scientific reports

Check for updates

OPEN

Expert evaluation of large language models for clinical dialogue summarization

David Fraile Navarro[1]✉, Enrico Coiera[1], Thomas W. Hambly[2], Zoe Triplett[3], Nahyan Asif[4], Anindya Susanto[1,5], Anamika Chowdhury[6], Amaya Azcoaga Lorenzo[7,8,9], Mark Dras[10] & Shlomo Berkovsky[1]

We assessed the performance of large language models' summarizing clinical dialogues using computational metrics and human evaluations. The comparison was done between automatically generated and human-produced summaries. We conducted an exploratory evaluation of five language models: one general summarisation model, one fine-tuned for general dialogues, two fine-tuned with anonymized clinical dialogues, and one Large Language Model (ChatGPT). These models were assessed using ROUGE, UniEval metrics, and expert human evaluation was done by clinicians comparing the generated summaries against a clinician generated summary (gold standard). The fine-tuned transformer model scored the highest when evaluated with ROUGE, while ChatGPT scored the lowest overall. However, using UniEval, ChatGPT scored the highest across all the evaluated domains (coherence 0.957, consistency 0.7583, fluency 0.947, and relevance 0.947 and overall score 0.9891). Similar results were obtained when the systems were evaluated by clinicians, with ChatGPT scoring the highest in four domains (coherency 0.573, consistency 0.908, fluency 0.96 and overall clinical use 0.862). Statistical analyses showed differences between ChatGPT and human summaries vs. all other models. These exploratory results indicate that ChatGPT's performance in summarizing clinical dialogues approached the quality of human summaries. The study also found that the ROUGE metrics may not be reliable for evaluating clinical summary generation, whereas UniEval correlated well with human ratings. Large language models may provide a successful path for automating clinical dialogue summarization. Privacy concerns and the restricted nature of health records remain challenges for its integration. Further evaluations using diverse clinical dialogues and multiple initialization seeds are needed to verify the reliability and generalizability of automatically generated summaries.

Keywords Natural language processing, Electronic health records, Primary care, Artificial intelligence

Clinical history taking is one of the pillars of medical practice. This is especially true in the context of primary care where clinical history taking and examination take a central role. Clinicians have traditionally kept a record of their patient consultations. With the advent of the informatics era, this record has evolved into an Electronic Health Record (EHR) where, among others, a synthesized summary of a clinical conversation is kept, and produced either during or after each clinical visit. EHR usage has become a major burden for clinicians worldwide[1,2]. Several reasons explain this situation, such as the ever-increasing complexity of records and poor implementation of EHR systems[3] as well as a growing shortage of clinicians, especially in primary care[4]. Effectively, EHRs consume a considerable amount of clinicians' time[5], and are considered one of the causes influencing their burnout[6]. The use of an automated approach to record keeping[7] may prove to be a viable alternative.

Abstractive summarization involves generating lengthy summaries by rephrasing or using new words. This approach differs from extractive summarization, which combines important text from the source. However,

[1]Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Level 6, 75 Talavera Road, North Ryde, Sydney, NSW 2113, Australia. [2]Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia. [3]School of Medicine, Faculty of Human and Health Sciences, Macquarie University, Sydney, Australia. [4]Macquarie University Hospital, Sydney, Australia. [5]Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia. [6]Cowra District Hospital, Cowra, Australia. [7]Health Centre Los Pintores, Madrid Health Services, Madrid, Spain. [8]Health Research Institute, Fundación Jimenez Díaz, Madrid, Spain. [9]University of St Andrews, St Andrews, Scotland, UK. [10]School of Computing, Macquarie University, Sydney, Australia. ✉email: david.frailenavarro@mq.edu.au

abstractive summarization comes with its own set of challenges. These challenges must be addressed to ensure the summaries are of high quality. The resulting summaries should be coherent and accurately represent the source text. To produce effective abstractive summaries, several key challenges need to be overcome: First, as the length of the source text increases, the risk of introducing irrelevant or repetitive content in the summary also rises, which can negatively impact the summary's coherence and usefulness. Second, longer summaries require a deeper understanding of the relationships between the source text concepts and ideas, making it difficult for language models to maintain consistency and avoid contradictions. Third, computational complexity and memory requirements for generating longer summaries are significantly higher, which can hinder the training and deployment of large-scale models, especially in resource-constrained environments. Fourth, evaluating the quality of lengthy summaries becomes challenging, as traditional metrics such as ROUGE[8,9] and BLEU[10] may not fully capture the nuances and subtleties of longer, more complex texts. Fifth, ensuring that lengthy summaries remain faithful to the original text while maintaining a high level of abstraction is a critical challenge, as the models need to strike a delicate balance between comprehensiveness and conciseness. Therefore, developing advanced abstractive summarization models capable of addressing these challenges will be crucial for unlocking their full potential in various domains, including clinical settings, scientific literature, and legal texts.

Natural Language Processing (NLP) has been developing steadily, especially over the last decade, with the advent of word-vectorization[11], transformer models[12], and their derivatives such as BART[13] and T5[14]. Previously deemed as challenging, the possibility of abstractive summarization has become increasingly feasible with newer language models. This is a result of their flexible architectures, ability to fine-tune pre-trained models and improvements in the long text processing capability. Previously, one of the main barriers to perform clinical dialogue summarization was the ability to process full clinical conversations with deep learning architectures. Thanks to recent developments in transformer models, processing medical conversations in full is now feasible. The progress has been exponential the last years, scaling from a few hundred words for older transformers[14] up to more than 1 million tokens (several books) for latest Google's Gemini model[15], which would suffice to process entirely a full patient record.

The aim of this exploratory study is to perform and compare a computational and human expert evaluation of abstractive summarization of clinical dialogues. This evaluation assesses the suitability of various NLP models for clinical use. The study compares the performance of the models using diverse evaluation metrics. Through this comparison, we seek to surface indicative empirical evidence regarding the most effective and reliable approaches for generating accurate, coherent, and contextually relevant summaries that can assist healthcare professionals in their daily practice. By evaluating state-of-the-art transformer-based architectures and Large Language Models (LLMs), we aim to gain insights into their strengths and weaknesses when applied to the domain-specific language and nuanced content of clinical dialogues. We assess the quality of the generated summaries using two approaches. First, we employ computational metrics that focus on common characteristics such as sentence similarity and semantic quality. Second, we conduct human expert evaluations. This dual approach ensures a comprehensive understanding of each model's performance. We specifically focus on how well the models capture and present the critical aspects of clinical dialogues. This approach allowed us to evaluate the quality of the generated summaries. It shows that one system (ChatGPT) outperforms the rest both by human judgment and with most comprehensive computational metrics. Further studies, using larger datasets and more diverse initialization seeds, may be needed to replicate and re-affirm our findings.

