MediSpeech

# D3.1. Clinical requirements and baseline systems

| | |
|---|---|
| Work package | WP3:AI-powered Automated Speech Processing Components |
| Type | Document |
| Contractual date of deliverable | 30 September 2026 |
| Actual date of deliverable | 19/08/2026 |
| Dissemination level | Public |
| Partner responsible | Radboud University |
| Status | Finalized |

| Document history | | | |
|---|---|---|---|
| **Version** | **Date** | **Author** | **Description** |
| 1.0 | November 2025 | Henk van den Heuvel . Mariia Zamyrova, Khiet Truong (RU) | Baseline system Dutch use case, speech related |
| 1.0 | November 2025 | Maxim Popov (AUMC) | Baseline system Dutch use case, NER related |
| 1.0 | June 2026 | Mehmet Kaan İldiz | Baseline system Turkish use case |
| 1.0 | July 2026 | Erica Yang | Baseline system UK use case |
| 1.0 | July 2026 | Filipe Bernardi | Baseline system Portuguese use case |

| Document review | | |
|---|---|---|
| **Date** | **Reviewer** | **Description** |
| 18 August 2026 | Henk van den Heuvel | Review |
| 18 August 2026 | Joanna Morozowska | Last polishing and review |

| Partners involved | | |
|---|---|---|
| **Company** | **Name and surname** | **Email address** |
| Radboud University | Henk van den Heuvel | Henk.vandenheuvel@ru.nl |
| University of Twente | Khiet Truong | k.p.truong@utwente.nl |
| Amsterdam UMC | Maxim Popov | m.popov@amsterdamumc.nl |
| Healthtalk | Joanna Morozowska | joanna@medrecord.io |
| SESTEK | Beyza Nur Hıdır | beyzanur.hidir@sestek.com |
| Turkcell | Emin Seckin | emin.seckin@turkcell.com.tr |
| NP Istanbul Brain Hospital | Mehmet Kaan İldiz | mehmetkaan.ildiz@uskudar.edu.tr |
| Chilton Computing | Erica Yang | erica.yang@chiltoncomputing.co.uk |
| University Porto | Filipe Bernardi | fbernardi@med.up.pt |
| BHI | Tim Dong | Qd18830@bristol.ac.uk |

# Table of Contents

# 1. EXECUTIVE SUMMARY

## 1.1. Purpose, scope and key outcomes

In this deliverable we define the clinical requirements for the speech technology components following from the use case definitions. Next, we define the baseline systems for the automated speech processing components involved in MediSpeech. The outcomes of this deliverable are relevant for the SoTA analysis carried out in D2.1 and as the first milestone for advancing these technologies in the project.

## 1.2 Abbreviations

| ABBREVIATION | MEANING |
|---|---|
| ASR | Automatic Speech Recognition |
| SD | Speaker Diarization |
| WER | Word Error Rate |
| CER | Character Error Rate |
| LM | Language Model |
| DER | Diarization Error Rate |
| JER | Jaccard Error Rate |
| DP | Diarization Purity |
| DC | Diarization Coverage |
| | |

# 2. INTRODUCTION

## 2.1 Background and context of deliverable

As a firm basis for the MediSpeech we need an overview of clinical requirements for the use cases of the various countries and partners involved. Also an overview of the baseline systems with which we start the project and which we will use and improve to achieve the project objectives are needed. This deliverable offers both for the contributing countries and use cases: The Netherlands, Turkey, UK and Portugal. For the Dutch use case the baseline system addresses ASR, acoustic event detection, speaker diarization, emotion recognition, and NER of medical terms. For the Turkish, use case is based on the existing manual clinical documentation workflow used during patient–doctor and clinician–caregiver consultations. Components addressed are for transcription, ASR, clinical NLP, and report generation. The UK use case deals with the extraction of semi-structured data from Echocardiography imaging systems currently by a non-AI-based NLP based free text extraction tool for converting Echocardiography medical notes into structured text. The Portuguese use case focuses on primary-care-related clinical documentation which includes the existing EHR and clinical information system environment, This baseline provides structured and semi-structured clinical data, clinical notes, coded information and integration points, without the envisaged end-to-end automated pipeline for real-time Portuguese clinical ASR, automatic speaker diarisation, clinical named entity recognition, concept mapping, summarisation and draft report generation from patient–doctor conversations.

## 2.2 Relation to the project objectives and work package

This deliverable is one of the ground stones of the project describing the basic needs and building blocks for the project. As such it is closely related to the deliverables of WP2:

| | |
|---|---|
| D2.1 | State of the art analysis |
| D2.2 | Use cases innovation description and high level requirements specification |
| D2.3 | Platform architecture design |

It this sets the basis for the data collection methods (D3.2) and protocols (D4.1) and the technical work of WP3 and WP4 which is then integrated in WP5.

# 3. CLINICAL REQUIREMENTS

Medical scribes are healthcare professionals who support the clinicians mostly with real-time documentation of the clinician-patient encounter by making notes that can be entered into the EHR. A digital scribe uses ASR, NLP and AI techniques to automate clinical documentation task that is currently solely conducted by humans. The tasks of a digital scribe are 1) to record the clinician-patient conversation, 2) to convert the audio to text, and 3) to extract and summarize important information from the text (Quiroz et al., 2019). As a whole, this system has specific requirements. However, in this section, we focus on requirements for ASR that aims to transcribe clinician-patient interactions. Clinician-patient interactions have specific characteristics that ASR models have to deal with and that may be challenging. Based on literature, we have identified several requirements for ASR as part of a digital scribe that aims to support the clinician in documenting clinician-patient interactions.

*ASR should be able to deal with variable and sub-optimal recording conditions* While clinician-patient encounters often take place in a relatively quiet room (e.g., the clinician's office), there is always the chance of high ambient noise in practice. Moreover, different clinicians use different microphones of varying quality. Speaker volume is also variable due to distance to the microphone and relative positioning (Quiroz et al., 2019).

*Recording of the audio should start in the clinician's waiting room* Some articles indicated that the clinician-patient encounter does not start in the clinician's office, but rather starts in the clinician's waiting room from the moment that the patient is being called into the clinician's office. During the walk from the waiting room to the clinician's office, the clinician is already informally chatting to the patient to build and preserve good relationship. This "medically irrelevant" information helps clinicians give context to previous and current encounters (Koopman et al., 2015).

*ASR should be able to deal with multiple speakers in a conversation* Often, patients bring their family members to the clinician's visit, sometimes out of necessity. Hence, in addition to ASR, we need speaker diarization technology that can identify who speaks when (Quiroz et al., 2019; Tran et al, 2023; Falcetta et al., 2023).

*ASR should be able to correctly transcribe conversational speech* Current ASR models still have some difficulties with transcribing spontaneous, conversational speech as they are usually trained on massive amounts of read speech. Several studies have illustrated that typical conversational phenomena such as disfluencies (e.g., false starts, interruptions, filled pauses) and other short, reduced lexical items (e.g., I, I'll) are erroneously transcribed (Quiroz et al., 2019; Tran et al., 2023).

*ASR should be able to correctly transcribe medical terminology* An additional complexity is the medical jargon used in clinician-patient interactions. ASR technology not explicitly trained on clinical conversations usually has difficulties recognizing medical terms since they are not part of our daily vocabulary (Quiroz et al., 2019; Tran et al., 2023; Van Buchem et al., 2021; Falcetta et al., 2023).

*ASR should be able to correctly transcribe accented, non-native speech* Since health issues occur in all layers of society, ASR should be robust against all types of speaker variation. For example, especially in multinational healthcare systems or regions with a high percentage of non-native speakers, their speech deviates from "the norm" and is hence challenging for ASR that is trained for "the norm" (Ng et al., 2025; Tran et al., 2023).

*Near real-time error correction should be possible* Give that ASR is not working perfectly yet, there might be some revision necessary post-interaction. To avoid lengthy revision processes, robust (near) real-time error correction should be made possible (Ng et al., 2025).

*Other metrics than WER should be used in the evaluation of the ASR* Many articles agree that more research is needed into what the clinicians' acceptance and satisfaction of the ASR in practice is, as well as what the clinical validity and usability of the ASR is (Ng et al., 2025; Evans et al., 2025; Van Buchem et al., 2021; Falcetta et al., 2023).

