Clinical BERTScore: An Improved Measure of Automatic Speech Recognition Performance in Clinical Settings

Joel Shor*
Verily Life Sciences, USA
joelshor@verily.com

Ruyue Agnes Bi*
MIT, USA
ruyuebi@mit.edu

Subhashini Venugopalan
Google Research, USA

Steven Ibara
Verily Life Sciences, USA

Roman Goldenberg
Verily Life Sciences, Israel

Ehud Rivlin
Verily Life Sciences, Israel

Abstract

Automatic Speech Recognition (ASR) in medical contexts has the potential to save time, cut costs, increase report accuracy, and reduce physician burnout. However, the healthcare industry has been slower to adopt this technology, in part due to the importance of avoiding medically-relevant transcription mistakes. In this work, we present the Clinical BERTScore (CBERTScore), an ASR metric that penalizes clinically-relevant mistakes more than others. We collect a benchmark of 18 clinician preferences on 149 realistic medical sentences called the Clinician Transcript Preference benchmark (CTP) and make it publicly available[1] for the community to further develop clinically-aware ASR metrics. To our knowledge, this is the first public dataset of its kind. We demonstrate that our metric more closely aligns with clinician preferences on medical sentences as compared to other metrics (WER, BLUE, METEOR, etc), sometimes by wide margins.

1 Introduction

Clinicians in a number of disciplines work in an overburdened healthcare system that leads to difficult working environments and an epidemic of physician burnout (Dzau et al., 2018). AI-related technologies have the potential for improving efficiency on repetitive tasks, therefore increasing both patient throughput and decreasing physician burnout. For example, physicians in a number of disciplines spend as much time doing paperwork as with patients (Tai-Seale et al., 2017). However, the adoption of speech technology in the medical community has been slow (Latif et al., 2021), and there are a number of speech technologies that could improve efficiency.

Speech technology can be applied to a number of medical problems including transcribing patient-physician conversations (Shafran et al., 2020), helping dysarthric patients communicate (Shor et al., 2020), and diagnosing medical conditions from speech (Shor et al., 2022; Shor and Venugopalan, 2022; Peplinski et al., 2021; Venugopalan et al., 2021). In this work, we focus on the task of generating a report after a colonoscopy procedure.

One of many reasons for the lower adoption of time-saving speech transcription technologies is that the ASR systems often don't perform as well in real-world clinical settings as they do on evaluation benchmarks. The most common metric for measuring ASR performance, Word Error Rate (WER), has significant practical drawbacks (Wang et al., 2003; Morris et al., 2004; He et al., 2011). First, all mistakes are treated equally. In clinical settings, however, medical words are more important (e.g. "had complete resection" → "had complete **c-section**" is a worse mistake than → "**has** complete resection", but both have equal WER). Second, some mistakes affect the overall intelligibility more than others (e.g. "was no perforation" → "was no **puffer age**" vs "was **not any** perforation"). Although researchers have proposed alternatives to the WER, no metric combines medical domain knowledge with recent AI advances in language understanding.

In this work, we make the following contributions:

1. Generate a collection of realistic medical sentences and transcripts with plausible ASR errors and collect preferences from 18 clinicians on 149 sentences. We publicly released this dataset for reference and future studies. This is the first public dataset of its kind.
2. Present the Clinical BERTScore (CBERTScore) and demonstrate that it more closely matches clinician preferences on medical transcripts than other ASR metrics (WER, BLEU, METEOR, BERTScore).
3. Demonstrate that CBERTScore does not perform worse than other metrics on non-medical

*Authors contributed equally
[1] https://osf.io/tg492/

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 1–7
July 14, 2023 ©2023 Association for Computational Linguistics

transcripts.

2 Related work

There are a number of ways to evaluate transcript quality. The Word Error Rate (WER), is the simplest to compute and most common. It counts the number of insertions, deletions, and substitutions between two text strings, and normalizes by the length of the reference string. The Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) measures the amount of n-gram overlap between two text strings (where *n* is often 4). It captures the intuition that groups of words are important in addition to individual words. METEOR (Banerjee and Lavie, 2005) focuses on unigrams, but computes an explicit alignment between two strings and takes both precision and recall into consideration. While these techniques are cheap to compute, they primarily focus on character or string similarity, not semantic similarity.