Materials and methods

Extending previous research[16], we evaluated fine-tuned and off-the-shelf language models deployed for a clinical dialogue summarization task. The summarized outputs were compared with a human-generated gold standard, and the quality of the output was further evaluated by clinician evaluators.

Experimental setup

We included five models pre-trained for the summarization task: two models were fine-tuned with clinical dialogues, one was trained only with general dialogues, one was trained for long summarization in books, and one was a general-purpose LLM. The main characteristics of the included models are as follows:

- *BART-LSG-conv* – This model is based on BART[13] pre-trained with snippets from clinical dialogues[16]. We utilized the LSG attention mechanism[17] to modify the model and increase the maximum allowed number of tokens to 8,000.
- *BART-DnC* (Divide-and-conquer) – Following the divide-and-conquer approach[18], we utilized the pre-trained model of[16] to produce summaries of snippets. Subsequently, a second summarization step was deployed to generate the final summary.
- *LongDialSumm* – Long Dialogue Summarization mode l[19] based on BART[13] This model was not further fine-tuned.
- *T5-Booksum* - LongT5[20] model pre-trained for long summarization using the BookSum[21] dataset.
- OpenAI's ChatGPT *(text-davinci-003 mode)l*[22].

The two fine-tuned models (*BART-LSG-Conv* and *BART-DnC*) using the clinical dialogue dataset were based on the best performing model in our previous study[16] that are fine-tuned versions of *BART (*Available at https://huggingface.co/dafraile/Clini-dialog-sum-BART). The other models were deployed in a domain-agnostic manner, without any further training. The fine-tuned models where trained using the Huggingface's transformer library with a Pytorch backend on Amazon Web Services (p2.xlarge) instances with an Nvidia Tesla K80 GPU with 12GB or VRAM or a p3.2xlarge with an NVIDIA V100 GPU and 16GB of VRAM. Both instances were configured with 4-core CPU and 61GB of conventional RAM. Given that no additional fine-tuning was

performed on longer dialogues, we did not utilise different initialization seeds beyond the fine-tuning already reported in[16]. Training scripts are available in the Supplementary File 1 including model configuration and training configuration details.

Dataset

The dataset exploited in this study was sourced from 27 anonymized clinical dialogues recorded from primary care consultations, a subset of data reported by Quiroz et al.[23]. This data was further processed for summarization tasks as described by Navarro et al.[16]. For the models that were fine-tuned (*BART-LSG-conv and BART-DnC*), as the dialogue transcripts exceeded the token limit (512 or 1024 tokens) set by these language models, a pre-processing phase was necessary. To preserve the structural integrity of the conversations, patients were assigned generic identities (Joe and Jane) while the clinician was generically named and referred to as "Doctor". The transcripts were divided into 400-word segments, which were further refined to maintain semantically consistent pairs of clinician-patient interaction, i.e., a doctor's question followed by a patient's answer. A small portion of the segmented snippets (< 5%) was discarded due to the lack of relevant clinical content. These snippets contained non-relevant clinical text (such as administrative discussions), or were part of the introductory or closing parts of the conversations, without any medical relevance (e.g. discussing holidays).

The dataset was subsequently split in an 80 − 20 ratio for training and evaluation purposes. Given the data split ratio, 22 dialogues were used for training and the remaining 5 dialogues were used for evaluation purposes. In total the dataset contained 56,158 tokens for the clinical dialogues (41,501 for training + 14,657 for testing) and 9,784 in the summaries (7,040 for training and 2,744 for testing). The average dialogue length was 391.52 tokens (median 418, standard deviation 85.28) and the average summary length was 66.42 tokens (median 67.5, standard deviation 21.42).

Evaluation

We kept the same evaluation subset used in our previous evaluation[16]. Given that the original snippet annotations did not generate full dialogue summaries, new expert-generated summaries were created by a clinician (DFN) and reviewed by a second clinician (AAL). To generate semantically self-contained summaries that were semantically contained in the dialogues, the human summaries conformed to two rules. First, they only used the vocabulary that was present in the dialogue (e.g. if a patient or doctor described a symptom as "out of breath" it would not use the more technical term dyspnea). Second, the human-produced summaries were created as a recollection of descriptive sentences, with minimal stylistic features (e.g. not splitting it into paragraphs or adding headings) and containing only a recollection of the clinically relevant information (e.g. past medical history, drug use, family history). An example of a generated summary by one of the language models and the human-generated summary (gold standard) is shown in Box 1 (the complete original dialogue is available in Supplementary File 1).

Evaluation metrics

Summaries were evaluated with quantitative and qualitative methods, using both automated and manual scores. For the automated scores, the ROUGE[24] family metrics were used. Additionally, we used UniEval[25] to automatically compute 4 quality dimensions (or, domains) of the produced summaries. UniEval, a unified multi-dimensional evaluator for Natural Language Generation (NLG), reframes evaluation as a Boolean question-answering task. By incorporating external knowledge and guiding the model with different questions, UniEval was shown to achieve a higher correlation with human judgements and demonstrated strong zero-shot learning capabilities in unseen evaluation dimensions and tasks[25].

Qualitative evaluation

Five medically trained clinicians performed the human evaluation (TH, ZT, NA, AC, AS) assessing the summaries generated by the five systems and the gold-standard human-generated summaries. They were blinded regarding which system they were evaluating including the human summaries. To avoid priming, systems were presented in a random order, allowing the evaluator to access them in a non-consecutive way.

Human evaluators were explained the task and the scoring with respect to the different evaluation domains. The human evaluators followed the same 4-domain structure proposed in UniEval[25]: coherence, consistency, fluency and relevance. The evaluators were encouraged to resolve with the investigators any difficulties arising upon completing the first dialogue to ensure consistency in the evaluation.

