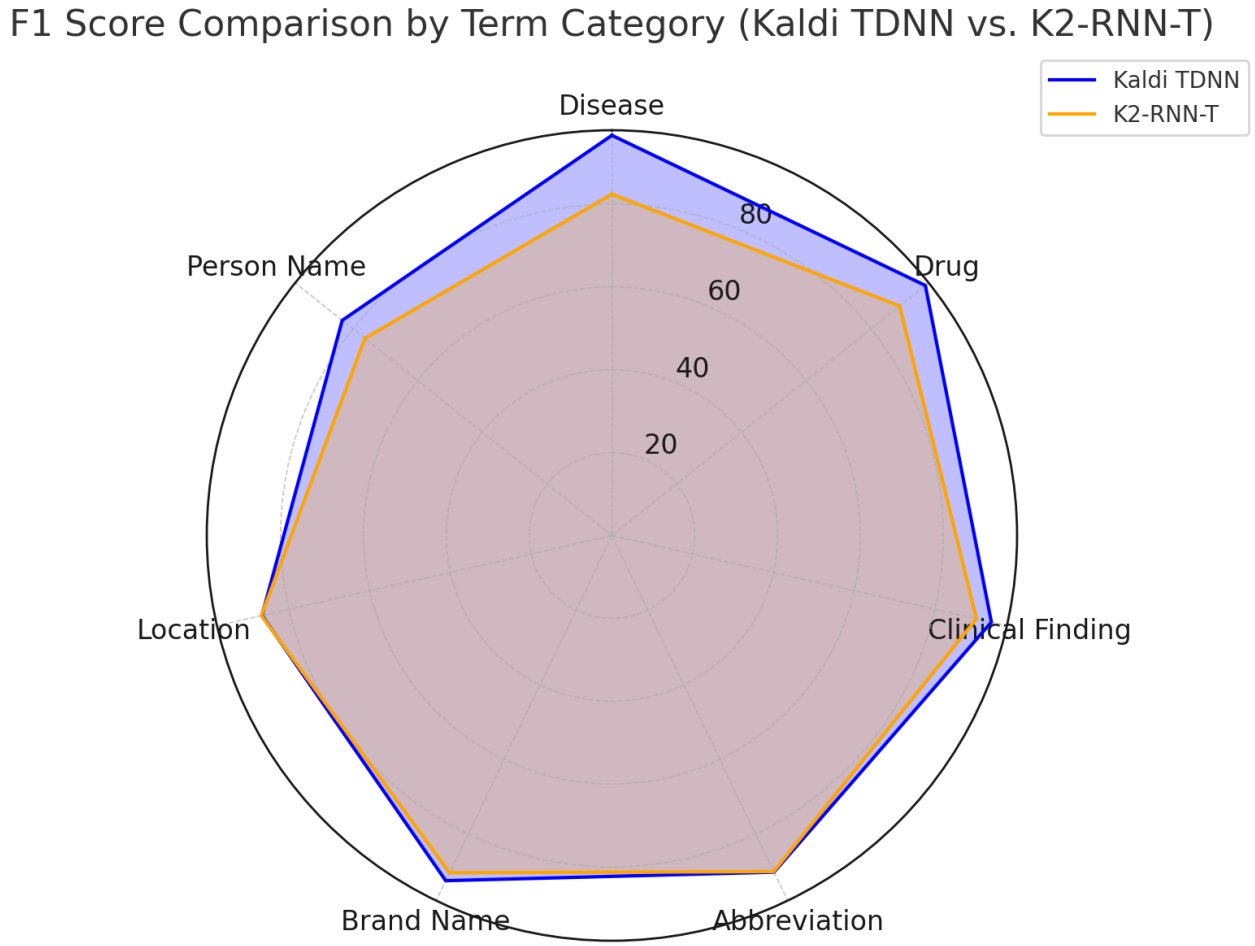

Figure 9

The F1 score, as the harmonic mean of precision and recall, provides a balanced measure of both correctness and completeness in recognition tasks. It is particularly informative in evaluating models' real-world usability, where both false positives and false negatives must be minimized. As illustrated in Figure 9, the Kaldi TDNN model achieves consistently higher F1 scores than K2-RNN-T across all term categories, with especially large margins observed in clinical categories such as Disease (96.60% vs. 82.41%), Drug (96.80% vs. 88.86%), and Person Name (83.29% vs. 76.31%), which suggests that Kaldi's closed-vocabulary design supports more stable and reliable recognition in domain-specific contexts. While the overall trend favors Kaldi, the performance gap between the two models narrows in entity types that are more linguistically diverse or loosely defined. For instance, Abbreviation (90.20% vs. 90.05%) and Location (86.67% vs. 86.83%) show nearly identical F1 scores. This indicates that open-vocabulary systems like K2-RNN-T may be better equipped to handle more variable or non-standardized entities, reflecting the influence of its broader recognition tendency combined with precision fluctuations across term types.

Overall, precision, recall, and F1 score are commonly evaluated together, as they offer complemen-

tary perspectives on model performance: precision reflects correctness, recall captures completeness, and F1 provides a balanced measure of both. The comparative results reveal that Kaldi TDNN tends to produce more consistent and higher precision and F1 scores across most clinical term categories, indicating a stable recognition behavior with fewer false positives. K2-RNN-T, on the other hand, achieves relatively higher recall in several categories, especially those with greater lexical variability, demonstrating broader coverage and a greater ability to detect non-standard or less predictable entities.