## REFERENCES

Evans, K., Papinniemi, A., Ploderer, B., Nicholson, V., Hindhaugh, T., Vuvan, V., Cowley, N., Tariq, A., & Thomson, H. (2025). Impact of using an AI scribe on clinical documentation and clinician-patient interactions in allied health private practice: Perspectives of clinicians and patients. Musculoskeletal Science and Practice, 78, 103333. https://doi.org/10.1016/j.msksp.2025.103333

Falcetta, F. S., De Almeida, F. K., Lemos, J. C. S., Goldim, J. R., & Da Costa, C. A. (2023). Automatic documentation of professional health interactions: A systematic review. Artificial Intelligence in Medicine, 137, 102487. https://doi.org/10.1016/j.artmed.2023.102487

Koopman, R. J., Steege, L. M. B., Moore, J. L., Clarke, M. A., Canfield, S. M., Kim, M. S., & Belden, J. L. (2015). Physician Information Needs and Electronic Health Records (EHRs): Time to Reengineer the Clinic Note. The Journal of the American Board of Family Medicine, 28(3), 316–323. https://doi.org/10.3122/jabfm.2015.03.140244

Quiroz, J. C., Laranjo, L., Kocaballi, A. B., Berkovsky, S., Rezazadegan, D., & Coiera, E. (2019). Challenges of developing a digital scribe to reduce clinical documentation burden. Npj Digital Medicine, 2(1), 114. https://doi.org/10.1038/s41746-019-0190-1

Tran, B. D., Mangu, R., Tai-Seale, M., Lafata, J. E., & Zheng, K. (2023). Automatic speech recognition performance for digital scribes: A performance comparison between general-purpose and specialized models tuned for patient-clinician conversations. AMIA Annual Symposium Proceedings, 2022, 1072.

Van Buchem, M. M., Boosman, H., Bauer, M. P., Kant, I. M. J., Cammel, S. A., & Steyerberg, E. W. (2021a). The digital scribe in clinical practice: A scoping review and research agenda. Npj Digital Medicine, 4(1), 57. https://doi.org/10.1038/s41746-021-00432-5

Below follows a listing of more general requirements for clinical applications of the speech technology and NLP components following from the use case definitions. They also result from the state of the art analysis in D2.1 and from documents such as from the European Commission (2025).

## 3.1 Functional

- Accurate transcription and contextualization of patient-doctor conversations
- Integration with Electronic Health Records (EHRs) for seamless upload, review, and signature
- Adaptability across medical specialties, including complex internal medicine and critical care
- Support for multilingual environments, especially for deployment across EU Member States
- Customization to individual clinician preferences, even within the same specialty
- Template flexibility to accommodate personalized documentation styles
- Terminology precision using accurate medical language rather than layman terms
- Real-time customization by clinicians during clinical workflows
- Support for clinician-editable draft documentation, ensuring that ASR/NLP/CDSS outputs remain assistive and subject to healthcare professional review
- Ability to distinguish between verbatim transcript, clinically relevant extracted information, and final clinical documentation
- Traceability between generated documentation items and the corresponding source transcript segments

## 3.2 Technical

- High-performance ASR models capable of handling ambient noise and multi-speaker environments
- High-performance NLP models capable of extracting medical terms following standard ontologies from spoken dialogues
- Scalability and flexibility to accommodate different documentation workflows
- Standardized data structures for interoperability and bias mitigation
- Support clinical speech, including spontaneous speech, clinical abbreviations, medication names, measurements, accents, interruptions and multi-speaker interaction
- Ability to process consultations involving clinician, patient and, where applicable, caregiver or family member
- Support for mapping extracted concepts to clinical terminologies and coding systems, where applicable
- Separation between raw audio, transcript, extracted entities, structured fields and final clinician-approved documentation
- Support for evaluation datasets with reference transcripts, speaker labels and clinician-annotated clinical entities

## 3.3 Legal & Regulatory

- GDPR compliant protocols (verbal or written) for ambient recording and data processing
- Compliance with AI Act, the Product Liability Directive (PLD), the Medical Device Regulation (MDR), the Health Technology Assessment Regulation (HTAR) and the European Health Data Space (EHDS)
- Targeted patient communication to minimize unnecessary concern
- Data privacy and security compliance (cloud storage, cybersecurity certifications)
- Assurance labs to validate AI tools before deployment

## 3.4 Training

- Training programs for clinicians to ensure effective use and trust
- Education on capabilities and limitations to manage expectations

## 3.5 Deployment & Evaluation

- Seamless EHR integration to reduce cognitive load leading to reduction in administration time and effort and to improvement of subjective quality of patient doctor contact
- Infrastructure readiness (hardware availability, digital systems)
- Pilot testing across specialties to ensure adaptability
- Post-deployment monitoring to assess (KPIs, usability, workflow impact)
- Health economics evaluation (ROI analysis, economic impact)
- Multidisciplinary collaboration (IT, data science, clinical teams)
- Budget allocation and selection criteria for equitable deployment
- Workflow integration support from IT personnel and innovation officers
- Scalable rollout strategy with feedback loops
- Evaluation should include WER/CER, diarisation metrics, entity-level precision/recall/F1-score, concept-mapping accuracy, report completeness, clinician correction effort, documentation time, usability, workflow fit and safety-critical error analysis
- WER should not be used as the only evaluation metric, because clinical impact depends on the type, meaning and context of errors
- Evaluation should distinguish between technical performance and clinical acceptability.
- AI-generated outputs should not be automatically incorporated into the EHR without healthcare professional review and approval

# 4. BASELINE SYSTEMS FOR PATIENT DOCTOR CONVERSATIONS, THE NETHERLANDS

## 4.1 Automatic Speech Recognition

In identifying the baseline for the ASR component we evaluated the performance of several models on Dutch medical conversations. We used a corpus of 11 recordings collected as part of the HoMed project (van den Heuvel & Pieters, 2024). The recordings capture 6-13 minute-long spontaneous monologues and dialogues with background noise and music. Tejedor-García et al. (2024) already measured the word error rate (WER) of Whisper (large-v2), wav2vec 2.0 and Kaldi_NL on the same dataset, finding that Whisper performed the best (10.9%) followed by wav2vec 2.0 (24.2%) and Kaldi_NL (28.4%).

In our experiments we evaluated XLSR-53-Dutch (with and without a language model (LM)), Whisper large-v2 and NVIDIA's NeMo NL FastConformer-CTC. We compared the models' performance to HealthTalk's medical reporting tool ("HealthTalk", n.d.). We found that Whisper again performed best, very close to the HealthTalk system. Therefore, we select Whisper as our baseline ASR model. The active development of new Whisper-based models that improve its latency, word-timestamp accuracy and ability to recognized speech disfluency (as mentioned in D2.1) also gives an optimistic outlook on what we can achieve using Whisper as our core ASR.

| | WER (%) | CER (%) |
|---|---|---|
| Whisper (large-v2) | **20.6** | **11.4** |
| XLSR-53 (no LM) | 40.1 | 18.0 |
| XLSR-53 (LM) | 35.4 | 17.2 |
| NeMo FastConformer | 61.3 | 40.9 |
| HealthTalk | 19.7 | 10.8 |

**Table 1: Results of evaluating Whisper, XLSR-53, NeMo FastConformer and HealthTalk's tool on the HoMed collection of recordings. For each system we report the average WER and CER. The best WER and CER compared to HealthTalk's values are highlighted in bold.**

## REFERENCES

HealthTalk. (n.d.). *HealthTalk*. https://healthtalk.ai/

Van den Heuvel, H., Pieters, A.H.L.M. (2024). Homed Medical Conversations Test Materials. Version 1. Radboud University. (dataset). https://doi.org/10.34973/vtaw-fm19

Tejedor-García, C. & van den Heuvel, H. & Hessen, A. & Dulmen, S. & Pieters, T. (2024). Comparative Analysis of State-of-the-Art Automatic Speech Recognition Systems for Transcribing Medical Consultation Audio Recordings.