LongT5-booksum summary: The Doctor meets with Joe, a student at the university who has been complaining of a feverish cough and sore throat for two weeks. He also complains about having pressure in his ear as well as a cold. The Doctor asks Joe if he is on any regular medicine, and Joe replies that he's an international student studying business and finance from the United States. They discuss his family history, which includes his mother and his siblings. His father has heart disease, and his sister has surgery just two days ago. When the Doctor looks at Joe's ear, he notices nodes of redness but doesn't know whether it's caused by a disease or a infection. The doctor decides to give Joe a penicillin to prevent him from getting sick again.
Human summary: *Joe had a cough and a sore throat for the last couple of days. Today he wasn't feverish when waking up, but he was the last two days and he has been taking ibuprofen. He has been sick like this a couple of times. A year and half ago, he had five tonsillitis episodes within the year. This time is different because he also has cough. He has no allergies. Occasionally he has difficulties with breathing when exercising but he has not been diagnosed of asthma. Joe is an international student from USA studying international business and finance. He doesn't smoke and he drinks occasionally. His dad has heart issues due to high blood pressure. His sister has a heart condition too, she was born with and is having surgery. He doesn't have a fever now. Tonsils are enlarged and lymph nodes are a bit sore, which is consistent with a viral infection. Doctor listened to Joe's chest and is fine. Doctor takes a throat swab to rule out a particular type of strep, in case that he's positive he would have to take penicillin for 10 days. Doctor orders penicillin one tablet twice a day. Joe will start the treatment until the results come back on Monday. He can take Ibuprofen, honey, lemon or Strepsils or Lozenges from the chemist. Joe will come back on Monday to check the results.*

Box 1. Generated summary by *LongT5-booksum* model (top) and human-generated summary (bottom).

The decimal scoring of UniEval was converted into a 5-point Likert scale to facilitate human evaluation. Instead of producing an overall score as in the UniEval metric, human evaluators were asked an additional question that reflected on the overall potential for clinical use: "To what extent do you think the summary could be used in a clinical setting? (e.g. to produce a progress, note in primary care)", which was also evaluated on a 5-point Likert scale. Following each question, a free-text box allowed the evaluators to justify their scores and provide examples. The obtained free-text feedback was evaluated by extracting the commonalities across the answers using a bottom-up thematic analysis[26].

Statistical analysis

We calculated inter-rater reliability using intra-class correlation coefficients (ICC) between human evaluators themselves and between humans and the automated (UniEval) metrics. For this calculation, we used the average score across the evaluated dialogues for each evaluator, for each system and each dimension (one intra-class correlation coefficient for each dimension and system). Additionally, individual dialogue scores were calculated and are also reported in Supplementary File 2.

For the comparison between human evaluations and UniEval, an average score (converted to a decimal scale) for the human scoring was compared with the automated UniEval score. To compare the performance of the studied systems across the evaluated domains, we used repeated measures ANOVA, with a significance level set at $p = .05$, and a post hoc analysis using a t-test with the Bonferroni correction. We established comparisons between the model-generated and gold-standard human summaries using the scorings obtained from the human evaluation. Additionally, we repeated the analysis using the scores obtained with UniEval.

All the study methods were carried out in accordance with Macquarie University research policies. Experimental protocol was deemed exempt from requiring additional ethics approval by the Macquarie University, School of Computing Ethics liaison. Original data collection Ethics Approval available at[27] where informed consent was obtained from all subjects and/or their legal guardian(s).

Results

In this exploratory study, when summaries were evaluated using ROUGE, the highest ROUGE-1, ROUGE-2 and ROUGE-L-SUM scores were obtained by the BART-LSG-conv model pre-trained with clinical dialogue snippets, while ChatGPT scored the lowest for ROUGE-1 and ROUGE-L-Sum. For the ROUGE-L metric, BART-LSG-conv and LongDialSumm scored similarly, while T5-Booksum was the lowest for ROUGE-2 and ROUGE-L. Figure 1 represents the ROUGE scores across the different systems and evaluation metrics.

UniEval Scoring

Applying the UniEval scoring, ChatGPT scored the highest for coherence, consistency, fluency, relevance and overall. LongDialSumm achieved the lowest scores in all the domains except for consistency, where the lowest score was obtained by T5-BookSum (Table 1). Note that human-generated summaries were excluded from the automated UniEval evaluation, as they were used by UniEval as the gold standard to calculate the scores of the automated summarization systems.

Human evaluation

When clinicians evaluated the summaries concerning the four UniEval domains and an additional item judging the overall potential for use in a clinical setting, ChatGPT scored the highest in coherency, consistency fluency and clinical use, while human-generated summaries scored the highest in the relevance domain (Table 1).

The intra-class correlation coefficient showed generally excellent reliability among the human evaluators when comparing the averaged ratings across the evaluated dialogues. When comparing the averaged human evaluation with UniEval, excellent reliability was found for coherency, relevance and overall, while consistency

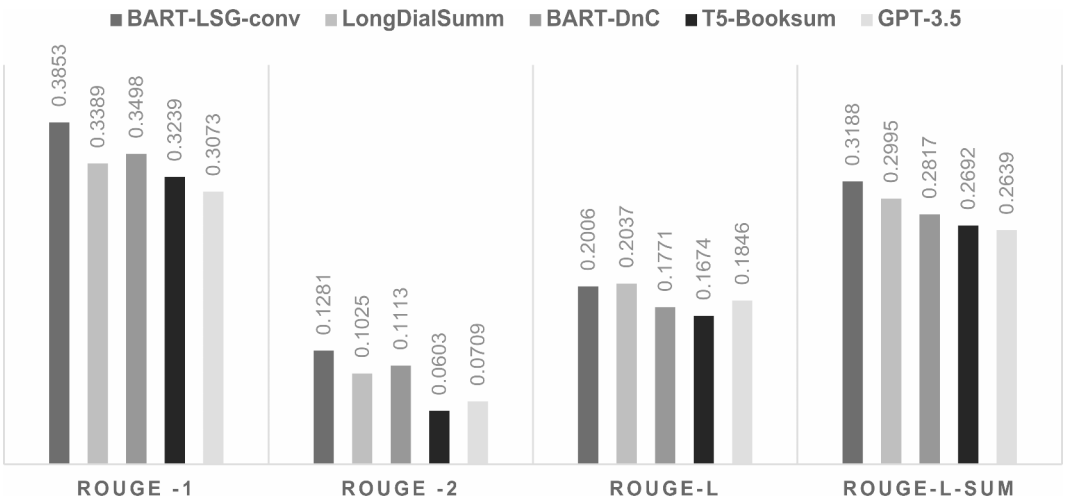

Fig. 1. ROUGE scores by system.