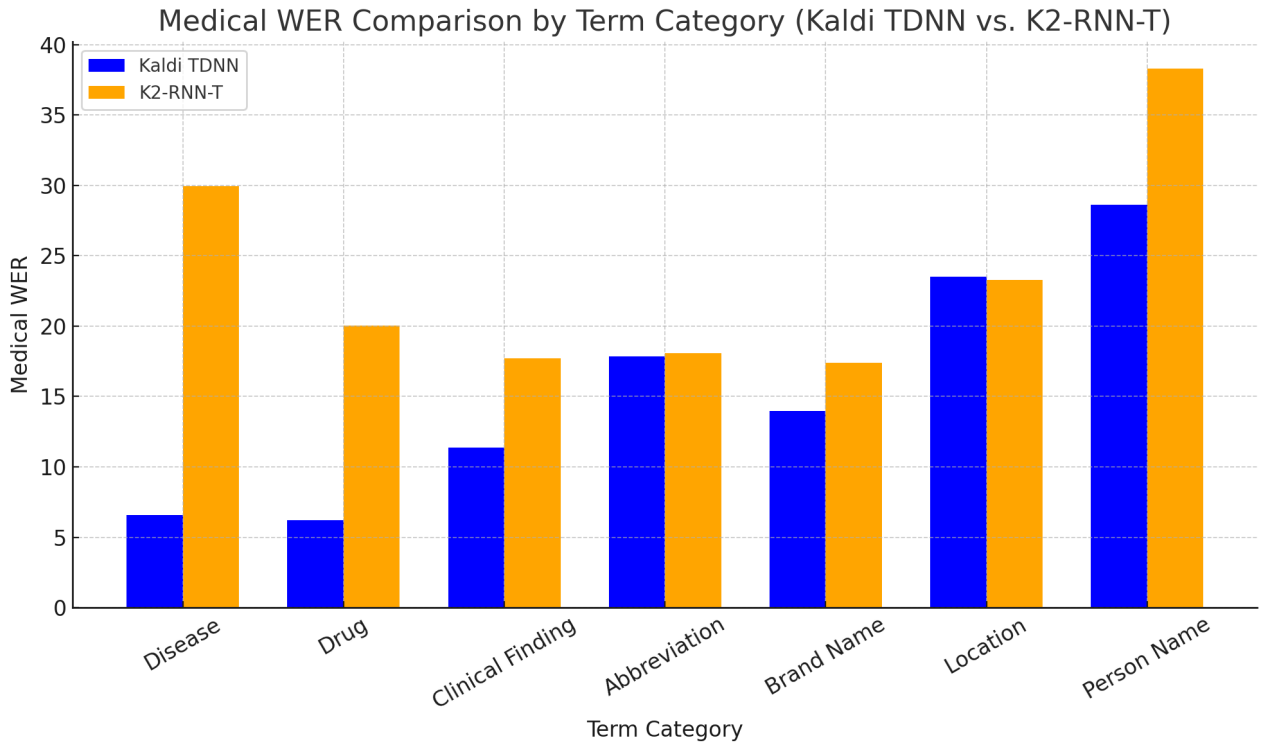

Figure 10

M-WER measures the word-level error rate for medically relevant terms, including substitutions, deletions, and insertions. A lower M-WER indicates better recognition accuracy, It is particularly critical in clinical contexts where term integrity is essential. As shown in Figure 10, the Kaldi TDNN model consistently achieves lower M-WER than K2-RNN-T across all term categories. The difference is especially pronounced in structured domains such as Disease (6.57% vs. 29.92%) and Drug (6.21% vs. 20.04%), while in categories like Abbreviation (17.85% vs. 18.06%) and Location (23.53% vs. 23.28%), the two models perform similarly.

These findings align with the earlier analysis of precision, recall, and F1 score. Kaldi TDNN shows more stable and balanced performance, particularly in domains with standardized terminology, whereas K2-RNN-T achieves broader recall in more variable term types. The correlation between Kaldi's high F1 scores and low M-WER values suggests not only metric consistency, but also greater robustness in preserving term structure and ensuring more consistent word-level outputs.