## 4.2 Paralinguistic event detection

For emotion recognition, given the available emotional speech datasets, it is rather doubtful whether the existing and available baseline systems fit the context of physician-patient conversations. As described in D2.1, many of the available emotional speech databases contain basic emotions (neutral, anger, fear, joy, boredom, sadness), and are acted or scripted. It is also known that emotions are highly context-dependent; emotions expressed during physician-patient encounters are likely different from the acted, basis emotions. The question is how meaningful baseline systems trained on acted, basic emotions can be in the context of physician-patient encounters. Nevertheless, we will mention several potential state-of-the-art models to be used, some concern emotion categories, and concern dimensions (which might be a bit more promising):

- https://github.com/HappyColor/Vesper (Chen et al., 2024): this is a pre-trained model, trained on LSSED.
- https://github.com/audeering/w2v2-how-to (Wagner et al., 2023): this is a wav2vec2 model fine-tuned on the dimensions of arousal, valence, and dominance of the MSP-Podcast dataset.
- https://huggingface.co/speechbrain/emotion-recognition-wav2vec2-IEMOCAP: a wav2vec2 model by SpeechBrain, fine-tuned on the IEMOCAP which achieved an averaged accuracy of 75.3% on the IEMOCAP test set.
- https://huggingface.co/speechbrain/emotion-diarization-wavlm-large: another model by SpeechBrain that uses a WavLM model fine-tuned on the Zaion Emotion Dataset (Wang et al., 2023), and that focuses on diarizing emotion (happy, sad, angry, neutral) in speech.
- https://huggingface.co/Dpngtm/wav2vec2-emotion-recognition: a wav2vec2 model fine-tuned on ~12,000 audiofiles from TESS, CREMA-D, SAVEE, and RAVDESS for a 7-class emotion classification task.

With respect to enhancing transcripts with nonverbal vocalisations, such as filler events, we have identified CrisperWhisper (Zusag et al., 2024) as a potential baseline system. CrisperWhisper is able to produce more verbatim speech transcriptions with word-level timestamped segmentation and aims to be robust against multiple speakers and background noise.

https://github.com/nyrahealth/CrisperWhisper

With respect to laughter detection, a potential baseline is the laughter detector developed by Gillick et al. (2021). However, this laughter detector is not integrated in the ASR pipeline and hence, some work needs to be carried out to integrate the detected laughter events into the ASR transcripts. The work by Ho et al. (2025) does integrate laughter detection into ASR transcription. However, their model is not available (yet).

https://github.com/jrgillick/laughter-detection

### REFERENCES

Chen, W., Xing, X., Chen, P., & Xu, X. (2024). Vesper: A Compact and Effective Pretrained Model for Speech Emotion Recognition. IEEE Transactions on Affective Computing, 15(3), 1711–1724. https://doi.org/10.1109/TAFFC.2024.3369726

Gillick, J., Deng, W., Ryokai, K., & Bamman, D. (2021). Robust Laughter Detection in Noisy Environments. Interspeech 2021, 2481–2485. https://doi.org/10.21437/Interspeech.2021-353

Ho, P. H., Bălan, D. A., Heylen, D. K. J., & Truong, K. P. (2025). Enhancing Transcripts of Open-Source Automatic Speech Recognition Models Through Fine-Tuning with Laughter and Speech-Laugh. Interspeech 2025, 4513–4517. https://doi.org/10.21437/Interspeech.2025-2193

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. W. (2023). Dawn of the transformer era in speech emotion recognition: Closing the valence gap. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(9), 10745–10759. https://doi.org/10.1109/TPAMI.2023.3263585

Wang, Y., Ravanelli, M., & Yacoubi, A. (2023, December). Speech emotion diarization: Which emotion appears when?. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 1-7).

Zusag, M., Wagner, L., & Thallinger, B. (2024). CrisperWhisper: Accurate Timestamps on Verbatim Speech Transcriptions. Interspeech 2024, 1265–1269. https://doi.org/10.21437/Interspeech.2024-731

## 4.3 Speaker diarization

Baseline for speaker diarization in Dutch (Radboud University) is the Pyannote package using the following models:

- https://huggingface.co/pyannote/speaker-diarization-3.1
- https://huggingface.co/pyannote/segmentation-3.0

To get some baseline estimates of its performance we evaluated it on a subset of 18 recordings of 2-3 speaker spontaneous conversations from the Corpus Gesproken Nederlands (CGN) dataset. In addition to diarization error rate (DER), for a detailed overview, we used Jaccard error rate (JER), diarization purity (DP) and diarization coverage (DC). JER is similar to DER but averages the error

rate over individual speakers. DP measures the overlap of the reference and predicted speech segments, such that ideally only one reference speaker makes up the majority of the predicted speech segment (high purity). DC also measures overlap but emphasises that ideally a speaker's reference segment should be covered by a single (large) predicted segment rather than several shorter ones. For both DP and DC the larger the overlapping regions the higher the metric and the better the segmentation results. We computed the measures with different collars which indicate the absolute allowed deviation of the predicted timestamps from the reference. For example, a collar of 250ms implies that the timestamp can deviate from the reference by ±250 ms (500ms error window).

| | DER % ↓ | JER % ↓ | DP % ↑ | DC % ↑ |
|---|---|---|---|---|
| Collar = 0 ms | 18.9 | 26.9 | 76 | 78.8 |
| Collar = 250 ms | 12.8 | 21.4 | 79.2 | 82.4 |

**Table 2: Performance of the Pyannote diarization pipeline on a set of Dutch CGN 2-3 speaker recordings.**

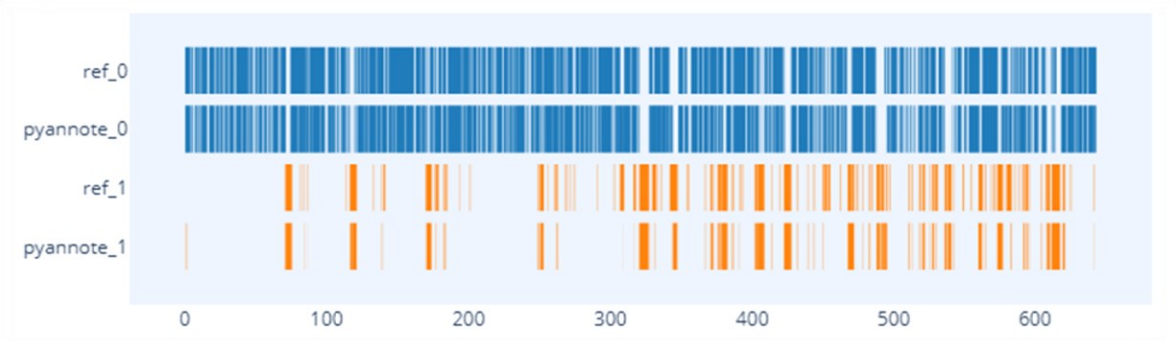


**Table 3: An example segmentation generated by Pyannote for a 2-speaker (labeled as speakers 0 and 1) recording, compared to the reference segmentation.**

Moving forward we aim to explore and compare new diarization solutions like the DiariZen model and diarizers from NVIDIA's NeMo library. NeMo's models are already being used by, for example, HealthTalk. As part of the onboarding process in their platform, HealthTalk asks doctors to record a short audio sample intended to support speaker diarization, so that their voices can be more easily distinguished. During a conversation, Whisper transcribes what is being said, while a separate diarization system (an NVIDIA model in Healthtalk's NeMo Framework) uses that voice profile to figure out who is speaking (the doctor or the patient). The platform then stitches those two pieces of information together to create the final transcript that shows who said what.

## REFERENCES

Han, J., et al. (2025). Leveraging self-supervised learning for speaker diarization.

NVIDIA NeMo Framerwork User Guide: speaker diarization models. (n.d.) https://docs.nvidia.com/nemo-framework/user-guide/latest/nemotoolkit/asr/speaker_diarization/models.html

## 4.4 Named Entity Recognition of Medical Terms

Named Entity Recognition task in Medical Domain can be generally split into 2 tasks – Biomedical Entity Linking, that can be used for integration of free-text clinical notes into the EHR systems, and Entity Recognition for de-identification.

To test the sensitive information identification capabilities of current methods, a group from Utrecht and Leiden universities has extracted a collection of files in JSON Lines text format from the UMC Utrecht EHR system. These were then used to evaluate 2 common approaches – Deduce, a rule-based approach, which is currently used in the Utrecht UMC, and Deidentify, a pre-trained Bidirectional Long ShortTerm Memory–Conditional Random Field (BiLSTM-CRF) model. The results of this evaluation can be found in the table below.

| PHI Category | Deduce P | Deduce R | Deduce $F_1$ | Deidentify P | Deidentify R | Deidentify $F_1$ |
|---|---|---|---|---|---|---|
| DATUM (date) | 0.94 | 0.8 | 0.86 | 0.97 | 0.86 | 0.91 |
| INSTELLING (institution) | 1 | 0.38 | 0.55 | 0.93 | 0.33 | 0.49 |
| LEEFTIJD (age) | 1 | 0.26 | 0.41 | 0.96 | 0.27 | 0.42 |
| LOCATIE (location) | 0.93 | 0.4 | 0.56 | 0.82 | 0.63 | 0.71 |
| PATIENTNUMMER (patient number) | 0.03 | 0.03 | 0.03 | 0.37 | 0.8 | 0.5 |
| PERSOON (person) | 0.98 | 0.79 | 0.87 | 0.2 | 0.77 | 0.32 |
| TELEFOONNUMMER (telephone number) | 0.96 | 0.8 | 0.87 | 0.83 | 0.91 | 0.87 |
| URL | 0.81 | 0.93 | 0.87 | 0.59 | 0.93 | 0.72 |
| macro avg | 0.74 | 0.6 | 0.56 | 0.67 | 0.65 | 0.49 |
| weighted avg | 0.95 | 0.66 | 0.76 | 0.87 | 0.46 | 0.46 |

As evident from this evaluation, currently deployed approaches are sub-optimal when it comes to the deidentification of sensitive patient information, with certain categories like “patient number” being identified by Deduce with a sensitivity of just 0.03, with an overall F1 score of 0.76, while a Neural Network-based approach only reached 0.46. This raises a concern about the performance of state-of-the-art approaches in current medical scenarios, and urges to develop new, more stable approaches to solve the problem of patient data safety.