Our work most closely follows the BERTScore (Zhang et al., 2019). This metric computes a neural word embedding for each word in the reference and candidate. Embeddings are matched using cosine distance instead of string similarity, and the final score takes precision and recall into account (see Fig.1). This method takes semantic similarity into account, but not that some words are more important to preserve in clinical contexts.

Structured graphs are one way to encode real-world knowledge in a machine-readable format. The Knowledge Graph (KG) (Singhal, 2012) is a publicly available structure that encodes medical knowledge. Previous work has used the medical subset of the KG to learn medical entity extraction (Shafran et al., 2020). We primarily follow this approach to determine which words are clinically significant.

3 Methods

3.1 Clinical BERTScore

Our proposed metric, the Clinical BERTScore (CBERTScore), combines the BERTScore (Zhang et al., 2019) and the medical subset of the Knowledge Graph (Shafran et al., 2020).

BERTScore is a relatively novel language generation evaluation metric proposed in (Zhang et al., 2019) based on pre-trained BERT contextual embeddings. It is designed to capture semantic similarity between two sentences, instead of simple string matching. Given a reference sentence $x = \langle x_1, ..., x_k \rangle$ and a candidate sentence $\hat{x} = \langle \hat{x}_1, ..., \hat{x}_l \rangle$, we first represent each token by a contextual embedding, and then calculate the cosine similarities between the tokens. Each token in the reference sentence is matched to the most similar token in the candidate sentence, and vice versa. The former is used to compute the recall R_{BERT}, and the latter to compute the precision P_{BERT}. Precision and recall are then combined into a single score BERTScore as follows:

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^T \hat{\mathbf{x}}_j,$$

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^T \hat{\mathbf{x}}_j$$

$$\text{BERTScore} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}$$

Building on this, we define CBERTScore:

$$\begin{aligned}\textbf{CBERTScore}(x, \hat{x}) =& k \times \text{BERTScore}_{\text{medical}}(x, \hat{x}) + \\ & (1-k) \times \text{BERTScore}_{\text{all}}(x, \hat{x}) \\ & \text{, where } 0 \leq k \leq 1\end{aligned}$$

$\text{BERTScore}_{\text{all}}$ is computed over all words in the sentences, and $\text{BERTScore}_{\text{medical}}$ is computed over a subset of them that are medically relevant. If there are no medical terms in either the reference or candidate sentence, we define the CBERTScore to be the standard BERTScore (on all words), i.e., k is set at 0.

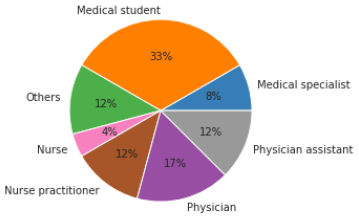
We inject medical information into this metric in two ways. First, we compute a weighted score on a subset of words involving medical terms, as determined by the Knowledge Graph (Shafran et al., 2020). Second, we tune the weight of the clinical term penalty to best match a clinician transcript dataset (CTP) that we collected. We describe our method for determining k in Sec. 3.1.2.

3.1.1 Medical Entities

Similar to (Shafran et al., 2020), we derive roughly 20K medically relevant words from Google's Knowledge graph (Singhal, 2012). These words come from entities with properties such as "/medicine/disease", "/medicine/drug", "/medicine/medical_treatment", and "/medicine/medical_finding". We also include numbers for the CBERTScore algorithm, since numerical accuracy is important in medical contexts.