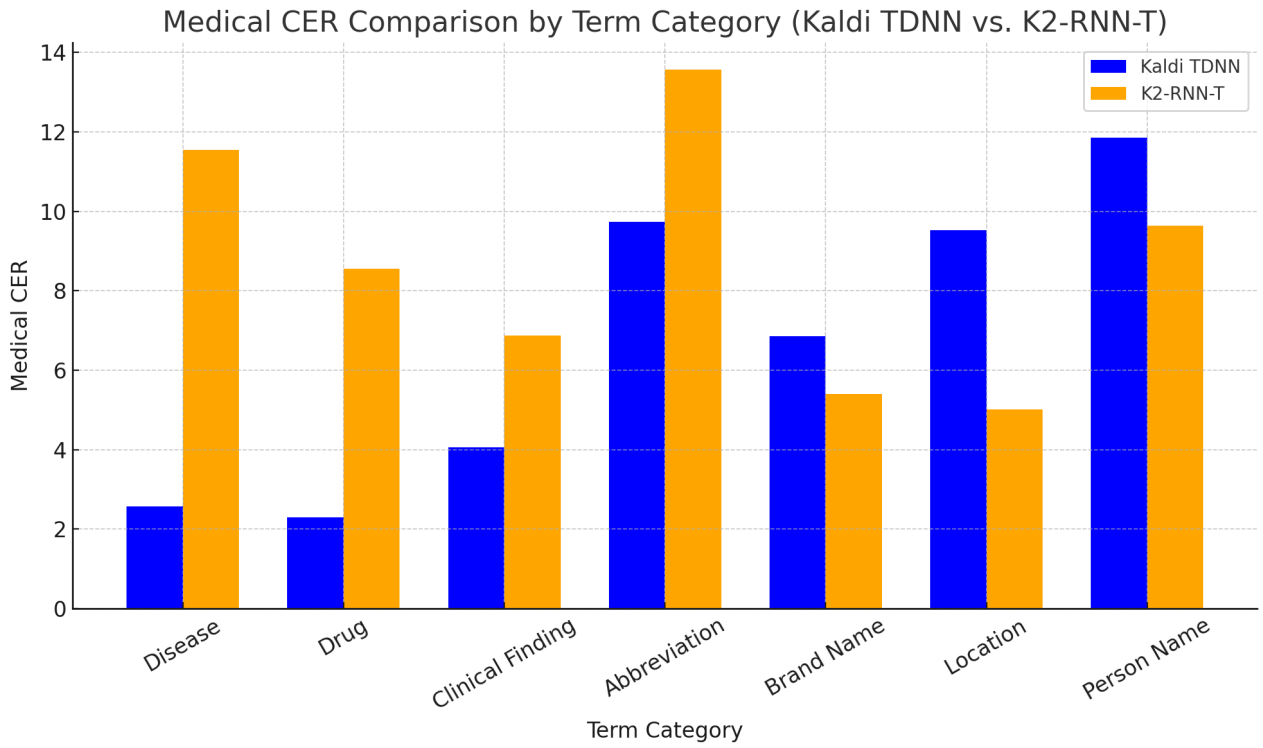

Figure 11

M-CER provides a fine-grained measure of recognition accuracy by evaluating character-level transcription fidelity. Unlike word-level metrics, M-CER captures partial recognition errors and is especially useful for analyzing performance on morphologically diverse or unpredictable term types. As illustrated in Figure 11, the Kaldi TDNN model shows lower M-CER scores in core clinical terminology categories such as Disease (2.57% vs. 11.54%), Drug (2.29% vs. 8.55%), and Clinical Finding (4.06% vs. 6.87%). These categories typically involve well-standardized and domain-specific terms with consistent lexical structures. Kaldi's closed-vocabulary architecture, coupled with strong alignment to pre-defined token sequences, likely contributes to its consistent handling of these terms at the character level. In contrast, K2-RNN-T demonstrates comparatively better M-CER performance in categories like Location (5.02% vs. 9.52%), Brand Name (5.40% vs. 6.86%), and Person Name (9.63% vs. 11.85%). These entity types are more lexically variable and context-dependent, which often involving proper names, place names, or proprietary terms that may not follow standardized linguistic patterns. The relatively lower M-CER achieved by K2-RNN-T in these cases suggests that its open-vocabulary design, which allows subword modeling and greater flexibility in decoding unfamiliar sequences, is better suited to capturing such variability.

These findings point to a nuanced shift: while Kaldi TDNN maintains consistent recognition in structured medical terms, K2-RNN-T begins to outperform in more variable entity categories when evaluated at the character level. This shift highlights the advantage of open-vocabulary models in adapting to linguistic diversity and capturing less predictable lexical items with finer granularity.

4.2.2 Category-wise Analysis

1. **Clinical Terminology**
   This subsection focuses on model performance in recognizing structured clinical terms, specifically disease, drug, and clinical finding categories. These categories are characterized by relatively standardized and domain-specific vocabulary, making them suitable for evaluating the models' ability to handle canonical medical expressions.

   Across disease, drug, and clinical finding categories, Kaldi TDNN consistently outperforms K2-RNN-T in precision, F1, M-WER, and M-CER. For example, in the disease category, Kaldi achieves an F1 of 96.60% with a M-WER of 6.57% and M-CER of 2.57%, compared to 82.41%, 29.92%, and 11.54% for K2-RNN-T. Similar patterns are observed for drug and clinical findings.While K2-RNN-T achieves higher recall, particularly for drug and clinical finding terms, this improvement often comes at the cost of increased recognition errors. This trade-off is likely attributable to its open-vocabulary design, which prioritizes term coverage over precision.

   The superior performance of Kaldi TDNN on these clinical categories may be attributed to its closed-vocabulary architecture, which is tightly aligned with a predefined lexicon and language model constraints. This setup favors accurate, consistent recognition of standardized terminology, particularly when term forms are relatively stable and appear frequently in training data. Furthermore, Kaldi's higher character-level accuracy, as evidenced by lower M-CER, suggests greater robustness in capturing the internal structure of complex medical terms. In contrast, K2-RNN-T's subword-based modeling offers greater flexibility, but this may lead to fragmented decoding or errors when encountering compound or morphologically rich medical words. As such, although recall is high, the model may face challenges in consistently preserving the exact structure of domain-specific terms.

2. **Named Entities**
   Compared to clinical terminology, the performance gap between Kaldi TDNN and K2-RNN-T narrows in named entity categories such as abbreviation, brand name, location, and person name. While Kaldi generally retains higher precision and F1, K2-RNN-T achieves comparable or better performance in recall, M-WER and M-CER, indicating more balanced outcomes in these categories.

   In the abbreviation category, both models achieve comparable F1 scores. However, K2 exhibits a higher M-CER, likely because abbreviations are short and character-sensitive by nature. Even minor errors can significantly affect the entire word, thereby reducing the typical distinction between character-level and word-level accuracy. For brand names and locations, K2 achieves higher recall and lower M-CER, suggesting better handling of less standardized, possibly institution-specific terms. In person names, K2 again leads in M-CER, while Kaldi maintains an advantage in F1 and M-WER, reflecting more stable recognition output.

   These patterns reflect the nature of named entities, which often involve greater lexical diversity, such as foreign names or non-standard spelling conventions. K2-RNN-T's open-vocabulary and subword-based modeling appears better suited to such variability, offering more flexible recognition and accurate character-level transcription. In contrast, Kaldi's fixed-vocabulary structure, while effective in minimizing false positives, may be less adaptive to

unpredictable or out-of-vocabulary forms.

4.3 Analysis of Error Types

To complement the quantitative evaluation, this section presents a qualitative analysis of recognition errors.We analyzed the errors and grouped them into three types: substitution (the transcribed word is different from the ground truth), approximation (the transcribed word differs by a few characters from the ground truth, but has a similar pronunciation), and truncation (the transcribed word omits certain characters from the ground truth). These error types were used in the literature (Luo, Zhou, Adelgais, et al., 2025) to summarize and categorize speech recognition errors.We aim to better understand the characteristic patterns and challenges each model encounters during medical term recognition.