One of the baseline models, trained for Dutch Medical NLP is https://huggingface.co/CLTL/MedRoBERTa.nl. It was trained on nearly 10 million anonymized hospital notes from the Amsterdam University Medical Centre.

As for the datasets it can be tuned on to accomplish a task of biomedical entity linking, the biggest and the most diverse ontology we can use is Dutch Medical Ontology, derived from a Dutch portion of UMLS (Unified Medical Language System, which was created to unify all of the existing ontologies and assign unique ids to their concepts) plus Dutch SNOMED vocabulary (that was not included in the UMLS) and common English drug names, resulting in 752,536 terms linked to over 366,000 unique concepts.

The datasets of medical texts that rely (fully or partially) on this ontology are WALVIS and Mantra GSC. WALVIS Corpus is a weakly labeled Dutch biomedical entity linking dataset that was automatically generated by combining structured knowledge from Wikidata with inter-article hyperlinks on Dutch Wikipedia. It contains 51,693 sentences and 53,781 mentions. While its label quality was deemed suboptimal for direct evaluation (manual sampling found 28% of mentions linked to related but not identical concepts), it proved effective for fine-tuning BioEL models.

Mantra GSC is a multilingual gold-standard corpus for biomedical concept recognition, including a Dutch subset. The texts are sourced from Medline abstract titles and drug labels. For Dutch, it includes 200 units (100 Medline titles, 100 drug labels), with 677 total annotations and 490 unique concepts (CUIs).

| Group | Example | Ontology Count | Ontology Perc. | WALVIS* Tra. Count | WALVIS* Tra. Perc. | WALVIS* Val. Count | WALVIS* Val. Perc. | Mantra* Count | Mantra* Perc. |
|---|---|---|---|---|---|---|---|---|---|
| DISO | MS (MULTIPLE SCLEROSIS) | 310057 | 41.3 | 957 | 49.8 | 224 | 46.7 | 149 | 39.3 |
| CHEM | Neupro | 185096 | 24.6 | 402 | 20.9 | 108 | 22.5 | 66 | 17.4 |
| PROC | Dialyse (DIALYSIS) | 124345 | 16.6 | 90 | 4.7 | 20 | 4.2 | 68 | 17.9 |
| ANAT | Heup (HIP) | 108622 | 14.5 | 391 | 20.4 | 105 | 21.9 | 33 | 17.4 |
| LIVB | Patiënt (PATIENT) | 7586 | 1 | 14 | 0.7 | 6 | 1.2 | 29 | 7.7 |
| PHEN | Licht (LIGHT) | 5997 | 0.8 | 4 | 0.2 | 1 | 0.2 | 7 | 1.8 |
| DEVI | IUD's | 3153 | 0.4 | 3 | 0.2 | 0 | 0 | 5 | 1.3 |
| PHYS | Groei (GROWTH) | 3125 | 0.4 | 33 | 1.7 | 11 | 2.3 | 19 | 5 |
| ACTI | Macht (POWER) | 1053 | 0.1 | 0 | 0 | 0 | 0 | 1 | 0.3 |
| OBJC | Stof (FABRIC) | 678 | 0.1 | 13 | 0.7 | 3 | 0.6 | 2 | 0.5 |
| GENE | Codon | 497 | 0.1 | 3 | 0.2 | 2 | 0.4 | 0 | 0 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| OCCU | Genomics | 464 | 0.1 | 10 | 0.5 | 0 | 0 | 0 | 0 |
| CONC | Retentie (RETENTION) | 304 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Oth. | | 1559 | 0.2 | 0 | 0 | 0 | 0 | 0 | 0 |
| TOTAL | | 752536 | | 1920 | | 480 | | 379 | |

The best model achieved **54.7% accuracy** for exact concept matches and **69.8% 1-distance accuracy** (meaning the prediction was a very closely related concept, like a parent or child term in the ontology).

| Group | # | Accuracy BM | Accuracy 3S10FT | 1-Dist Acc. BM | 1-Dist Acc. 3S10FT |
|---|---|---|---|---|---|
| DISO | 149 | 49.3 | 59.6 | 63 | 77 |
| PROC | 68 | 29.7 | 39.5 | 41.5 | 56.1 |
| CHEM | 66 | 48.2 | 57.6 | 58.5 | 67.3 |
| ANAT | 33 | 57.6 | 66.7 | 66.7 | 78.2 |
| LIVB | 29 | 33.8 | 48.3 | 48.3 | 61.4 |
| PHYS | 19 | 56.8 | 58.9 | 66.3 | 71.6 |
| PHEN | 7 | 57.1 | 76.2 | 71.2 | 82.4 |
| DEVI | 5 | 20 | 20 | 20 | 28 |
| OBJC | 2 | 0 | 0 | 50 | 70 |
| ACTI | 1 | 0 | 20 | 0 | 100 |
| Total | 379 | 44.6 | 54.7 | 56.7 | 69.8 |

However, in this study, the quality of the Entity Recognition was not directly evaluated, as the authors only reported that when they manually evaluated a random sample of 100 extracted mentions, they found that 42 of them were errors from the NER step, and so the authors conclude that the named entity recognition step remains the "main source of error" in the overall process.

## REFERENCES

Kors, J. A., Clematide, S., Akhondi, S. A., van Mulligen, E. M., & Rebholz-Schuhmann, D. (2015). A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. *Journal of the American Medical Informatics Association*, *22*(5), 948–953.

University Medical Center Utrecht. (2022). *Dutch-medical-concepts* (Version 1.11.0) [Computer software]. GitHub. https://github.com/umcu/dutch-medical-concepts

University Medical Center Utrecht. (n.d.). *Dutch-medical-wikipedia* [Computer software]. GitHub. Retrieved September 8, 2025, from https://github.com/umcu/dutch-medical-wikipedia

Hartendorp, F., Seinen, T., van Mulligen, E., & Verberne, S. (2024). Biomedical entity linking for Dutch: Fine-tuning a self-alignment BERT model on an automatically generated Wikipedia corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 1656–1665). ELRA and ICCL.

Mosteiro, P., Wang, R., Scheepers, F., & Spruit, M. (2025). Investigating de-identification methodologies in Dutch medical texts: A replication study of deduce and deidentify. *Electronics*. (Citation based on anticipated publication).

University Medical Center Utrecht. (2024). *Clinlp: A Python library for performing NLP on clinical text written in Dutch* (Version 0.9.4) [Computer software]. GitHub. https://github.com/umcu/clinlp

# 5 BASELINE SYSTEMS FOR PATIENT DOCTOR CONVERSATIONS, TURKEY

## 5.1 Baseline system of use cases

In the Turkey use case, the baseline system is the existing manual clinical documentation workflow used during patient–doctor and clinician–caregiver consultations. Clinical information is collected through face-to-face consultation, clinical observation, anamnesis, caregiver or parent interviews, structured assessment forms, neuropsychological and behavioral tests, and existing hospital information system or electronic health record records. This baseline reflects routine clinical practice before the introduction of an end-to-end automated medical reporting pipeline.

The current workflow does not include a validated Turkish clinical automatic speech recognition system, automatic transcription of patient–doctor conversations, automatic clinical named entity recognition, automated summarization, or automatic draft report generation. Speech recognition has been widely studied for clinical documentation, but reported benefits and limitations depend on the clinical setting, documentation task, accuracy, workflow integration, and human review process (Blackley et al., 2019). Therefore, for the Turkey use case, the manual reporting workflow will be treated as the operational baseline.