UNIEVAL scores	Coherence	Consistency	Fluency	Relevance	Overall
LongDialsumm	0.53153	0.61116	0.6438	0.2597	0.5115
ChatGPT	**0.9547**	**0.73634**	**0.9576**	**0.9437**	**0.8981**
T5booksum	0.68484	0.56454	0.9341	0.6486	0.708
BART-DnC	0.59903	0.61694	0.7925	0.4575	0.6165
BART-LSG-conv	0.52529	0.69167	0.8451	0.4632	0.6313
Human evaluation scores	**Coherence**	**Consistency**	**Fluency**	**Relevance**	**Clinical use**
Human summaries	0.904	0.872	0.832	**0.896**	0.848
BART-LSG-conv	0.48	0.488	0.544	0.496	0.408
LongDialSumm	0.452	0.562	0.474	0.432	0.414
T5-BookSum	0.544	0.456	0.536	0.472	0.392
Bart-DnC	0.504	0.616	0.496	0.52	0.456
ChatGPT	**0.973**	**0.908**	**0.96**	0.886	**0.862**

Table 1. Automated UniEval and human evaluation scores for each system (average scores across 5 dialogues). Top result values are in bold.

Domain	ICC	F	df1	df2	p-value	CI95%
Average fixed raters ICC human evaluation						
Coherency	0.971	34.201	5	20	<**0.001**	[0.9, 1.0]
Consistency	0.961	25.818	5	20	<**0.001**	[0.87, 0.99]
Fluency	0.959	24.691	5	20	<**0.001**	[0.87, 0.99]
Relevance	0.947	18.748	5	20	<**0.001**	[0.82, 0.99]
Overall (clinical use)	0.979	47.759	5	20	<**0.001**	[0.93, 1.0]
Average fixed raters ICC human evaluation vs. UniEval						
Coherency	0.974	37.816	4	4	**0.002**	[0.75, 1.0]
Consistency	0.658	2.925	4	4	0.162	[-2.28, 0.96]
Fluency	0.689	3.213	4	4	0.142	[-1.99, 0.97]
Relevance	0.901	10.105	4	4	**0.023**	[0.05, 0.99]
Overall	0.89	9.132	4	4	**0.027**	[-0.05, 0.99]

Table 2. ICC coefficients for averaged scores by domain for all systems. Significant values are in bold.

and fluency reached moderate reliability (Table 2). Supplementary file 2 contains the individual scores for each dialogue.

The repeated measures ANOVA test indicated a significant difference in the scores between the systems across all domains: $F(5,20) = 28.18$, $p < .001$, for the coherency domain, $F(2.23,8.93) = 7.58$, $p = .011$ for consistency, $F(2.26,9.05) = 27.6$, $p < .001$ for fluency, $F(5,20) = 22.47$, $p < .001$ for relevance and $F(1.99,7.94) = 20.41$, $p < .001$ for the overall clinical use (Table 3 and supplementary file 3). The post hoc paired t-test using a Bonferroni correction showed that the means of several pairs were significantly different for human-generated and ChatGPT-generated summaries compared to all the other systems (Table 3 and supplementary file 3). When analysing UniEval automated scores, repeated measures ANOVA found significant differences for fluency ($p = .008$), relevance ($p < .001$) and overall ($p = .003$) (supplementary file 3).

Qualitative evaluation

Human evaluators provided comprehensive justifications for the scores they assigned across various domains, along with the overall clinical usefulness. These explanations offered valuable insight into the distinct types of errors and limitations that surfaced, as well as the main factors influencing their scoring decisions. We analyzed the feedback received for the potential clinical use and categorized it into three scoring categories. For scores 1 and 2 (Poor Quality / Limited Usefulness), for score 3 (Adequate but Needs Improvement) and for summaries scores of 4 and 5 (Good to Excellent Quality).

For the category of coherency, evaluators frequently cited non-sequential sentence ordering and inconsistent verb tense usage as factors contributing to lower scores. Medium scores were given in cases where the system unexpectedly incorporated elements of the original dialogue. Higher scores were attributed to the presence of logically ordered sentences.

Regarding consistency, evaluators noted that low scores were primarily due to the usage of grammatically incorrect sentences, inaccuracies, factual errors, and the inclusion of hallucinatory content. Medium scores were triggered by the absence of certain clinical information, while high scores were reserved for summaries

	Sum of square	Mean square	F Statistic (df1, df2)	*p*-value
Coherency	32.276	6.4552	28.1845 (5,20)	<**0.0001**
Consistency	20.8323	4.1665	7,5815 (5,20)	<**0.001**
Fluency	245,524	49,105	27,6025 (5,20)	<**0.0001**
Relevance	35.0293	7.0059	22.4653 (5,20)	<**0.0001**
Overall	33.9617	6.7923	20.4076 (5,20)	<**0.0001**
	BART-LSG-Conv	**BART-DnC**	**LongDialSumm**	**T5-Booksum**
Coherency				
Human		*		*
ChatGPT	*	*	*	*
Consistency				
Human	*			*
ChatGPT				
Fluency				
Human		*		
ChatGPT	*	*	*	*
Relevance				
Human				*
ChatGPT	*	*	*	*
Overall (clinical use)				
Human				*
ChatGPT	*	*	*	*

Table 3. Repeated measurements ANOVA scores for between systems differences and post hoc t-tests for model differences by domain using human evaluation scoring. *Denotes a significant difference (with Bonferroni correction applied for a baseline alpha = 0.05). Significant values are in bold.

that demonstrated consistent information flow, even if the subject matter order was lacking. When evaluating fluency, evaluators determined low scores by the presence of nonsensical sentences and repetitive content. Summaries containing non-technical words received medium scores. High scores were assigned to summaries that were well-articulated and exhibited correct grammar. In terms of relevance, summaries with missing factual content typically resulted in low scores, whereas summaries with incorrect minor facts led to medium scores. High scores were given to summaries that captured key information, even if they omitted minor details.