Error Type	Category	Ground Truth Transcript	Kaldi-TDNN Inference Transcript	K2-RNN-T Inference Transcript
Substitution	Disease	**Hemorroïden**	(Missing)	Variesnemerïde
	Disease	**Hemiparese**	**Hemiparese**	Hemi parijse
	Drug	**Naproxen**	**Naproxen**	Proxen
	Drug	**Colecalciferol**	**Colecalciferol**	Colicalsi voor
	Brand Name	**Oromorph**	**Oromorph**	Ongemak
Approximation	Clinical Finding	**Hypertonie**	Hypotonie	Hypotonie
	Disease	**Angiopathie**	**Angiopathie**	Anchiopathie
	Drug	**Pantoprozol**	Pantoprazol	Pantro prosool
	Brand Name	**Roho**	**Roho**	Rohol
	Location	**Berkenlaan**	**Berkenlaan**	Berkelen
	Person Name	**Simons**	**Simons**	Siemens
Truncation	Abbreviation	**IADL**	**IADL**	ADL
	Abbreviation	**AAT**	**AAT**	AT
	Clinical Finding	**Lewy bodydementie**	**Lewy bodydementie**	Bodydementie
	Location	**Emmeloord**	**Emmeloord**	Mmeloord

Figure 12: Examples of recognition errors by type for both models

Figure 12 provides representative examples of these error types. In the table, terms in the gold truth transcript are bolded. In the model outputs (Kaldi and K2), words that correctly match the gold reference are also bolded, while mistranscribed words are left unformatted. A notable pattern emerges when comparing the two models. The Kaldi TDNN model, being a closed-vocabulary system, tends to omit or entirely miss terms that are not present in its lexicon, resulting in silence or unrelated words. This behavior reflects its dependency on a predefined word list and phonetic dictionary, limiting its flexibility when encountering uncommon or OOV terms. In contrast, the K2 model, based on an open-vocabulary subword architecture, exhibits more flexible decoding behavior. It attempts to approximate unfamiliar terms by leveraging subword units or common phoneme patterns. However, this flexibility sometimes leads to near-homophone errors, or overgeneralizations, as seen in the example where “Hemiparese” (“Hemiparese” in English) was recognized as “Hemi parijse”(“Hemi Parisian” in English), introducing both a phonetic distortion and a semantic shift.

The next six examples in Figure 12 represent approximation errors, where the ASR models transcribe terms into phonetically similar but semantically different alternatives. While these errors may appear minor in form, they can carry major clinical implications. For instance, transcribing “Hypertensie” (high blood pressure) as “Hypotensie” (low blood pressure) or “Hypertonie” (increased muscle tone) as “Hypotonie” (reduced muscle tone) may completely reverse the intended meaning, potentially affecting diagnosis or treatment decisions. A closer comparison between the two models reveals that Kaldi TDNN tends to produce more orthographically complete and standardized spellings, especially for clinical or named entity terms. Even when variations exist, such as “Pantoprozol” vs. “Pantoprazol”, both referring to “Pantoprazole” in English, Kaldi’s outputs generally align more closely with expected medical vocabulary entries. In contrast, the K2 model shows greater variability in spelling, often generating outputs that approximate the correct term based on phonetic cues but diverge from standardized forms. This behavior is likely due to K2’s open-vocabulary subword-based architecture, which allows it to reconstruct unfamiliar terms creatively, but not always accurately. As a result, it is more prone to producing plausible-looking but incorrect medical terms, especially in cases where the acoustic signal is ambiguous.

The final four examples in Figure 12 illustrate truncation errors, where only part of a term is transcribed, resulting in incomplete or clipped forms. Although these may involve the omission of just a single character or syllable, the consequences can still be significant, particularly in the medical domain, where even subtle distortions can impact the clarity or accuracy of documentation. A recurring pattern in these examples is the vulnerability of the K2 model to such truncations. As an open-vocabulary, subword-based system, K2 relies on probabilistic combinations of subword units. While this enables it to handle unseen words more flexibly, it also makes it more susceptible to generating incomplete forms, especially for long or compound terms or named entities like locations and personal names, where variability in spelling and unfamiliarity further increase the risk of early cutoffs. In contrast, the Kaldi TDNN model demonstrates a relatively more stable transcription pattern in these cases. Being a closed-vocabulary model, Kaldi tends to either fully recognize a known term or omit it entirely if it is not in the lexicon, thus avoiding partial outputs. This can be seen as a built-in filtering mechanism that maintains lexical integrity, albeit at the expense of recall. These findings highlight another critical trade-off: while K2’s architecture promotes broader coverage, it does so at the risk of fragmenting complex terms, whereas Kaldi’s rigid structure, though more conservative, may better preserve term boundaries when recognition succeeds.

5 Discussion

5.1 Validation of the Hypothesis

The central hypothesis of this study posited that although open-vocabulary ASR models often achieve lower overall WER in general transcription tasks, closed-vocabulary models may exhibit more stable and controlled performance in recognizing domain-specific clinical terminology.

The empirical results support this hypothesis. As shown in Figure 6, the open-vocabulary model (K2 RNN-T) achieved lower overall WER (8.67%) and CER (4.03%) across all terms, indicating strong general transcription ability. However, the closed-vocabulary model (Kaldi TDNN) showed lower M-WER (16.78% vs. 25.16%) and M-CER (6.68% vs. 8.86%), suggesting more consistent recognition of structured clinical terms at both word and character levels.

In terms of evaluation metrics, Kaldi obtained higher precision (88.00%) and F1 score (90.84%), while K2 achieved higher recall (95.80%). These results reflect two different recognition strategies: one favoring coverage and flexibility, and the other favoring selectivity and control. The relatively higher recall of K2 may reflect its subword-based decoding mechanism, which facilitates recognition of variable and out-of-vocabulary terms. Meanwhile, Kaldi's performance remains more consistent in structured terminology, likely due to its fixed vocabulary design.

Together, these results support the hypothesis in structured domains such as medical term recognition, while also highlighting the complementary strengths of both systems: coverage versus consistency. The findings further justify the use of both word-level and character-level metrics to capture recognition performance across multiple linguistic dimensions.

5.2 Validation of the First Subquestion

Subquestion 1: How can clinical terms and named entities be effectively extracted from medical transcripts for evaluation, and which methods are appropriate for identifying different term categories?