Clinical documentation burden is a relevant baseline issue because EHR-related and documentation-related workload has been reported as a contributor to clinician burnout and time pressure (Budd, 2023). In the Turkey pilot, MediSpeech will address this baseline limitation by supporting speech-to-text conversion, Turkish clinical NLP, medical entity extraction, text summarisation, structured draft report generation, and clinician-controlled validation.

The starting point for MediSpeech in Turkey is the manual preparation of clinical reports after or during neuropsychiatric and child/adolescent psychiatry consultations. The clinician listens to the patient and/or caregiver, observes clinical behaviour, reviews previous records and test results, identifies clinically relevant information, and prepares the report manually.

The baseline process is human-led and clinician-controlled. It depends on the clinician's interpretation of consultation content, clinical observations, anamnesis, symptoms, behavioural indicators, diagnoses, assessment scale results, treatment recommendations, and follow-up requirements. NLP-based clinical entity extraction is not currently used as part of the routine

baseline, although named entity recognition is a recognised method for extracting structured clinical information from electronic health records and other clinical texts (Durango et al., 2023).

For MediSpeech, this manual baseline will be compared against an AI-supported workflow in which patient–doctor conversation audio and related clinical text sources are converted into structured draft documentation. Research on generating medical reports from patient–doctor conversations supports the technical relevance of using ASR-generated conversational transcripts and sequence-to-sequence summarisation approaches as a basis for automated medical reporting (Enarvi et al., 2020).

| Baseline system / tool | Current use in Turkey use case | Role in the baseline |
|---|---|---|
| Face-to-face clinical consultation | Primary source of clinical information | Captures patient complaints, caregiver input, clinical history, symptoms, and clinician observations. |
| Clinical observation | Used during psychiatric and neurodevelopmental assessment | Supports identification of attention, behaviour, communication, activity level, eye contact, social interaction, and functional indicators. |
| Anamnesis | Collected manually by the clinician | Provides developmental history, symptom onset, family observations, school-related information, previous diagnoses, and treatment history. |
| Structured assessment forms and clinical scales | Used where clinically required | Supports standardised assessment of ADHD, autism spectrum disorder, behavioural symptoms, and functional impairment. |
| Neuropsychological and behavioural tests | Used according to patient age and clinical indication | Provides additional evidence for diagnostic reasoning and reporting. |
| Existing HIS / EHR environment | Used for storing clinical records and reports | Serves as the baseline documentation and record management environment. |
| Manual report writing | Current reporting method | Produces the final clinical report without automatic speech-to-text, NLP extraction, or automatic summarisation. |
| Clinician review and approval | Used for all reports | Ensures clinical responsibility, correctness, and final validation. |

## 5.2 Baseline limitations

The baseline is clinically reliable because it remains fully controlled by the healthcare professional; however, it is time-consuming and difficult to benchmark computationally. The main limitation is that consultation content remains largely unstructured until the clinician manually converts it into a report. This limits reuse of clinical text for downstream analytics and makes report completeness dependent on available time, documentation style, and workload.

Another limitation is the absence of benchmarked Turkish clinical ASR and NLP performance for the Turkey pilot data. Since speech recognition accuracy and clinical documentation quality are affected by domain terminology, speaker variation, accents, background noise, and workflow context, the baseline must first be measured on representative pilot data (Blackley et al., 2019).

| Limitation | Description |
|---|---|
| Manual documentation burden | Clinicians spend time manually writing reports after or during consultations. |
| No automated conversation transcription | Patient–doctor conversations are not automatically transformed into clinical text. |
| No validated Turkish clinical ASR baseline | There is no validated baseline system for automatic speech recognition in Turkish neuropsychiatric patient–doctor conversations. |
| No automatic clinical entity extraction | Symptoms, diagnoses, treatments, test names, observations, and recommendations are extracted manually. |
| No automated report generation | Structured draft medical reports are not generated automatically from consultation content. |
| Limited benchmarking data | Baseline WER, entity extraction F1-score, summarisation quality, and report generation accuracy are not yet available for the Turkey use case. |
| Variability in reporting style | Report completeness and terminology may vary depending on clinician, clinical scenario, and documentation workload. |
| Limited reuse of unstructured text | Free-text notes are not systematically converted into structured data for downstream analysis. |

## 5.3 Benchmark performance and methodology

The current baseline performance is not available as a computational benchmark because the routine Turkey workflow is manual and does not generate automatic ASR, NLP, or summarisation outputs. Therefore, benchmark values such as Word Error Rate, entity extraction precision, recall, F1-score, automatic summary quality, report completeness, and clinician correction rate will be measured during the Turkey pilot.

The benchmark will compare the manual baseline with the MediSpeech-supported workflow. For transcription, ASR output will be compared with reference transcripts using Word Error Rate. For clinical NLP, extracted symptoms, diagnoses, tests, treatments, recommendations, and other medical entities will be compared with clinician-annotated reference data using precision, recall, and F1-score. For report generation, draft reports will be evaluated using completeness against predefined report sections and clinician review outcomes. These benchmark dimensions are consistent with the literature on clinical speech recognition, medical conversation summarisation, and clinical named entity recognition (Blackley et al., 2019; Enarvi et al., 2020; Durango et al., 2023).

| Benchmark dimension | Current baseline | Baseline value | MediSpeech target / expected improvement |
|---|---|---|---|
| Speech-to-text accuracy | No automated Turkish clinical ASR system currently used in routine baseline workflow | Not available / to be measured | WER target to be measured on Turkey pilot data; target below 15% WER where technically feasible. |
| Word Error Rate | Manual documentation; no ASR benchmark available | Not available / to be measured | Reference transcript comparison after ASR component deployment. |

| | | | |
|---|---|---|---|
| Clinical entity extraction | Manual identification by clinician | Not available / to be measured | Entity-level precision, recall, and F1-score to be measured. |
| Symptom extraction | Manual identification by clinician | Not available / to be measured | Improved structured capture of symptoms and behavioural indicators. |
| Diagnosis-related information extraction | Manual identification by clinician | Not available / to be measured | Automatic extraction of diagnosis-related terms and diagnostic clues. |
| Test and scale extraction | Manual identification from clinical records and reports | Not available / to be measured | Automatic detection of test names, scales, and assessment references. |
| Summarisation quality | Manual summarisation by clinician | Not available / to be measured | Clinician-rated quality of generated summaries. |
| Report completeness | Manual report writing | Not available / to be measured | Higher completeness against predefined report template items. |
| Documentation time | Manual preparation of clinical report | To be measured during baseline data collection | Reduction in report preparation time. |
| Clinician correction effort | Not applicable | To be measured after MediSpeech prototype use | Lower correction effort after model improvement. |
| Clinician acceptance | Not applicable | To be measured through pilot feedback | Positive usability and acceptance scores. |

The Turkey benchmark will be conducted in two stages. First, representative manual reporting scenarios will be documented to establish the baseline. Second, comparable scenarios will be processed with MediSpeech components. The comparison will include transcription quality, clinical entity extraction quality, report completeness, documentation time, correction effort, and clinician acceptance.

The final medical report will remain under the responsibility of the healthcare professional. MediSpeech will function as a clinical documentation support tool and will not replace medical judgement, diagnosis, or final report approval. Human review is required because clinical speech recognition and automatically generated documentation may contain errors that require professional validation (Blackley et al., 2019).

| **Metric** | **Definition** | **Baseline measurement approach** |
|---|---|---|
| WER | Word Error Rate of automatic transcription against reference transcript | Not applicable to manual baseline; to be measured once Turkish ASR output is available. |
| Precision | Proportion of extracted clinical entities that are correct | To be measured against clinician-annotated reference data. |
| Recall | Proportion of reference clinical entities correctly extracted | To be measured against clinician-annotated reference data. |
| F1-score | Harmonic mean of precision and recall | To be measured for symptom, diagnosis, test, treatment, and recommendation categories. |
| Report completeness | Proportion of required report sections correctly populated | To be assessed against predefined report templates. |
| Documentation time | Time required to prepare the report | Manual baseline time versus MediSpeech-supported workflow time. |

| | | |
|---|---|---|
| Correction rate | Amount of clinician correction required in generated draft report | To be measured during prototype validation. |
| Clinician acceptance | Clinician perception of usefulness, correctness, and workflow fit | To be measured through structured feedback forms. |

### REFERENCES

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. https://doi.org/10.1093/jamia/ocy179