3.1.2 Tuning the medical entities weight factor

CBERTScore has a parameter controlling the weight of the clinical component. To determine



Original	Transcript 1	Transcript 2	Clinician preference	WER	BLEU	METEOR	BERTScore	CBERTScore
We see a serrated five millimeter hyperplastic polyp during a u-turn.	We see a terraced five millimeter hyperplastic polyp during a u-turn.	We see a serrated five millimeter hyperplastic polyp doing a u-turn.	2	Same	Same	Same	2	2
A biopsy forceps is being used.	A biopsy force is being used.	A biopsy forceps was used.	2	1	1	2	2	2
Three clips were applied to close these defects.	Three clips were applied to close these side effect.	Three clips were replied to close these defects.	2	2	1	2	1	2
Coming down from the descending colon and to the sigmoid colon.	Came down from the descending colon and to the sigmoid colon.	Coming down from the ascending colon and to the sigmoid colon.	1	Same	1	1	2	1

Figure 1: **Left:** Background of the clinicians who were surveyed to create the Clinician Transcript Preference (CTP) dataset. **Right:** Some examples of triplet medical sentences, which transcript clinicians prefer, and which transcript scores better based on different metrics.

this factor, we picked the best performing k on the training subset of the Clinician Transcript Preference (CTP) dataset (Sec. 3.2). We evaluated k using 11 points evenly spaced between 0 and 1, and performed the evaluation methodology in Sec. 3.2 for each. We then used this value for all subsequent results and analyses.

3.2 Clinician Transcript Preference (CTP) Dataset

In order to compare CBERTScore's agreement with human preference, we sent out a Qualtrics survey to elicit judgment specifically from clinicians[2]. We call this dataset the Clinician Transcript Preference dataset (CTP), and we make it publicly available on the Open Science Framework (OSF). To our best knowledge, this is the first publicly available dataset with clinician preferences of transcript errors.

We collected data on 150 sentences. They were divided into three groups, each containing 50 trials. 18 subjects with clinical backgrounds responded to more than half the questions. Fig. 1 (left) describes clinician backgrounds. Each participant was randomly assigned to a group to ensure approximately uniform response coverage. For each trial, participants are given a ground truth sentence and two "transcripts" and asked to select the less useful one or to indicate the two are about the same. An example of such a triplet is as follows: *"Patient elects to go under Propofol sedation."*

#1: Patient elects to go under Prilosec sedation.

#2: Patient selects to go under Propofol sedation.

The survey was designed to take no more than 20 min to minimize the cognitive strain on participants. One sentence was malformed, resulting in 149 sentences for the final dataset.

3.2.1 Constructing the CTP triplets

To generate the triplets of (target, transcript #1, transcript #2) used in the survey, we started by downloading publicly available YouTube videos on colonoscopies created by GI physicians and educational institutes. The target sentences were transcribed by Google's publicly available Speech-to-Text medical dictation model (Soltau et al., 2021) and manually checked for accuracy. Filler words such as "uh" and repeated words were edited out. Sentences longer than 30 words or less than 5 were discarded.

For each target sentence, transcript #1 was generated by one of Google's other, non-medical, publicly available ASR models. Transcripts with an edit distance(edi) outside [1, 3] were discarded. This procedure generated 1220 candidate sentences.

To ensure that the two transcripts were roughly comparable in terms of fidelity, transcript #2 was generated synthetically. We used a publicly available English word frequency dictionary(Goldhahn et al., 2012) to select words in the target sentence that were candidates for synthetic errors. Candidate words were at least 5 characters, appeared in the 1M word dictionary fewer than 10 times, and were not proper nouns. 486 candidate sentences matched these criteria. Finally, transcript #2 was generated by deleting the candidate word or manually substituting it with a phonetically similar word or phrase[3]. We discarded similar sentences and selected 150 triplets for the final survey. The ordering of the two transcripts was randomized, and so were the sentences.

3.2.2 Evaluating metrics on the CTP

To compare the ability of different metrics to agree with rater preference from the CTP, we define a

[2]Broadly defined as a person with extensive clinical experience or from a clinical research background, for our purpose.

[3]A Python fuzz search algorithm based on CMU Pronouncing Dictionary was used for consistency.