This study employed a dual extraction framework combining the SNOMED CT clinical terminology system and a Dutch-language Named Entity Recognition (NER) model (spaCy's nl_core_news_lg) to identify two major lexical content types within medical transcripts: (1) standardized clinical terms (e.g., diseases, drugs, clinical findings), and (2) semantically relevant named entities (e.g., personal names, locations, brand names, abbreviations).

To accommodate the variability of ASR outputs, a fuzzy matching strategy was used using RapidFuzz and n-gram expansion. This approach allowed for approximate string matches, accounting for common ASR errors such as phonetic variation, spelling inconsistencies, and fragmentary outputs. A similarity threshold of 75 was adopted to balance between recall and specificity. This method-

ology successfully enhanced retrieval robustness, as evidenced by increased coverage of drug and clinical finding terms under fuzzy matching.

5.3 Validation of the Second Subquestion

Subquestion 2: How do the two ASR models compare in recognizing these terms, based on evaluation metrics and error patterns, and do they show clear strengths or limitations in specific categories? The comparative evaluation of the two ASR models demonstrated distinct performance patterns across term categories and metric dimensions:

The closed-vocabulary Kaldi-TDNN model demonstrated consistently higher precision, F1 scores, and lower M-WER and M-CER in clinical term categories such as Disease (Precision: 95.72%, F1: 96.60%, M-WER: 6.57%, M-CER: 2.57%) and Drug (F1: 96.80%, M-WER: 6.21%, M-CER: 2.29%). Similar advantages are observed in Clinical Findings (F1: 93.98%, M-WER: 11.36%, M-CER: 4.06%). These results indicate that Kaldi's structured decoding and reliance on a predefined lexicon offer substantial benefits when transcribing standardized medical terms. However, the model tends to omit or misrecognize out-of-vocabulary terms, a limitation inherent to its closed-vocabulary design.

The open-vocabulary K2-RNN-T model, on the other hand, showed higher recall across all categories, especially in those with greater lexical variability. For instance, in Location (Recall: 98.34%) and Person Name (Recall: 97.19%), K2 achieved significantly higher recall than Kaldi, alongside lower M-CER in Person Name (9.63% vs. 11.85%) and Location (5.02% vs. 9.52%). In Brand Name, both models performed comparably in F1 scores (Kaldi: 92.49% vs. K2: 90.38%), but K2 achieved lower M-CER (5.40% vs. 6.86%). This suggests that K2's subword-based and flexible decoding strategies may offer advantages in handling non-standardized or institution-specific terms. Despite these strengths, K2-RNN-T's gains in recall often come with elevated word-level error rates. For example, in Disease, its M-WER (29.92%) and M-CER (11.54%) are considerably higher than Kaldi's, pointing to greater substitution or approximation errors. Notably, Person Names yielded the lowest F1 scores for both models (Kaldi: 83.29%, K2: 76.31%) and the highest M-WER values (Kaldi: 28.64%, K2: 38.31%), reflecting shared difficulty in recognizing unfamiliar or acoustically ambiguous names.

Qualitative error analysis further corroborated these findings: Kaldi-TDNN tended to fully omit unknown terms, whereas K2-RNN-T often generated near-miss approximations or truncated outputs. These differences highlight a core trade-off: Kaldi-TDNN prioritizes lexical integrity at the expense of coverage, whereas K2-RNN-T sacrifices standardization for recall. In conclusion, both models exhibit complementary strengths. Kaldi-TDNN is more suited for high-precision tasks involving standardized terminology, while K2-RNN-T excels in recognizing lexically diverse or spontaneously expressed terms. These findings suggest that future ASR development in the clinical domain may benefit from hybrid or task-adaptive strategies that leverage the strengths of both modeling paradigms.

5.4 Limitations

While this study presents a comprehensive evaluation of closed- and open-vocabulary ASR systems in the context of Dutch medical speech, several limitations should be acknowledged that may influence the generalizability and scope of the findings.

1. **Data Scope and Representativeness**
   The dataset used in this study, although rich in authentic Dutch clinical dialogue and professionally transcribed, was drawn exclusively from long-term care institutions using Gerimedica's platform. As such, it may not fully represent the linguistic diversity, acoustic conditions, and interactional styles found in other healthcare settings such as emergency rooms, surgical units, or outpatient clinics. Furthermore, the dataset contains relatively structured and formal speech, which may limit insights into more spontaneous or noisy clinical interactions.

2. **Domain and Language Constraints**
   The dataset used in this study, although rich in authentic Dutch clinical dialogue and professionally transcribed, was drawn exclusively from long-term care institutions using Gerimedica's platform. As such, it may not fully represent the linguistic diversity, acoustic conditions, and interactional styles found in other healthcare settings such as emergency rooms, surgical units, or outpatient clinics. Furthermore, the dataset contains relatively structured and formal speech, which may limit insights into more spontaneous or noisy clinical interactions.

3. **Model Design and Decoding Constraints**
   The closed-vocabulary and open-vocabulary systems evaluated in this study differ not only in their lexical design (lexicon-based vs. subword-based) but also in decoding behavior. For example, the open-vocabulary model (K2-RNN-T) encountered instability when processing longer audio files, leading to decoder crashes and incomplete transcript generation. This affected transcript alignment: while the gold standard reference set contained 8595 utterances, only 8160 hypothesis transcripts were generated by the K2 model. This discrepancy may bias comparative metrics and underrepresent errors in longer or more complex inputs.

4. **Level of Evaluation and Entity Coverage**
   Although the study employed fine-grained metrics (e.g., M-WER, entity-specific F1 scores) and fuzzy matching for approximate recognition, certain semantic errors may still evade detection, particularly those involving paraphrased or partially expressed concepts that fall outside strict term boundaries. Moreover, the entity extraction process was limited to selected categories (e.g., SNOMED-based terms and named entities such as PERSON, LOC, BRAND, ABBR), and did not include all clinically relevant types such as temporal expressions, numeric quantities, or nested multi-entity constructs. This selection, while practical, introduces some bias in error attribution and recall estimation.