In relation to the overall potential for clinical use, summaries falling under low scores were characterized by factual inaccuracies, deficiencies in crucial clinical information, incoherent sentence construction, disjointed information presentation, and poor language or grammar usage. These summaries are considered to have limited use or are deemed unusable in a clinical setting without substantial improvement. Summaries that garnered medium scores generally encapsulated the main points of the consultation. However, they could potentially be lacking in key details or contain inaccuracies. However, with a thorough review and necessary corrections or additions, these summaries could still be employed in a clinical setting. Finally, summaries awarded high scores exhibited high levels of accuracy, excellent structure, and effectively captured the central topics of the consultation. While they may contain minor errors, omissions, or areas necessitating improvement, they largely satisfy the requirements for clinical use, requiring little to no revisions.

Table 4and Supplementary File 4 provides a detailed breakdown of these factors, accompanied by specific examples that illustrate the feedback from the evaluators.

Discussion

The reported expert evaluation of clinical dialogue summarization suggests that while fine-tuned with clinical dialogues models (*BART-LSG-conv*) outperform those that were not fine-tuned with respect to classical evaluation scores (ROUGE metrics), this did not translate to improved quality or usefulness when more granular metrics such as UniEval or human evaluation were applied. Strikingly, the worst performing model with the ROUGE metrics (ChatGPT) consistently performed best when evaluated both with UniEval and human evaluation, outperforming all other models. ChatGPT results show significant improvements in the quality of the summaries compared to previously developed BART and T5 transformer models. Additionally, our findings suggest that the quality of the ChatGPT summaries may be comparable to the quality of human-generated summaries when assessed by clinicians. We have also shown a strong correlation between human evaluation and the automated UniEval metrics, validating the usefulness of this metric in the clinical summarization scenario.

While the existence of clinical dialogue datasets for summarization remains a major obstacle to the practical implementation of these approaches, thanks to the above developments, especially LLMs, the potential for producing, accurate, clinical dialogue summaries is within reach. Ultimately, they have the potential to advance

	Low (1–2)	Medium (3)	High (4–5)
Coherency	**Non-sequential order**: Each sentence is not in sequential order. *(BART-DnC, Evaluator 5, Dialogue 1)* **Mix of verb tenses**: The mix of past and present tense makes this difficult to read and understand. *(BART-DnC, Evaluator 5, Dialogue 1)*	**Incorporating dialogue**: Starts somewhat coherent, but degenerates about halfway through when it incorporates some dialogue ("Doctor:") and begins discussing 22q11. *(T5-Booksum, Evaluator 1, Dialogue 5)*	**Adequate sequential order**: The text coherently documents the key findings from the conversation in a sequential matter from the patient's symptoms to their allergies and family history and then to the examination and plan. *(ChatGPT, Evaluator 5, Dialogue 1)*
Consistency	**Incorrect sentences: "**Sometimes it is just the right side. Sometimes it is just the click." isn't correct - it is only the right side, and the clicking is consistent. (BART-LSG-conv, *Evaluator 1, Dialogue 2)* **Lack of accuracy**: Sometimes it is a combination of pain relief, time, time and possibly some specialized physiotherapy." appears to be saying that he has used these treatments before, which isn't accurate. (BART-LSG-conv, *Evaluator 1, Dialogue 2)* **Factual errors**: He is allergic and takes Diazepam five times a week." this is wrong - he is not allergic to anything, and only takes the diazepam 1–2 times per week. The name is incorrect (listed as John, when the document says Joe). (BART-LSG-conv, *Evaluator 1, Dialogue 2)* **Hallucination**: […]makes up multiple surgeries that the patient didn't have." (LongDialogSumm, *Evaluator 1, Dialogue 4)*	**Omitting (some) clinical information**: The majority of facts listed are correct, although some modest errors above. Omits significant amounts about diagnosis, treatment and history/exams. *(BART-DnC, Evaluator 1, Dialogue 3)*	**Consistent but lacking subject matter order**: Consistent, however, misses a lot of important clinical detail compared to other models. And does not arrange subject matter within sentences in a sequence as one would expect for a clinical summary. *(LongDialogSumm, Evaluator 2, Dialogue 1)*
Fluency	**Lack of sense**: Very poor-quality sentences, at times using phrases that are themselves non-sensical such as "feverish cough". (T5-Booksum, *Evaluator 3, Dialogue 1)* **Repetition**: Not fluent. Long sentences, repetition in sentences. (BART-LSG-conv, *Evaluator 5, Dialogue 4)*	**Use of non-technical words**: […] Text also uses words such as 'a whole bunch of colds' and 'bugs' which are not of very high quality for medical documentation purposes: 'He has head airiness and pressure in his ears as well as a couple of times.' (BART-LSG-conv, *Evaluator 5, Dialogue 1)*	**Well-written, correct grammar**: All the sentences make sense and are well-written. (ChatGPT, *Evaluator 1, Dialogue 1)* Well-composed sentences make sense, and grammar is good. (ChatGPT, *Evaluator 1, Dialogue 1)*
Relevance	**Missing facts**: The summary is missing the majority of the relevant facts including examination findings, diagnosis, medication prescribed, follow-up plan and previous tonsillitis episodes […]. (LongDialogSumm, *Evaluator 5, Dialogue 1)*	**Incorrect facts**: Most facts are relevant, if somewhat butchered. More problematic is that the facts are often wrong. Missing much of the content it should have. […] Interprets examination finding incorrectly ("little lymph nodes consistent with a virus or infection") (T5-Booksum, *Evaluator 1, Dialogue 2)*	**Important items are present**: Summarizes many of the important features of the case. Could mention the past medical history and could also mention the plan for follow-up as well. Could also include exam findings. (BART-DnC, *Evaluator 1, Dialogue 1)* **Missing detail**: Contains most components of a good history […]. However, it did not contain information including lack of regular medications, lack of past medical history and social history[…] It also did not document the examination findings or the time frame of his symptoms (2 weeks). (CHATGPT, *Evaluator 5, Dialogue 1)*
Clinical use	**Unsuitable**: I do not think this summary could be used in a clinical setting[…]. Its largest downfall is the lack of discussion regarding the management and follow-up plan for this patient and in documenting an impression or diagnosis. (BART-LSG-conv, *Evaluator 5, Dialogue 1)* **Irrelevant information**: It also documents a lot of irrelevant information and misinterprets some information discussed pertaining to pathogens and the symptoms experienced currently versus in previous similar episodes of illness. (BART-LSG-conv, *Evaluator 5, Dialogue 1)* **Misinterpretation and lack of clinical information**: Not usable in a real-world clinical scenario due to misinterpretation of examination findings and paucity of other clinical information required for a medical summary. (BART-DnC, *Evaluator 2, Dialogue 1)*	**Missing important findings**: This doesn't include key exam findings and treatment plans. The facts are mostly correct. Discusses x-ray results at too much length. (LongDialogSumm, *Evaluator 1, Dialogue 3)* Reasonable summary, however, some obvious mistakes. Missing some content about examination findings and the whole treatment plan/advice. (T5-Booksum, *Evaluator 1, Dialogue 4)*	**Concise enough**: Can be used to provide a concise summary of the consultation, that the physician can then corroborate/explore further with the original medical progress note/consult notes. (CHATGPT, *Evaluator 2, Dialogue 3)* **Useful**: Good note overall, but several minor mistakes as mentioned previously. This could actually be useful for a clinician. (CHATGPT, *Evaluator 1, Dialogue 1)*