Budd, J. (2023). Burnout related to electronic health record use in primary care. Journal of Primary Care & Community Health, 14, 21501319231166921. https://doi.org/10.1177/21501319231166921

Durango, M. C., Torres-Silva, E. A., & Orozco-Duque, A. (2023). Named entity recognition in electronic health records: A methodological review. Healthcare Informatics Research, 29(4), 286–300. https://doi.org/10.4258/hir.2023.29.4.286

Enarvi, S., Amoia, M., Del-Agua Teba, M., Delaney, B., Diehl, F., Hahn, S., Harris, K., McGrath, L., Pan, Y., Pinto, J., Rubini, L., Ruiz, M., Singh, G., Stemmer, F., Sun, W., Vozila, P., Lin, T., & Ramamurthy, R. (2020). Generating medical reports from patient–doctor conversations using sequence-to-sequence models. Proceedings of the First Workshop on Natural Language Processing for Medical Conversations, 22–30. https://doi.org/10.18653/v1/2020.nlpmc-1.4

# 6 BASELINE SYSTEMS FOR PATIENT DOCTOR CONVERSATIONS, UNITED KINGDOM

The British use cases involves the extraction of semi-structured data from Echocardiography imaging systems. Our use cases do not include requirements for ASR because English ASR is a mature field with substantial market offerings for medical domains, especially from a transcription accuracy standpoint. Some market leading AI-based AI scribe for medical domains vendors say they can reliably achieve up to an accuracy of 98.6%[1].

It is also important to note that this section only describes the baseline situation of the British use cases and possible evaluation methodology. The actual clinical decision support and reporting capabilities using advanced AI-based text analytics, including design, development, and implementation) will be undertaken in WP4 – Advanced text analytics using AI.

## 6.1 Baseline System of Use Cases

Currently, it is difficult for researchers and clinicians at the Bristol Heart Institute and University Hospitals Bristol, Weston NHS Foundation Trust (UHBW) and other NHS hospitals and to obtain imaging report related data in a structured format for research or risk flagging purposes as this is generally not available in the routinely collected datasets or national cardiac surgery audit database. Whilst we have access to data generated from collaborative efforts such as the Health Informatics Collaborative (HIC), the data collected multi-institutionally contains only a limited set of imaging data, specifically only the following echocardiography variables: DATE ECHOCARDIOGRAM, TIME ECHOCARDIOGRAM, LV EJECTION FRACTION, LVEDD, LVESD, LVFS and LV function. As can been seen in our previous publication,(https://doi.org/10.3390/bioengineering10111307) the number of variables for only Echocardiography extends far beyond the dimensionality of this HIC dataset. Whilst recent advances in Large language modelling (LLM) have made the use of ChatGPT common place in commercial applications, these are limited since clinically sensitive patient information that are protected by GDPR and other data protection regulations do not allow patient data to be uploaded into these publicly available AI systems. The existing available imaging datasets in the BHI include for example Echocardiography data, although with the pre-allocated funds available to Chilton for additional data collation, this may potentially allow the seeking of additional datasets such as Magnetic Resonance Imaging (MRI) datasets. However, another challenge is that currently, there are no available approaches that can extract free text imaging reports with clinically usable accuracy without additional clinician input. Opportunities to seek additional datasets from available resources (e.g. consortium partners) to this project shall also be considered. The output of the structured dataset generated from the historical data at BHI from imaging reports using this approach if not unsuccessful is expected to have considerable value to clinical research and patient risk flagging.

The data required are available in the (semi-structured) free text format in clinical medical notes recorded from for example, patient Echocardiography (echo) tests. In addition, risk factors in relation to structural deformities, conduction disturbances and blood flow abnormalities are present in these medical notes and may offer additional insight into clinical diagnosis that are currently difficult for researchers and clinicians to analyse in free text database format. We have collected a subset of echo medical notes at UHBW that are ready to be processed using AI approaches. Through the novel AI approaches, we aim to provide clinicians with high-dimensionally, accurate and reliable structural clinical indicators that could potentially be useful for clinical risk identification or research purposes such as for atrial fibrillation in the future. Although audit based registries could be considered for collating such structure information, these processes are not only expensive in terms of time and effort (i.e. requiring hand-typing/entry each individual measurement) but also do not enable capture of high-dimensionality of the clinical indicators available, since only a limited set of variables can be manually entered.

At the BHI, we have developed a non-AI-based NLP based free text extraction tool for converting Echocardiography medical notes into structured text. However, the ability for the tool to capture important clinical information is limited by the need for clinical expertise as input into the development of the tool and for annotating validation data to evaluate the tool. In addition, the tool is ill-suited to capturing structured information from different hospitals due to different local recording formats of medical notes. In this project, we will aim to assess the AI (e.g. Large language model) based solution using this non-AI-based NLP tool as benchmark. Through the novel AI-based solution, we envisage improved visualisation of medical notes through insight deriving analytics methods as well as performance and usability improvements.

Further details on the baseline system can be found in our BHI published paper: https://www.mdpi.com/2306-5354/10/11/1307

The current baseline system of the British use cases, provided by British Heart Institute, is an NLP system that do not employ AI. The current landscape is fragmented, unstandardized, and unsustainable. The lack of central management, evaluation, and comparability hinders the reliability and scalability of clinical research, creating a need for more robust, automated, and standardized AI-driven solutions.

Clinical academics use non-AI-based NLP extraction tools to extract clinical events and key information from the notes manually. The rules are handcrafted individually and there is no central system to systematically manage the rules and capture the outputs. There is also no systematic evaluation framework to evaluate the performance, such as consistency, reliability, and accuracy of the manual extraction. The rules used by the tools are often handcrafted by individual researchers as part of a larger clinical study to access the information from the notes to fulfil the needs of the specific study. Current tools struggle to scale across diverse studies, topics, and interests for different research projects. Significant variances also exist between the quality of the information extracted across different studies. Currently, there is no systematic comparison between different extraction tools used by different studies either.

It is also worth noting that some of these notes may come from clinician approved transcriptions and initial reports/letters produced by ASR and transcription software. Many British hospitals have now adopted ASR and AI scribe tools to assist clinicians to write reports. However, automated clinical text extraction for clinical decision support, robust clinical research and routine reporting remain immature.

## 6.2 Baseline limitations

The current non-AI-based NLP extraction tools are primarily rule-based, based on a combination of the following methods:

- Human crafted/hardcoded rules, dictionaries (white- or blacklists)
- Pattern matching, based on regular expressions assuming specific arrangements of dates, identifiers, and phone number arrangements
- Basic NLP processing to perform tokenization, identify part-of-speech, carry out language specific checks, such as deduplication, lemmatization, stammering

Rules and patterns are structure-based methods to extract information. Firstly, the rigidness of the approach determines the performance of the extraction. Any structural derivation e.g. misspelling, different sequences of wording, or even just additional whitespaces, will lead to a mismatch. This is common when a document is written by a different person and therefore, the same content may be presented structurally different. This can lead to a mismatch, thus reducing the accuracy of extraction. Secondly, it can also be subjective, subject to the text understanding capability of the individual who sets the rules, defines the word patterns, or data formats (e.g. date, mobile phone numbering style). Thus, the rules that work well in one hospital may not be readily transferrable to another. Finally, rules and patterns can get complicated to a point that maintaining them becomes a delicate task, making it difficult to validate their consistency and reliability. Especially, when it comes to negation, these methods perform poorly.

NLP based methods rely on grammar and language structures for extraction. Thus, the performance is subject to the grammatical structure of the language. Informal text which can appear in follow-up notes from post-surgery patient-doctor interactions. Such text is difficult to handle with NLP based methods due to its informality. The models used by popular NLP tools, such as spaCy,

will need additional tuning to work with specialized medical text to understand the specific terminologies, short forms, synonyms, and relationships.

Even after expert guided adaptations, existing NLP based methods can still struggle to capture the rich information from the clinical notes due to the lack of semantic, contextual and domain understanding of the text. These methods currently fall short of clinical research and routine reporting requirements.