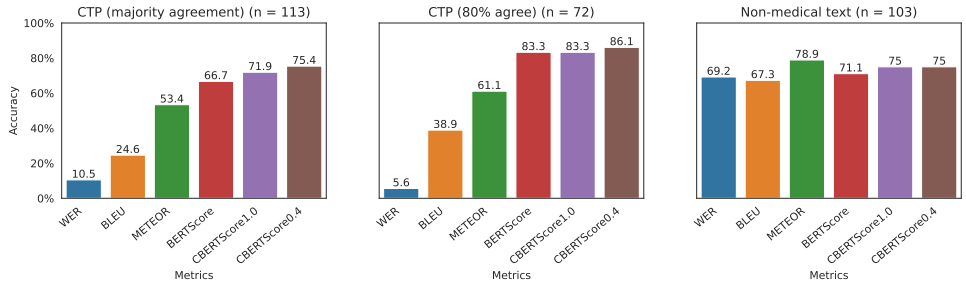


Figure 2: Comparison of different metrics' agreement with human rater transcript preferences. Process of deriving a prediction from metric values is described in Sec. 3.2. In all plots, "CBERTScore1.0" is the performance from only the medical term component ($k = 1.0$ in Sec. 3.1). "CBERTScore0.4" uses the optimal value of k according to the train set. **Left:** Agreements with clinicians on the CTP benchmark when labels are derived using majority voting. **Center:** Agreements with clinicians on the CTP benchmark when restricted to questions with unanimous answers. **Right:** Agreement with speech pathologist raters on the non-medical dataset, when restricting the data to cases where there is a fidelity difference between two candidate transcripts.

3-class classification problem as follows:

$$\text{Predicted better transcript}(M)(gt, t_1, t_2) = \begin{cases} t_1 & M(gt, s_1) - M(gt, s_2) > l \\ t_2 & M(gt, s_1) - M(gt, s_2) < -l \\ \text{same} & \text{else} \end{cases}$$

where M is an evaluation metric, gt is the ground truth sentence, and t_i are the transcripts. Note the predictions are reversed for the WER, since lower values indicate higher fidelity. l is a free variable, which we optimize separately for each metric. We split the data into two halves, choose the best performing l on one half, and report the accuracy using that l on the second half.

3.2.3 Non-medical sentences

To demonstrate that CBERTScore doesn't degrade on non-medical speech, we compare the metrics' agreement with rater preferences on a dataset with annotations similar to (Tobin et al., 2022). Part of this dataset consists of 5-tuples of (ground truth sentence, transcript 1, transcript 2, assessment 1, assessment 2), where the sentence assessments describe how much of the ground truth sentence's meaning is captured in the transcript. We used a subset of 103 utterances from our annotated data where the ratings were not the same, and at least one transcript was rated as having "Major errors". We report performance using a similar formulation as on the CTP evaluation in Sec. 3.2.2: we frame this as a 2-way classification problem (no cutoff is needed since we exclude tuples that have the same rating).

4 Results

4.1 Clinician responses

18 clinicians responded to a total number of 149 triplet questions. Each question had 5 or 6 responses. 78% of questions had more than half agreement on which transcript was less useful and 42% had more than 80% agreement. Clinicians thought transcripts were the same usefulness in 21% of cases.

4.2 Metric agreement on medical text

We report 3-way accuracy classification on the CTP dataset using two labeling schemes (Fig. 2). In the first, we only look at the questions where more than half the respondents agreed. In the second, we report accuracy on the questions where more than 4/5 of the respondents agreed. For both numbers, we determine the cutoff from one half the data and report accuracy on the second half.

First, the metric ordering by performance is the same using both labeling schemes, and the best CBERTScore medical weighting factor was the same using both label schemes. Second, BERTScore and CBERTScore are significantly more closely aligned with clinician preferences than other metrics. Third, CBERTScore weighted entirely toward medical terms outperforms or ties with BERTScore agreement. Fourth, the weighted combination of medical and non-medical terms outperforms other metrics in terms of clinician agreement. Fifth, the medical component meaningfully improves the performance of CBERTScore over BERTScore (75.9% vs 67.2% and 87.5% vs 84.4%).

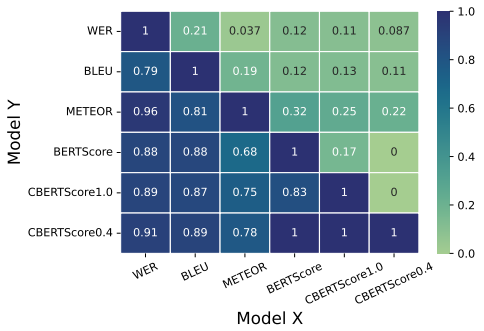


Figure 3: Fraction of cases where metric Y is correctly conditioned on metric X and Y disagreeing. An indicator of how similar the pattern of mistakes is between metrics.