Table 4. Human evaluation: examples by domain and reasons provided for each scoring.

the development of tailored abstractive summarization tools for the healthcare domain, enhance communication among medical professionals, improve documentation accuracy, and consequently facilitate better patient care.

Strength and limitations

This exploratory study has several strengths. First, it produces an empirical comparison of pre-trained models deployed for clinical dialogue summarization tasks evaluated by expert users (clinicians) which provided a thoroughly evaluation of the models' outputs, both quantitative and qualitative. Our study adds value to previous ones as it is able to produce full-dialogue summarization, compared to previous research[16,28] that only produced summaries of clinical dialogue snippets, as models were unable to capture the entire clinical conversations, given the lack of powerful models with enough context length. It additionally compares various automated summarization evaluation methods, the established ROUGE family of metrics, and the newer, more comprehensive UniEval. One of the key limitations of our study concerns the lack of multiple initialization seeds for the fine-tuned models (BART-LSG-Conv and BART-DnC). Given that transformer models like BART are sensitive to their initial weights, using only a single initialization might not fully capture the models' performance variability. While the pronounced performance gap between ChatGPT and other models suggests that different initialization parameters would unlikely affect our main conclusions, this limitation should be addressed in future work.

Our findings also align with previous findings regarding the advantages of using UniEval[25], which are also supported by our work, demonstrating its applicability to new data with context-specific domain knowledge (clinical medicine) and a distinct textual structure (dialogues). Moreover, we also produced an original

evaluation setting, comparing automated systems with human-generated summaries demonstrating that ChatGPT produces summaries comparable to the human ones, with important implications for the clinical documentation applications.

Several other limitations need to be highlighted. First, given the small number of dialogues in the dataset, we could not extensively fine-tune models for the summarization task, dismissing potential improvements achievable if a larger dataset was used. Likewise, our small number of clinical dialogues samples for evaluation, limits the generalizability of these findings and points to further evaluation with larger sets and across different specialties and medical domains. However, in this exploratory study, we have detected a strong signal favouring summaries generated by LLMs, consistently with the emerging literature exploring the capabilities of those models. Additionally, we have not explored in-depth parameter tuning, evaluated performance in a k-fold validation manner, possibly limiting the generalizability of results. In relation to reproducibility aspects and ChatGPT, while additional configurations or prompting strategies could have been explored, after initial testing and given the onerous task of manual evaluation by human experts we decided to focus on the best performing prompt and default configuration (reflecting the most likely scenario for regular non-technical users). Lastly, it is important to note the subjective nature of human evaluation scores. We mitigated this by pooling estimates of 5 clinical expert evaluators, presenting the responses in a random order to avoid priming, and masking the automatically generated summaries as well as the human-produced ones, to minimize potential interference with the scoring.

The strength of our ratings is confirmed by the standardized metrics of reliability (ICC), which is also maintained when compared with the automated metric UniEval. An additional limitation may have emerged from the quality of the human-generated summaries. As a single clinician produced those summaries, they may not be representative of the summaries produced in the clinical setting, as well as they may vary from one clinician to another. Considering that these summaries were produced also with some stylistic constraints, this might have penalized the ROUGE scoring of ChatGPT while also picturing the human-generated summaries as less stylistically adequate than expected. While not removing completely this limitation, we mitigated it by using another clinician to review and suggest corrections with the human-generated summaries.

Although the results of ChatGPT are encouraging an important limitation to its use surrounds the potential data privacy concerns that may arise from using it to process private medical information. In the light of this, it is important to note that open-source alternative and smaller-scale models that can run on premises or on the device may pose better alternatives and will require further evaluation, which we plan to conduct in the future.

Relation to previous research

Previous studies focusing on clinical dialogue summarization[16,29] have not produced conclusive results. First, the length of the clinical conversations was unsuitable for earlier language models to process into a full, coherent summary and were limited to summarize snippets of conversations. Second, given the paucity of clinical data, it remained challenging to produce models that were reliable and performed the task consistently[28], while summarization of other types of documents such as news[30] or law texts[31] had larger corpora available for training. Recent advancements can be seen in the creation of a contest of synthetic clinical notes and summarization evaluation[32].

Our study also confirms the strong capabilities of the newer LLMs, illustrated by the GPT-derived models[33]. These models have outperformed previous approaches in several tasks[34] including summarization, while not requiring special fine-tuning or retraining of a custom model and being able to respond to a wide range of questions and use cases. More recently, studies applied to the context of clinical summarization had also appeared[35,36], showcasing the benefits of this approach in different types of medical text such as radiology reports[37] and clinical dialogues.

The results of our study are also in line with previous summarization metric evaluations that showed the suboptimal quality of the ROUGE metrics[9] especially when applying to the medical domain[38] and increasing context length for multi-document summarization[39]. Among the limitations that ROUGE encounters when processing long text, is its reliance on overlapping n-grams to calculate its scores. This may explain partially how models using a zero-shot approach (such as ChatGPT) may underperform when measured with ROUGE, as they diverge more broadly from the expected summaries, particularly for long texts such as clinical dialogues. Further exploration of ROUGE as a metric is needed, especially for evaluating when the generated summaries differ from the gold standard while maintaining their quality. Likewise, this study confirms previous findings proposing new, automated, unified summarization UniEval metrics[25]. When deployed in a different evaluation scenario of summarizing clinical dialogues, these metrics still produced similar results to human evaluators.