## 6.3 Evaluation Metrics

The baseline performance evaluation are provided in the BHI paper. However, there are a range of benchmark evaluation metrics that we are closely examining in terms of their applicability to the British use cases, and they are provided in the table below.

| Metrics | Applicability | Description |
|---|---|---|
| Accuracy | AI model | Measure the total percentage of correct predictions (both positive and negative) |
| Precision and recall | AI model | Measure how many extracted entities and relationships are correct (medicine, doses, frequency, diagnosis, treatment, operation, lab tests etc) vs how many entities and relationships that the model misses |
| F1 score | AI model | Measure the overall robustness of the extraction, taking into account both precision and recalls |
| Scalability | AI model | Measure the throughputs of the AI model, in terms of requests/second (request = extraction request) |
| Speed | AI model | Measure how fast the AI model responds, in terms of seconds/request |
| Consistency (repeatability) | AI model | Measure how consistent (repeatable) the responses are given the same set of requests |
| Alignment correctness | AI model | Measure how many concepts (entities) extracted by the AI model are correctly mapped to a valid, and correctly grounded name entity in a given knowledge graph (e.g. SNOMED-CT)<br><br>Similar to entity/relation extraction, it is also possible to refine this metric to cover accuracy, precision, recall, F1, and other related metrics, but coming from a semantic angle. |
| Patient journey map accuracy* | Tool (system) | Measure how accurately the exaction tools can extract key events (e.g. medication, lab tests, diagnosis, treatments, side effects, adverse reactions, recovery status, mortality) to establish the chronological patient journey (admission, hospitalization, discharge, follow-up, recovery, rehabilitation, readmission etc) |
| Adaptability | Tool (system) | Measure how many different clinical studies the system can be configured to support, through providing patient journey related features/data and patient characteristics.<br><br>Example studies include structural deformities impact on patients' risk of post-surgery AF, or conduction disturbances impact on patients' risk of post-surgery AF, or pre-operative and post-operative factors. |

Similar to AI model evaluation metrics, it is well worth noting that the patient journey map metric has several other dimensions too, going beyond accuracy evaluation. When resources permitted, it can also include consistency, speed, scalability, precision/recall at a tool/system level.

[1] https://www.notta.ai/en/blog/medical-transcription-market

# 7 BASELINE SYSTEMS FOR PATIENT DOCTOR CONVERSATIONS, PORTUGAL

## 7.1 Baseline system of Portuguese use case

The Portuguese use case focuses on primary-care-related clinical documentation. In the current baseline workflow, clinical information is obtained through face-to-face patient–doctor interaction, review of previous clinical information, manual entry of findings into the electronic health record, and clinician-led coding or structuring of information where applicable.

The operational baseline for Portugal is therefore the existing manual or semi-structured clinical documentation workflow used in primary care and related clinical settings before the introduction of a MediSpeech-supported pipeline. In this baseline, patient–doctor conversations are not automatically recorded, transcribed, diarised, summarised or converted into structured draft reports. Clinical documentation remains under the direct responsibility of the healthcare professional, who listens to the patient, asks follow-up questions, interprets clinical information, records relevant findings, and produces the final clinical note or report.

The Portuguese baseline also includes the existing EHR and clinical information system environment, with MedicineOne as the relevant EHR and integration partner for the Portuguese use case. This baseline provides structured and semi-structured clinical data, clinical notes, coded information and integration points, but does not currently provide an end-to-end automated pipeline for real-time Portuguese clinical ASR, automatic speaker diarisation, clinical named entity recognition, concept mapping, summarisation and draft report generation from patient–doctor conversations.

For MediSpeech, the Portuguese starting point should therefore be understood as a combination of:

- the current patient–doctor consultation workflow in primary care;
- manual clinical documentation and clinician-controlled EHR entry;
- existing EHR structures and coding practices, including use of ICPC-2, ICD-10, LOINC and SNOMED CT where applicable;
- absence of a validated Portuguese clinical ASR baseline specific to primary-care patient–doctor conversations;
- absence of a validated Portuguese clinical NLP/NER baseline for extracting clinical concepts from ASR-generated consultation transcripts in this use case;
- clinical and methodological validation capacity from FMUP/University of Porto;
- EHR integration and clinical coding expertise from MedicineOne, alongside its role as a data-technology, AI/ML, platform and CDSS-development partner for the Portuguese use case;
- AI/ML, data-technology, platform and CDSS-development capacity from MTG, including AI validation and monitoring capacity;

- clinical/end-user contribution from UFP and other Portuguese clinical settings where applicable.

The baseline is clinically safe because it is fully human-led and clinician-controlled. However, it is time-consuming, difficult to benchmark computationally, and dependent on clinician documentation style, workload, consultation complexity and EHR usability. MediSpeech will compare this baseline with an AI-supported workflow in which patient–doctor conversation data and related clinical text sources are transformed into structured, clinician-reviewable draft documentation.

**Table 7.1. Baseline systems and workflows in the Portuguese use case**

| Baseline system/workflow | Current use in the PT use case | Role in the baseline |
|---|---|---|
| Face-to-face primary care consultation | Main source of clinical information | Captures symptoms, concerns, history, clinical reasoning, treatment plan and follow-up needs |
| Manual clinical documentation | Current documentation method | Clinician manually records relevant information in the EHR during or after the consultation |
| MedicineOne / EHR environment | Relevant Portuguese EHR/integration environment | Stores structured and semi-structured clinical data, clinical notes and coded information |
| Manual review of previous records | Used before or during consultation | Provides context on diagnoses, medication, laboratory results, previous encounters and follow-up plans |
| Manual Clinical coding and terminology mapping | Used where applicable | Supports the structuring of clinical information using ICPC-2, ICD-10, LOINC, and SNOMED CT, depending on the data type and clinical workflow |
| Manual report or note preparation | Current reporting process | Produces the final clinical documentation without automated ASR, diarisation, NER or summarisation. |
| Clinician review and approval | Required for all documentation | Ensures clinical responsibility, accuracy, error correction, and final validation |
| FMUP scientific and methodological input | Available for project validation | Supports requirements definition, annotation strategy, data quality assessment, clinical validation and evaluation methodology |
| MTG AI validation and monitoring input | Available for project validation | Supports monitoring and validation of AI components and outputs |

| | | |
|---|---|---|
| HE-UFP / Portuguese clinical site(s) | Clinical/end-user environment where applicable | Provides clinical workflow input, end-user perspective and potential pilot context under approved protocols |

## 7.2 Baseline limitations

The Portuguese baseline has the advantage of being clinically controlled and aligned with current professional responsibility. However, it presents several limitations that MediSpeech aims to address.

First, consultation information remains largely unstructured until the clinician manually transforms it into a clinical note, report or coded record. This reduces the possibility of reusing consultation content for quality improvement, clinical analytics, continuity of care, decision support or research.

Second, there is no currently validated European Portuguese clinical ASR baseline for spontaneous, multi-speaker primary-care conversations in the MediSpeech Portuguese use case. Existing ASR models for Portuguese, including recent European Portuguese benchmarks, are useful starting points, but they do not replace use-case-specific benchmarking on clinical conversations, medical terminology, background noise, multiple speakers, accents, abbreviations and EHR-oriented documentation.

Third, there is no validated baseline for extracting European Portuguese clinical entities directly from ASR-generated consultation transcripts in this use case. Clinical NLP and NER resources exist for Portuguese, particularly Brazilian Portuguese clinical corpora and domain-specific medical record datasets, but differences in language variety, clinical settings, terminology, coding practices, and documentation styles mean that the European Portuguese MediSpeech use case requires its own validation.

Fourth, the Portuguese use case involves heterogeneous primary-care domains, with different terminology, consultation structures, levels of patient/caregiver participation, clinical measurements, and different terminology, different consultation structures, different levels of patient/caregiver participation, different clinical measurements and different documentation requirements. The baseline must therefore be evaluated by ASR/NLP problem.

**Table 7.2. Main limitations of the Portuguese baseline**

| Limitation | Description |
|---|---|
| Manual documentation burden | Clinicians manually transform consultation content into EHR documentation, which may increase administrative workload |
| No automated PT clinical ASR in routine baseline | Patient–doctor conversations are not automatically transcribed in the current Portuguese baseline workflow |
| No validated PT primary-care ASR benchmark | Baseline WER/CER for spontaneous Portuguese clinical conversations in this use case is not yet available |

| | |
|---|---|
| No automatic speaker diarisation in the routine workflow | Speakers turn between the clinician, patient and caregiver/family member are not automatically identified |
| No automatic clinical entity extraction from conversations | Symptoms, diagnoses, measurements, medications, tests, procedures and follow-up plans are identified manually |
| No automatic draft report generation | Structured draft notes or reports are not generated automatically from the consultation transcript |
| Variable documentation style | Note structure, terminology, completeness and level of detail may vary between clinicians and consultation types |
| Limited reuse of free-text information | Unstructured notes are not systematically converted into reusable structured data |
| Need for clinical safety evaluation | ASR/NLP errors may affect clinically relevant information, including negation, medications, dosages, diagnoses, and follow-up instructions |
| Need for human review | Any MediSpeech-generated output must remain assistive and subject to clinician review before clinical use or EHR incorporation |

## 7.3 Benchmark performance and methodology

No computational baseline values are currently available for the Portuguese use case, because the routine baseline workflow is manual and does not generate ASR, diarisation, NER, summarisation or automatic report-generation outputs. Therefore, benchmark values must be established during the Portuguese pilot or validation phase.