4.3 Metric agreement on non-medical text

CBERTScore was the second best-performing metric on non-medical text. Importantly, the addition of the medical component did not degrade the performance compared to BERTScore.

5 Discussion

5.1 Knowledge Graph medical terms wins and losses on the CTP

The CTP (Sec. 3.2) had 127 distinct words that were the source of transcript errors, and 684 distinct other words. The medically-relevant terms used in the CBERTScore algorithm, identified primarily from the Knowledge Graph as described in Sec. 3.1.1, intersected with 99 of the 127 transcript error words. By manual inspection, 25 of the 28 transcript error words in the CTP not included in the CBERTScore word list were used in a medical context but were not only medical in meaning (ex. “surveillance", “tethered", and “longitudinal"). 3 of the 28 missed words did have a primarily medical meaning, but were not included in the CBERTScore list either due to errors in the KG or errors in the queries generating the list (“cologuard", “colonoscope", “protuberance"). Some of the words have a clear meaning in a medical context, and could be manually added to the list for future applications (“snare", “suctioning", etc.).

The CBERTScore word list included 100 words that weren’t selected for transcript errors. Many of these are medical in nature, but were not selected for synthetic transcript errors via the method described in Sec. 3.2 (ex. “endoscope", “hypoplastic", “lymphoma").

5.2 CBERTScore performance on the CTP

5.2.1 CBERTScore wins

Fig. 3 left shows the degree to which better-performing metrics subsume other metrics, or make a different pattern of mistakes. The plot shows the (Metric Y correct)/(Metric X and Y disagree). Metrics that have higher clinician agreement and a high fraction on this plot are strictly better, whereas metrics with higher agreement but a low value in this plot indicate that another metric might have an additional signal. We see that CBERTScore is nearly strictly better than the other metrics, with the possible exception of METEOR (when they differ, METEOR gives the correct rating in roughly a third of cases).

There were some triplets that CBERTScore got correct that no other metric did. The improvements over BERTScore always involved a medical term, and sometimes involved encouraging the metric to prioritize medical mistakes (ex. "Marked the site with 5 cc's of indigo carmine.“ → "Marked the site with 5 cici's of indigo carmine.“ vs "Marked the sight with 5 cc's of indigo carmine.“)

There were thirteen triplets that the neural word embeddings predicted correctly that other metrics did not. Many of these wins came from the strength of neural word embeddings penalizing less for semantically similar mistakes (ex. "Small burst of coagulation to create a darkish white ablation.“ → "Small burst of coagulation to create a darkish white oblation.“ vs "Small burst of coagulation to create a dark white ablation.“). Furthermore, BERTScore agreed with clinicians on some medical word mistakes, likely due to the BERT embedding somewhat understanding when a transcript error leads to a large semantic change in a medical term (ex. "No ongoing infection or coagulopathy.“ → "No on going infection or coagulopathy.“ vs "No ongoing infection or glomerulopathy.“).

5.2.2 CBERTScore mistakes

Fig. 3 shows that METEOR made the most correct predictions when CBERTScore was incorrect. Some mistakes are due to the KG medical list being incomplete. For example, "longitudinal“ was not included, but has medical meaning in clinical contexts (ex. "The longitudinal extent of the hot snare.“ → "The long eternal extent of the hot snare.“ vs "The longitudinal extend to the hot snare.“).

Another pattern of mistake is when a non-medical adjective contains an error, but the ad-

jective modifies a medical term in an important way. For example, "vessel“ is a medical term, but "feeding“ is not (ex. "This polyp is at high risk of bleeding, with multiple feeding vessels.“ → "This polyp is at high risk of bleeding, with multiple seeding vessels.“ vs "This polyp is at high risking bleeding, with multiple feeding vessels.“). This suggests that future work might include modifications and dependencies when calculating clinical importance.