Follow up research

Given the exploratory nature of our research, evaluating with a wider variety of clinical settings (primary vs. secondary care), different clinical specialties (with specific medical vocabulary and acronyms), as well as site- and context-specific variations (different hospitals using different vocabulary) are vital to ensuring the generalizability of our findings. Additional research needs to explore whether LLMs consistently summarize clinical facts present in clinical dialogues, ensuring accuracy, completeness and clinical usefulness. Recently, open-source LLMs have been released, including LLaMA[39] and its derivatives[40], and models trained on medical texts[41]. The use of open-source LLMs may provide similar performance advantages in summarization, while maintaining control over the model ownership and data governance. Moreover, the potential to be deployed locally, or be fine-tuned for specific tasks is particularly important in a highly regulated environment such as healthcare. In terms of metrics and evaluation, future research may include additional evaluation metrics suited for transformer-based architectures such as BERTScore[42] or BARTScore[43]. Further metrics proposed for LLMs may include a more comprehensive approach such as an "ecosystemic" evaluation[44]. Additional benefits

may be ripped from an ensemble approach that combines a pretrained transformer to perform Named Entity Recognition, with an LLM such as a GPT or LLAMA tasked with the summarization and natural language generation parts.

While not the focus of this study, an important area of research refers to integrating these text automation tools in day-to-day clinical practice. As previous findings have shown, there is sometimes little correlation between model development, its reported performance, and its implementation in practice[45]. Exploring clinicians' needs, the fit of automated summarization tools into clinical pipelines, and safe implementation in a high-stake scenario such as medicine remain open questions. In relation to these findings, integration with current or future EHRs and exploring user experience and interaction aspects of clinical dialogue summarization is a crucial step towards their swift and meaningful adoption, to ensure safety and usefulness, while complying with medico-legal issues and potential clinicians' resistance to change[46].

Conclusions

Our exploratory findings suggest that LLMs such as ChatGPT can effectively perform clinical dialogue summarization tasks, consistently producing summaries not differing in quality from human-generated ones and outperforming previous approaches, while not requiring fine-tuning. In the LLM era, and especially when performing long-text summarization, the performance of ROUGE-based metrics may not reflect the real performance of the models, unfairly penalizing the models that have not seen domain-specific training data, while being more capable than their pre-trained counterparts. These findings question the usefulness of such metrics, pointing at potentially replacing them with more comprehensive metrics, such as UniEval. Lastly, our results indicate that clinical dialogue summarization is a feasible task in the era of LLMs. Exploring how to bring summarization into practice, especially considering privacy concerns and the restricted nature of health records, remains an open question.

Data availability

The data that support the findings of this study are not openly available due to reasons of sensitivity and are available from the corresponding author upon reasonable request. Data are located in controlled access data storage at Macquarie University.

Received: 16 July 2024; Accepted: 27 December 2024
Published online: 07 January 2025

References

1. Dymek, C. et al. Building the evidence-base to reduce electronic health record–related clinician burden. *J. Am. Med. Inform. Assoc.* **00**, 1–5 (2020).
2. Frintner, M. P. et al. The effect of electronic health record burden on pediatricians' work–life balance and career satisfaction. *Appl. Clin. Inf.* **12**, 697–707 (2021).
3. Fortune, F. S. & Fry, E. Death by 1,000 clicks: Where electronic health records went wrong. *Kaiser Health News* (2019). https://fortune.com/longform/medical-records/ Accesed 4 November 2024.
4. Cerny, T., Rosemann, T., Tandjung, R. & Chmiel, C. Reasons for general practitioner shortage: A comparison between France and Switzerland. *Praxis (Bern 1994)*. **105**, 619–636 (2016).
5. Farber, J., Siu, A. & Bloom, P. How much time do physicians spend providing care outside of office visits? *Ann. Intern. Med.* **147**, 693–698 (2007).
6. Yan, Q., Jiang, Z., Harbin, Z., Tolbert, P. H. & Davies, M. G. Exploring the relationship between electronic health records and provider burnout: A systematic review. *J. Am. Med. Inform. Assoc.* **28**, 1009–1021 (2021).
7. Coiera, E., Kocaballi, B., Halamka, J. & Laranjo, L. The digital scribe. *NPJ Digit. Med.* **1**, 58 (2018).
8. Lin, C-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* 74–81 (Association for Computational Linguistics, Barcelona, Spain, 2004).
9. Akter, M., Bansal, N. & Karmaker, S. K. Revisiting automatic evaluation of extractive summarization task: Can we do better than {ROUGE}? In *Findings of the Association for Computational Linguistics: ACL 2022* 1547–1560 (Association for Computational Linguistics, Dublin, Ireland, 2022).
10. Papineni, K., Roukos, S., Ward, T. & Zhu, W-J. Bleu: A method for automatic evaluation of machine translation 311–318 (2002).
11. Pennington, J., Socher, R. & Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 1532–1543 (2014).
12. Vaswani, A. *Attention is all you need* (Advances in Neural Information Processing Systems, 2017).
13. Lewis, M. et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. https://arxiv.org/abs/1910.13461 (2019).
14. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer https://arxiv.org/abs/1910.10683(2019).
15. Gemini Team, Anil R, Borgeaud S, et al. Gemini: A family of highly capable multimodal models. https://doi.org/10.48550/arXiv.2312.11805 (2023).
16. Navarro, D. F., Dras, M. & Berkovsky, S. Few-shot fine-tuning SOTA summarization models for medical dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop* 254–266 (Association for Computational Linguistics, Hybrid: Seattle, Washington, 2022).
17. Condevaux, C. & Harispe, S. LSG Attention: Extrapolation of pretrained transformers to long sequences. https://doi.org/10.48550/arXiv.2210.15497 (2022).
18. Gidiotis, A. & Tsoumakas, G. A Divide-and-conquer approach to the summarization of long documents. https://doi.org/10.48550/arXiv.2004.06190 (2020).
19. Zhang, Y. et al. *An Exploratory Study on Long Dialogue Summarization* (What Works and What's Next, 2021).
20. Guo, M. et al. LongT5: Efficient text-to-text transformer for long sequences (2022).
21. Kryściński, W., Rajani, N., Agarwal, D., Xiong, C. & Radev, D. BookSum: A collection of datasets for long-form narrative summarization. https://doi.org/10.48550/arXiv.2105.08209 (2022).
22. Brown, T. et al. Language models are few-shot learners. *Adv. Neural. Inf. Process. Syst.* **33**, 1877–1901 (2020).
23. Quiroz, J. C. et al. Identifying relevant information in medical conversations to summarize a clinician-patient encounter. *Health Inf. J.* **26**, 2906–2914 (2020).
24. Lin, C-Y. Rouge: A package for automatic evaluation of summaries 74–81 (2004).