5. **Occurrence of Code-switching**
   Some transcripts in the dataset include English terms that appear within otherwise Dutch sentences. These terms often relate to physiotherapy exercises or medications, such as leg press, range of motion. This shows that intra-utterance code-switching does occur in real-world Dutch clinical speech.

While this study did not focus on analyzing or evaluating code-switched utterances in detail, a small number of English terms were identified in the data, as shown in Figure 13. These examples confirm the presence of mixed-language usage in this domain. Future work may explore this further, especially in more spontaneous or complex cases of code-switching. It may also be valuable to compare how different ASR models handle such code-switched inputs and to assess their relative strengths in mixed-language conditions.

Gold term	Kaldi-TDNN	K2-RNN-T
leg press	legpress	leg press
sit to stand	sit to stand	sit to stand
range of motion	range of motion	range of motion
nutridrink	nutridrink	nutridrink
juice	juice	juice
style	stijl	style
pulley	pully	poeie

Figure 13: Instances of Intra-utterance Code-switching in the Dataset

In summary, while the methodological rigor and focused design of this study provide valuable insights into ASR performance in Dutch medical contexts, future work should address these limitations by incorporating more diverse clinical datasets, multilingual evaluation frameworks, code-switching and broader semantic coverage to enhance both internal validity and external generalizability.

6 Conclusion

6.1 Summary of the Main Contributions

This study provides a detailed comparative analysis of closed- and open-vocabulary automatic speech recognition (ASR) systems in the clinical domain, specifically targeting Dutch healthcare speech. Several core contributions can be highlighted:

1. **Use of Real-World Clinical Speech Data**
   Unlike prior studies that rely heavily on benchmark datasets, this research utilized professionally transcribed Dutch medical consultation data from Gerimedica. This real-world source enhances the ecological validity and practical relevance of the findings, contributing to evidence-based evaluation of ASR deployment in actual healthcare environments.

2. **Recognition Challenges of Clinical and Named Entities**
   The study moves beyond aggregate WER metrics and focuses on the recognition accuracy of medically salient terms, such as diseases, drugs, brand names, and personal identifiers. This term-level granularity offers critical insight for downstream clinical NLP applications, where recognition fidelity of domain-specific vocabulary is paramount.

3. **Closed- vs. Open-Vocabulary Modeling Comparison**
   A direct performance comparison between a lexicon-based closed-vocabulary model (Kaldi-TDNN) and a subword-based open-vocabulary model (K2 RNN-T) was conducted. The analysis reveals complementary strengths—Kaldi demonstrated greater precision and lower M-WER and M-CER, while K2 offered broader lexical recall, overall WER and CER underscoring theoretical trade-offs and guiding practical system selection for medical ASR tasks.

4. **Customized Evaluation Metrics and Error Categorization**
   This study adopts a comprehensive suite of evaluation metrics to assess ASR performance, extending beyond standard WER and CER to include domain-specific measures such as M-WER, M-CER, and entity-level F1 scores. This multi-metric approach offers a more granular and clinically meaningful evaluation than relying on surface-level scores alone. In addition to these quantitative metrics, the study incorporates a qualitative analysis of recognition errors, categorizing them into substitution, approximation, and truncation types. This layered approach not only reveals surface-level performance differences but also provides deeper insights into the models' behavior in real-world clinical contexts, highlighting practical implications for downstream healthcare applications.

5. **Multi-Source Terminology Recognition Design**
   A multi-step extraction pipeline was developed, integrating SNOMED CT, spaCy-based NER, and rule-based filters. This ensured high recall and semantic precision for both clinical terms and named entities, and demonstrates methodological innovation in domain-specific ASR evaluation.

6.2 Future Work

Building upon the contributions of this study, several directions are proposed for future research to further enhance the robustness, adaptability, and clinical relevance of ASR systems in medical settings:

1. **Clinical Domain Fine-Tuning and Model Fusion**
   Future work should explore domain-specific fine-tuning of open-vocabulary models using Dutch medical corpora, including unsupervised and self-supervised learning techniques. Additionally, hybrid or ensemble models that combine lexicon-constrained and subword-based architectures may better balance the precision of closed-vocabulary models with the flexibility of open-vocabulary systems.

2. **Broader Entity and Ontology Integration**
   The current evaluation focuses on a subset of medical and named entity types. Future extensions may include temporal expressions, dosage instructions, institutional identifiers, and complex nested entities. Integration with additional ontologies such as ICD, LOINC, or UMLS could enable richer semantic evaluation and cross-domain interoperability.

3. **Downstream Task Evaluation and Clinical Impact Analysis**
   While this study emphasizes ASR term recognition, future research should examine how recognition errors affect downstream clinical NLP tasks, such as automatic diagnosis coding, decision support, or summarization. Measuring the semantic and practical consequences of misrecognitions will help contextualize ASR performance within clinical workflows.

In sum, future work should aim not only to improve model performance in isolation, but also to expand the ecological scope and translational value of clinical ASR systems. This includes designing context-aware, terminology-sensitive, and institution-scalable models capable of supporting real-world healthcare delivery.

References

Afonja, T., Olatunji, T., Ogun, S., Etori, N. A., Owodunni, A., & Yekini, M. (2024). *Performant asr models for medical entities in accented speech.* Retrieved from https://doi.org/10.48550/arXiv.2406.12387 doi: 10.48550/arXiv.2406.12387

Alexiadou, A. (2020). Compound formation in language mixing. *Frontiers in Psychology*, *11*, 1021. Retrieved from https://doi.org/10.3389/fpsyg.2020.01021 doi: 10.3389/fpsyg.2020.01021

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in neural information processing systems* (Vol. 33, pp. 12449–12460).

Banga, A., Hanseen, E., Neijit, A., & Schreuder, R. (2013). Preference for linking element-en in dutch noun-noun compounds: native speakers and second language learners of dutch. *Morphology*, *23*, 33–56. doi: 10.1007/s11525-0139211-y

Blackley, S. V., Huynh, J., Wang, L., Korach, Z., & Zhou, L. (2019). Speech recognition for clinical documentation from 1990 to 2018: A systematic review. *Journal of the American Medical Informatics Association*, *26*(4), 324–338. Retrieved from https://doi.org/10.1093/jamia/ocy179 doi: 10.1093/jamia/ocy179

Boumans. (1998). *The syntax of code-switching: Analysing moroccan arabic/dutch conversation*. Tilburg: Tilburg University Press.