The benchmark should compare the manual baseline with the MediSpeech-supported workflow. The evaluation should include both technical and clinical dimensions: transcription accuracy, speaker diarisation, clinical entity extraction, concept mapping, report completeness, documentation time, clinician correction effort, usability, workflow fit and clinical safety.

For ASR, automatic transcripts should be compared against reference transcripts using Word Error Rate (WER) and Character Error Rate (CER). Because clinical consequences depend not only on the number of errors but also on their meaning, the Portuguese evaluation should include error typology and safety-critical error analysis. Particular attention should be paid to errors involving negation, medication names, dosage, measurements, diagnoses, temporal expressions, pregnancy-related information, child/adolescent health information, diabetes and hypertension follow-up parameters.

For speaker diarisation, the system should identify speaker turns among the clinician, patient, and, where applicable, caregiver or family member. Metrics such as Diarisation Error Rate (DER), Jaccard Error Rate (JER), diarisation purity and diarisation coverage can be used where annotated reference data are available.

For clinical NLP/NER, extracted entities should be compared with clinician-annotated reference data using precision, recall, and F1 Score. Relevant entity groups for the Portuguese use case may include symptoms, diagnoses, clinical measurements, medication, laboratory parameters, procedures, risk factors, pregnancy follow-up items, child/adolescent health indicators, recommendations and follow-up plans. Mapping to ICPC-2, ICD-10, LOINC and SNOMED CT should be evaluated where applicable.

For report generation or structured draft documentation, the evaluation should include completeness, correctness, traceability to the source transcript, clinician correction effort, and ability to support the intended EHR workflow. Generated outputs should be treated as draft assistive documentation, not as autonomous clinical documentation.

**Table 7.3. Proposed benchmark dimensions for the PT use case**

| **Benchmark dimension** | **Current baseline** | **Baseline status / initial value** | **MediSpeech evaluation / expected improvement** |
|---|---|---|---|
| Speech-to-text accuracy | No automated Portuguese clinical ASR in the routine workflow | No routine automated transcription baseline exists. Manual documentation is the current reference | WER and CER were measured against reference transcripts from Portuguese pilot data |
| Speaker diarisation | No automatic diarisation in the routine workflow | Not applicable in the manual workflow. Speaker roles are interpreted by the clinician during the consultation | DER, JER, diarisation purity and diarisation coverage where annotated data are available |
| Clinical entity extraction | Manual identification by a clinician | Manual clinical interpretation is the baseline. No automatic entity extraction score exists | Precision, recall and F1-score for clinically relevant entity types |
| Concept mapping | Manual or semi-structured coding where applicable | Existing coding practices depend on the EHR workflow and clinical context. No automatic mapping benchmark exists | Accuracy of mapping to ICPC-2, ICD-10, LOINC and SNOMED CT, where applicable |
| De-identification / sensitive information detection | Manual governance and data protection procedures | No automated de-identification benchmark is currently defined for Portuguese pilot transcripts | Precision, recall and F1-score for identifying personal/sensitive data in transcripts |
| Report completeness | Manual note/report preparation | Current baseline is clinician-authored documentation; completeness varies by consultation type and documentation style | Completeness against predefined primary-care documentation items |

| | | | |
|---|---|---|---|
| Documentation time | Manual documentation during/after consultation | Current time burden should be established through baseline observation or clinician self-report | Reduction in documentation time or post-consultation administrative effort |
| Clinician correction effort | Not applicable to manual baseline | Not applicable before AI-generated drafts are introduced | Time and extent of edits required before accepting AI-generated draft outputs |
| Clinical safety errors | Manual clinician-controlled documentation | Baseline relies on clinician responsibility and review; no automated safety-error baseline exists | Reduction or control of safety-critical errors after human review |
| Clinician acceptance | Not applicable to manual baseline | Baseline perception can be captured as satisfaction/frustration with the current documentation workflow | Perceived usefulness, trust, usability and workflow integration |
| EHR integration readiness | Manual/semi-structured EHR entry | The existing MedicineOne/EHR workflow is the integration baseline | Ability to insert or transfer validated structured outputs into the EHR workflow |

The Portuguese benchmark should be conducted in stages. First, representative baseline documentation scenarios should be characterised in the current workflow. Second, a reference dataset should be prepared using approved data sources and annotation procedures. Third, MediSpeech outputs should be compared against the reference data and against the manual baseline. Fourth, clinician review should be used to assess whether outputs are accurate, complete, safe, editable and useful in the clinical workflow.

The final clinical note or report must remain under the responsibility of the healthcare professional. MediSpeech should function as a clinical documentation support tool. It should not replace clinical judgement, diagnosis, treatment decisions or final approval of information entered into the EHR.

## References

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. https://doi.org/10.1093/jamia/ocy179

Carvalho, C., Teixeira, F., Botelho, C., Pompili, A., Solera-Ureña, R., Paulo, S., Julião, M., Rolland, T., Mendonça, J., Pereira, D., Trancoso, I., & Abad, A. (2025). CAMÕES: A Comprehensive Automatic Speech Recognition Benchmark for European Portuguese. arXiv:2508.19721. https://doi.org/10.48550/arXiv.2508.19721

Da Silva, D. P., Fröhlich, W. R., De Mello, B. H., Vieira, R., & Rigo, S. J. (2023). Exploring named entity recognition and relation extraction for ontology and medical records integration. Informatics in Medicine Unlocked, 41, 101381. https://doi.org/10.1016/j.imu.2023.101381

Enarvi, S., Amoia, M., Del-Agua Teba, M., Delaney, B., Diehl, F., Hahn, S., Harris, K., McGrath, L., Pan, Y., Pinto, J., Rubini, L., Ruiz, M., Singh, G., Stemmer, F., Sun, W., Vozila, P., Lin, T., & Ramamurthy, R. (2020). Generating medical reports from patient–doctor conversations using sequence-to-sequence models. Proceedings of the First Workshop on Natural Language Processing for Medical Conversations, 22–30. https://doi.org/10.18653/v1/2020.nlpmc-1.4

Just, S. A., Elvevåg, B., Pandey, S., Nenchev, I., Bröcker, A.-L., Montag, C., & Morgan, S. E. (2025). Moving beyond word error rate to evaluate automatic speech recognition in clinical samples: Lessons from research into schizophrenia-spectrum disorders. Psychiatry Research, 352, 116690. https://doi.org/10.1016/j.psychres.2025.116690

Kodish-Wachs, J., Agassi, E., Kenny, P., Overhage, J. M., & Wilcox, A. B. (2018). A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech. AMIA Annual Symposium Proceedings, 2018, 683–689.

Oliveira, L. E. S., Peters, A. C., Silva, A. M. P., Gebeluca, C. P., Gumiel, Y. B., Cintho, L. M. M., Carvalho, D. R., Hasan, S. A., & Moro, C. M. C. (2020). SemClinBr: A multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. arXiv:2001.10071. https://doi.org/10.48550/arXiv.2001.10071

# 8 CONCLUSIONS

This deliverable establishes a fundamental foundation for the MediSpeech project by providing a comprehensive overview of the clinical requirements and baseline systems across the four participating use cases in the Netherlands, Turkey, the United Kingdom, and Portugal. Together, these analyses define the starting point from which the project will develop, validate, and evaluate its innovations.

The assessment of the existing clinical workflows, technical infrastructures, and functional capabilities highlights both the common challenges and the country-specific requirements that must be addressed to achieve the project objectives. While the baseline systems already provide varying levels of support for activities such as speech recognition, clinical documentation, natural language processing, information extraction, and report generation, significant opportunities remain for the development of more advanced, integrated, and AI-driven solutions. The identified gaps and requirements therefore serve as clear targets for the technical developments to be undertaken within MediSpeech.

The findings presented here directly inform the design of data collection methods in Deliverable D3.2 and the development of data collection and evaluation protocols in Deliverable D4.1. Furthermore, they guide the technical activities in WP3 and WP4, whose outcomes will ultimately be integrated and validated within WP5. In this way, the deliverable provides the strategic and technical basis upon which the MediSpeech project will build its future innovations and achieve its expected impact.