25. Zhong, M. et al. Towards a unified multi-dimensional evaluator for text generation (2022). https://arxiv.org/abs/2210.07197v1. Accessed 18 Mar 2023.
26. Braun, V. & Clarke, V. Using thematic analysis in psychology. *Qual. Res. Psychol.* **3**, 77–101 (2006).
27. Kocaballi, A. B. et al. A network model of activities in primary care consultations. *J. Am. Med. Inform. Assoc.* **26**, 1074–1082 (2019).
28. Chintagunta, B., Katariya, N., Amatriain, X. & Kannan, A. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Proceedings of the 6th Machine Learning for Healthcare Conference* 354–372 (PMLR, 2021).
29. Xie, Q., Luo, Z., Wang, B. & Ananiadou, S. A survey for biomedical text summarization: From pre-trained to large language models. https://arxiv.org/abs/2304.08763 (2023).
30. See, A., Liu, P. J. & Manning, C. D. Get to the point: summarization with pointer-generator networks. https://arxiv.org/abs/1704.04368 (2017).
31. Ben Abacha, A., Yim, W., Fan, Y. & Lin, T. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (eds Vlachos, A. & Augenstein, I.) 2291–2302 (Association for Computational Linguistics, Dubrovnik, Croatia, 2023).
32. Abacha, A. B., Yim, W., Fan, Y. & Lin, T. An empirical study of clinical note generation from doctor-patient encounters 2291–2302 (2023).
33. OpenAI. GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774 (2023)
34. Bubeck, S. et al. Sparks of artificial general intelligence: Early experiments with gpt-4. https://arxiv.org/abs/2303.12712 (2023).
35. Chen, Y-W. & Hirschberg, J. Exploring robustness in doctor-patient conversation summarization: An analysis of out-of-domain SOAP notes. In *Proceedings of the 6th Clinical Natural Language Processing Workshop* (eds Naumann, T., Ben Abacha, A., Bethard, S., Roberts, K., Bitterman, D.). 1–9 (Association for Computational Linguistics, Mexico City, Mexico, 2024).
36. Van Veen, D. et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* **30**, 1134–1142 (2024).
37. López-Úbeda, P., Martín-Noguerol, T., Díaz-Angulo, C. & Luna, A. Evaluation of large language models performance against humans for summarizing MRI knee radiology reports: A feasibility study. *Int. J. Med. Inform.*. **187**, 105443 (2024).
38. Campillos-Llanos, L. et al. Replace, Paraphrase or fine-tune? Evaluating automatic simplification for medical texts in Spanish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (eds Calzolari N, Kan M-Y, Hoste V, Lenci A, Sakti S, Xue N) 13929–13945 (ELRA and ICCL, Torino, Italia, 2024).
39. Touvron, H. et al. LLaMA: Open and efficient foundation language models. https://doi.org/10.48550/arXiv.2302.13971 (2023). Accesed 4 November 2024.
40. Taori Rohan, G. et al. Hashimoto tatsunori B alpaca: A strong, replicable instruction-following model. https://crfm.stanford.edu/2023/03/13/alpaca.html. Accessed 10 May 2023.
41. Wu, C., Zhang, X., Zhang, Y., Wang, Y. & Xie, W. PMC-LLaMA: Further finetuning LLaMA on medical papers. https://doi.org/10.48550/arXiv.2304.14454 (2023). Accesed 4 November 2024.
42. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. Bertscore: Evaluating text generation with bert. https://arxiv.org/abs/1904.09675 (2019).
43. Yuan, W., Neubig, G. & Liu, P. Bartscore: Evaluating generated text as text generation. *Adv. Neural. Inf. Process. Syst.* **34**, 27263–27277 (2021).
44. Coiera, E. & Fraile-Navarro, D. *AI as an Ecosystem—Ensuring Generative AI Is Safe and Effective* (NEJM AI AIp2400611, 2024).
45. Verma, A. A. et al. Implementing machine learning in medicine. *Cmaj* **193**, E1351–E1357 (2021).
46. Navarro, D. F., Kocaballi, A. B., Dras, M. & Berkovsky, S. Collaboration, not confrontation: Understanding general practitioners' attitudes towards natural language and text automation in clinical practice. *ACM Trans. Comput-Hum Interact.* https://doi.org/10.1145/3569893 (2022).

Author contributions

D.F.N. conducted the conceptualization, investigation, data curation, formal analysis, methodology, project administration, writing of the original draft, and writing review and editing. E.C. performed formal analysis, methodology, and writing review and editing. T.W., Z.T., N.A., A.S., A.C., and A.A.L. handled investigation, data curation, and writing review and editing. M.D. contributed to conceptualization and writing review and editing. S.B. managed investigation, supervision, methodology, conceptualization, formal analysis, and writing review and editing. All authors reviewed the manuscript.

Funding

This study was not funded. David Fraile Navarro was supported by an iMQRES scholarship.

Declarations

Ethics approval and consent to participate

Human Ethics and Consent to Participate declarations: not applicable.

Competing interests

The authors declare no competing interests.
Original data collection Ethics Approval available at: Kocaballi AB, Coiera E, Tong HL, White SJ, Quiroz JC, Rezazadegan F, Willcock S, Laranjo L (2019) A network model of activities in primary care consultations. Journal of the American Medical Informatics Association 26:1074–1082.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-84850-x.

Correspondence and requests for materials should be addressed to D.F.N.