Chan, W., Jaitly, N., Le, Q. V., & Vinyals, O. (2015). *Listen, attend and spell.* Retrieved from https://doi.org/10.48550/arXiv.1508.01211 doi: 10.48550/arXiv.1508.01211

Chang, E., & Sung, S. (2024). Use of snomed ct in large language models: Scoping review. *JMIR Medical Informatics*, *12*(1), e62924.

Chiu, C. C., Tripathi, A., Chou, K., Co, C., Jaitly, N., Jaunzeikare, D., & Zhang, X. (2017). *Speech recognition for medical conversations.*

Chu, W., Chang, P., & Xiao, J. (2021). Extending pronunciation dictionary with automatically detected word mispronunciations to improve paii's system for interspeech 2021 non-native child english close track asr challenge. In *Proceedings of interspeech 2021* (pp. 1319–1323).

Garg, D. (2019). *Compressed dnn based automatic speech recognition engine* (Doctoral dissertation). Indian Institute of Technology, Madras.

Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Huh, J., Park, S., Lee, J. E., & Ye, J. C. (2023). *Improving medical speech-to-text accuracy with vision-language pre-training model.*

Juan. (2024). Retrieved from https://publications.idiap.ch/attachments/papers/2024/Juan_THESIS_2024.pdf (Accessed March 16, 2025)

Karita, S., Watanabe, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., & Yamamoto, R. (2019). A comparative study on transformer vs rnn in speech applications. In *Ieee automatic speech recognition and understanding workshop (asru)* (pp. 449–456).

Khassanov, Y. (2020). *Language model domain adaptation for automatic speech recognition systems* (Master's thesis, Nanyang Technological University). Retrieved from https://dr.ntu.edu.sg/bitstream/10356/141323/2/Thesis_Khassan.pdf

Kuang, F., Guo, L., Kang, W., Lin, L., Luo, M., Yao, Z., & Povey, D. (2022). *Pruned rnn-t for fast,*

memory-efficient asr training.

Laparra, E., Mascio, A., Velupillai, S., & Miller, T. (2021). A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records. *Yearbook of Medical Informatics*, *30*(01), 239–244.

Li, J. (2022). Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, *11*(1).

Liu, F., Tur, G., Hakkani-Tur, D., & Yu, H. (2011). Towards spoken clinical-question answering: evaluating and adapting automatic speech-recognition systems for spoken clinical questions. *Journal of the American Medical Informatics Association*, *18*(5), 625–630. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3168303/ doi: 10.1136/amiajnl-2010-000071

Luo, X., Zhou, L., Adelgais, K., et al. (2025). Assessing the effectiveness of automatic speech recognition technology in emergency medicine settings: a comparative study of four ai-powered engines. *Journal of Healthcare Informatics Research*. Retrieved from https://doi.org/10.1007/s41666-025-00193-w doi: 10.1007/s41666-025-00193-w

Melton, G. B., Parsons, S., Morrison, F. P., Rothschild, A. S., Markatou, M., & Hripcsak, G. (2006). Inter-patient distance metrics using snomed ct defining relationships. *Journal of Biomedical Informatics*, *39*(6), 697–705.

Meripo, N. V., & Konam, S. (2022). *Asr error detection via audio-transcript entailment.*

Mustafa, M. B., Yusoof, M. A., Khalaf, H. K., Abushariah, A. A. R. M., Kiah, M. L. M., Ting, H. N., & Muthaiyah, S. (2022). Code-switching in automatic speech recognition: The issues and future directions. *Applied Sciences*, *12*(19), 9541.

Navarro, D. F., Ijaz, K., Rezazadegan, D., Rahimi-Ardabili, H., Dras, M., Coiera, E., & Berkovsky, S. (2023). Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. *International Journal of Medical Informatics*, *177*, 105122.

Peddinti, V., Povey, D., & Khudanpur, S. (2015, September). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Proceedings of interspeech* (pp. 3214–3218).

Poder, T. G., Fisette, J.-F., & Déry, V. (2018). Speech recognition for medical dictation: Overview in quebec and systematic review. *Journal of Medical Systems*, *42*(5), 89. Retrieved from https://doi.org/10.1007/s10916-018-0947-0 doi: 10.1007/s10916-018-0947-0

Popović, B., Pakoci, E., & Pekar, D. (2020). *Automatic speech recognition system for dictating medical findings.* (Unpublished manuscript)

Prabhavalkar, R., Hori, T., Sainath, T. N., Schlüter, R., & Watanabe, S. (2023). End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *32*, 325–351.

Scharenborg, O., Besacier, L., Black, A., Hasegawa-Johnson, M., Metze, F., Neubig, G., ... Dupoux, E. (2020). Speech technology for unwritten languages. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 964–975. Retrieved from https://doi.org/10.1109/TASLP.2020.2973896 doi: 10.1109/TASLP.2020.2973896

Seiffert, M. (2021). *Rapidfuzz: Rapid fuzzy matching.* https://github.com/maxbachmann/RapidFuzz. (Accessed: 2025-06-07)

Sennrich, R., Haddow, B., & Birch, A. (2016). *Neural machine translation of rare words with subword units.* Retrieved from https://doi.org/10.48550/arXiv.1508.07909 doi: 10

.48550/arXiv.1508.07909

Shen, G. (2022). *Does where words come from matter? leveraging self-supervised models for multilingual asr and lid* (Doctoral dissertation).

Sitaram, S., Chandu, K. R., Rallabandi, S. K., & Black, A. W. (2019). *A survey of code-switched speech and language processing.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems (neurips)* (Vol. 30).

Wang, Q. (2023). *Code-switching detection techniques and language modeling strategies for automatic speech recognition* (Doctoral dissertation). National University of Singapore.