Finally, a third pattern of mistake involves the fact that METEOR penalizes complex correspondences between candidate and reference sentences, while CBERTScore only considers the best pairwise word matches. One example in the CTP preserves most of the words, but reorders them (ex. "Inject into the head of the polyp, another 1 to 2 cc.“ → "Injectant the head of the polyp, another 1 to 2 cc.“ vs "Inject into the head of the polyp, another 1 2 to cc.“).

6 Conclusions

We present CBERTScore, a novel metric that combines medical domain knowledge and recent advances in neural word embeddings. We collect and release a benchmark of clinician rater preferences on transcript errors, demonstrate that CBERTScore is more closely aligned with clinician preferences, and release the benchmark for the research community to continue to improve ASR in medical contexts.

References

Edit distance. https://en.wikipedia.org/wiki/Edit_distance. Accessed: 2023-03-03.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Victor J Dzau, Darrell G Kirch, Thomas J Nasca, et al. 2018. To care is human—collectively confronting the clinician-burnout crisis. *N Engl J Med*, 378(4):312–314.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Xiaodong He, Li Deng, and Alex Acero. 2011. Why word error rate is not a good metric for speech recognizer training for the speech translation task? In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5632–5635.

Siddique Latif, Junaid Qadir, Adnan Qayyum, Muhammad Usama, and Shahzad Younis. 2021. Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering*, 14:342–356.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jacob Peplinski, Joel Shor, Sachin Joglekar, Jake Garrison, and Shwetak Patel. 2021. FRILL: A non-semantic speech embedding for mobile devices. In *Interspeech 2021*. ISCA.

Izhak Shafran, Nan Du, Linh Tran, Amanda Perry, Lauren Keyes, Mark Knichel, Ashley Domin, Lei Huang, Yu-hui Chen, Gang Li, Mingqiu Wang, Laurent El Shafey, Hagen Soltau, and Justin Stuart Paul. 2020. The medical scribe: Corpus development and model performance analyses. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2036–2044, Marseille, France. European Language Resources Association.

Joel Shor, Aren Jansen, Wei Han, Daniel Park, and Yu Zhang. 2022. Universal paralinguistic speech representations using self-supervised conformers. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Joel Shor, Aren Jansen, Ronnie Maor, Oran Lang, Omry Tuval, Félix de Chaumont Quitry, Marco Tagliasacchi, Ira Shavitt, Dotan Emanuel, and Yinnon Haviv. 2020. Towards Learning a Universal Non-Semantic Representation of Speech. In *Proc. Interspeech 2020*, pages 140–144.

Joel Shor and Subhashini Venugopalan. 2022. TRILLsson: Distilled universal paralinguistic speech representations. In *Interspeech 2022*. ISCA.

Amit Singhal. 2012. Introducing the knowledge graph: things, not strings. https://blog.google/products/search/introducing-knowledge-graph-things-not. Accessed: 2023-03-03.

Hagen Soltau, Mingqiu Wang, Izhak Shafran, and Laurent El Shafey. 2021. Understanding medical conversations: Rich transcription, confidence scores & information extraction. In *Interspeech*.

Ming Tai-Seale, Cliff W Olson, Jinnan Li, Albert S Chan, Criss Morikawa, Meg Durbin, Wei Wang, and Harold S Luft. 2017. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. *Health affairs*, 36(4):655–662.

Jimmy Tobin, Qisheng Li, Subhashini Venugopalan, Katie Seaver, Richard Cave, and Katrin Tomanek. 2022. Assessing ASR Model Quality on Disordered Speech using BERTScore. In *Proc. 1st Workshop on Speech for Social Good (S4SG)*, pages 26–30.

Subhashini Venugopalan, Joel Shor, Manoj Plakal, Jimmy Tobin, Katrin Tomanek, Jordan R. Green, and Michael P. Brenner. 2021. Comparing Supervised Models and Learned Speech Representations for Classifying Intelligibility of Disordered Speech on Selected Phrases. In *Proc. Interspeech 2021*, pages 4843–4847.

Ye-Yi Wang, A. Acero, and C. Chelba. 2003. Is word error rate a good indicator for spoken language understanding accuracy. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, pages 577–582.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. *arXiv e-prints*, page arXiv:1904.09675.