Whetten, R., & Kennington, C. (2023). Evaluating and improving automatic speech recognition using severity. In *Proceedings of the 22nd workshop on biomedical natural language processing and bionlp shared tasks* (pp. 79–91). Retrieved from https://aclanthology.org/2023.bionlp-1.6/ doi: 10.18653/v1/2023.bionlp-1.6

Wu, Y., Jiang, M., Xu, J., Zhi, D., & Xu, H. (2018, April). Clinical named entity recognition using deep learning models. In *Amia annual symposium proceedings* (Vol. 2017, p. 1812).

Yılmaz, E. e. a. (2018). Semi-supervised acoustic model training for speech with code-switching. *Speech Communication*, *105*, 12–22.

Zhou, W., Zeineldeen, M., Zheng, Z., Schlüter, R., & Ney, H. (2021). *Acoustic data-driven subword modeling for end-to-end speech recognition.*

Zuluaga-Gomez, J. P. (2024). *Low-resource speech recognition and understanding for challenging applications* (Doctoral dissertation). École Polytechnique Fédérale de Lausanne.

Appendices

A SNOMED Category Definitions

Term Type	SNOMED Concept Description	SNOMED CT ID
Drug	Pharmaceutical / biologic product	373873005
Drug	Substance	105590001
Disease	Infectious diseases	40733004
Disease	Chronic diseases	27624003
Disease	Cardiovascular disorders	49601007
Disease	Nervous system disorders	118940003
Disease	Respiratory disorders	50043002
Disease	Musculoskeletal disorders	928000
Disease	Mental disorders	74732009
Clinical Finding	Functional finding	118228005
Clinical Finding	Nutritional finding	300893006
Clinical Finding	General symptom description	162408000
Clinical Finding	Mental state or behavioral findings	384821006

Figure 14: SNOMED Category Definitions

B Clinical Terminology Coverage Tables (Variety&Frequency)

Term Type	SNOMED Terms	Exact Match (Unique)	Fuzzy Match (Unique)	Exact Coverage(%)	Fuzzy Coverage (%)
Disease	141,100	30	67	0.021%	0.047%
Drug	20,703	12	160	0.058%	**0.773%**
Clinical Finding	44,203	56	194	**0.127%**	0.439%

Figure 15: Clinical Terminology Coverage Analysis–Variety

Term Type	Transcript Count	Exact Matches (All Mentions)	Fuzzy Matches (All Mentions)	Exact mention Rate (%)	Fuzzy mention Rate (%)
Disease	3828	38	160	0.99%	4.18%
Drug	3828	19	719	0.50%	18.78%
Clinical Finding	3828	107	800	**2.79%**	**20.90%**

Figure 16: Clinical Terminology Coverage Analysis–Frequency

C Definitions of Named Entities

Label	Description
PERSON	Names of individuals such as patients or healthcare workers
ABBR	Abbreviations commonly used in medical speech
LOC	Geographic or institutional references
BRAND	Product or manufacturer brand names
CARDINAL_UNIT	Dosage and quantity expressions (excluded)
OTHER	Miscellaneous terms with low semantic relevance (excluded)

Figure 17: Definitions of Named Entities

D Named Entities Coverage Tables (Variety&Frequency)

Label	Count	Percentage
PERSON	505	27.51%
CARDINAL_UNIT	442	24.07%
ABBR	375	20.42%
LOC	253	13.78%
OTHER	176	9.58%
BRAND	85	4.63%

Figure 18: Named Entities Coverage Analysis – Frequency

Label	Count	Percentage
PERSON	326	40.75%
CARDINAL_UNIT	138	17.25%
ABBR	132	16.50%
LOC	131	16.38%
OTHER	37	4.62%
BRAND	36	4.50%

Figure 19: Named Entities Coverage Analysis–Variety

E Evaluation and Comparision of Kaidi TDNN and K2-RNN-T Recognition Results

Category	Model	Precision	Recall	F1	Medical WER	Medical CER
Disease	Kaldi	95.72	97.51	96.60	6.57	2.57
	K2	73.08	94.48	82.41	29.92	11.54
Drug	Kaldi	95.61	98.02	96.80	6.21	2.29
	K2	83.62	94.80	88.86	20.04	8.55
Clinical Finding	Kaldi	91.08	97.07	93.98	11.36	4.06
	K2	85.76	95.32	90.29	17.70	6.87
Abbreviation	Kaldi	87.07	93.57	90.20	17.85	9.74
	K2	85.89	94.64	90.05	18.06	13.57
Brand Name	Kaldi	92.86	92.13	92.49	13.97	6.86
	K2	83.93	97.92	90.38	17.39	5.40
Location	Kaldi	82.06	91.82	86.67	23.53	9.52
	K2	77.73	98.34	86.83	23.28	5.02
Person Name	Kaldi	79.07	87.98	83.29	28.64	11.85
	K2	62.81	97.19	76.31	38.31	9.63

Figure 20: Evaluation and Comparision of Kaidi TDNN and K2-RNN-T Recognition Results (Blue fonts represent better performance)

F Declaration of AI Use

During the preparation of this thesis, I used *ChatGPT (OpenAI, GPT-4, 2025)* to support the development and presentation of this work in the following ways:

- Improving academic writing quality by refining grammar, clarity, and formal tone across all chapters.
- Helping format and generate LaTeX-compatible figures.
- Assisting with Python-based term extraction workflows using the spaCy model, and offering guidance with the use of RapidFuzz for fuzzy matching tasks, including restructuring scripts to reduce memory usage and improve reproducibility.
- Providing support on how to use concept relationships in SNOMED CT for extracting relevant categories of clinical terms.
- Offering guidance on the interpretation of ASR metrics such as M-WER and M-CER, including how to apply them meaningfully and report them consistently.

All AI-generated suggestions were critically reviewed and revised by me. The design of the experiments, interpretation of the results, and final conclusions reflect my own work and judgment. I take full responsibility for the content presented in this thesis.

Name: Shiran Sun

Date: 11.06.2025