The number of clinician burnouts is increasing and has been linked to a high administrative burden. Automatic speech recognition (ASR) and natural language processing (NLP) techniques may address this issue by creating the possibility of automating clinical documentation with a "digital scribe". We reviewed the current status of the digital scribe in development towards clinical practice and present a scope for future research. We performed a literature search of four scientific databases (Medline, Web of Science, ACL, and Arxiv) and requested several companies that offer digital scribes to provide performance data. We included articles that described the use of models on clinical conversational data, either automatically or manually transcribed, to automate clinical documentation. Of 20 included articles, three described ASR models for clinical conversations. The other 17 articles presented models for entity extraction, classification, or summarization of clinical conversations. Two studies examined the system’s clinical validity and usability, while the other 18 studies only assessed their model’s technical validity on the specific NLP task. One company provided performance data. The most promising models use context-sensitive word embeddings in combination with attention-based neural networks. However, the studies on digital scribes only focus on technical validity, while companies offering digital scribes do not publish information on any of the research phases. Future research should focus on more extensive reporting, iteratively studying technical validity and clinical validity and usability, and investigating the clinical utility of digital scribes.

npj Digital Medicine (2021) 4:57; https://doi.org/10.1038/s41746-021-00432-5
Objective
The purpose of the present study is to perform a scoping review of the literature and contact companies on the current status of digital scribes in healthcare. The specific research questions are:

- Which methods are being used to develop (part of) a digital scribe? (Suitability)
- How accurate are these methods? (Technical validity)
- Have any of these methods been evaluated in clinical practice? (Clinical validity and usability, clinical utility)

METHODS
Data search
We performed a scoping review based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) statement. We searched Medline, Web of Science, Arxiv, and ACL for all relevant articles until December 25, 2020. Furthermore, we scanned reference lists of relevant publications for additional articles. Search terms included terms describing the setting (clinical conversations) in combination with relevant methods (NLP, ASR) and usage of the output (clinical documentation). We also included "digital scribe" and "automated scribe" as search terms because these incorporate the setting, method, and goal. The full search queries can be found in Supplementary Table 1.

Besides, we aimed to include real-world data on existing digital scribes to bridge the gap between research and practice. Quiroz et al. provided a list of active companies in the digital scribe space: Robin Healthcare, DeepScribe, Saykara, Sopris Health, Amazon, Nuance. These companies were requested to provide unpublished performance data for their digital scribe.

Inclusion and exclusion criteria
Our definition of a digital scribe is any system that uses a clinical conversation as input, either as audio or text, and automatically extracts information that can be used to generate an encounter note. We included articles that describe the performance of either ASR or NLP on clinical conversational data. A clinical conversation was defined as a conversation—in real life, over the phone, or via chat—between at least one patient and one healthcare professional. Because ASR and NLP are different fields of expertise and will often be described in separate studies, we chose to include studies that only focused on part of a digital scribe. Studies that described NLP models that were not aimed at creating an encounter note but, for example, extracted information for research purposes, were excluded. Articles written in any language other than English were excluded. Because of the rapidly evolving research field and the time lag for publications, proceedings were included.

Study selection
Two reviewers (M.M.v.B. and S.A.C.) independently screened all articles on title and abstract, using the inclusion and exclusion criteria. The selected articles were assessed for eligibility by reading the full text.

Data extraction and synthesis
The first reviewer extracted information from the included articles and the unpublished data provided by companies. The second reviewer verified the extracted information. The following aspects were extracted and assessed:

1. Setting and research phase
2. ASR models and performance
3. NLP tasks, models, and performance

RESULTS
Study selection
Our search resulted in 2348 articles. After screening the titles and abstracts of these articles, we assessed 144 full-text articles for eligibility. We included 20 articles for our analysis (Fig. 1 and Supplementary Table 2). Of these, ten were conference proceedings, seven were workshop proceedings, and two were journal articles, and three were Arxiv preprints.

Of the six contacted companies, DeepScribe was the only one to provide unpublished data on their digital scribe system's performance. We were unable to obtain performance data from other companies.

Setting and research phase
Although all 20 studies aimed to decrease the administrative burden of clinical documentation in some way, the specific approaches and the setting differed greatly among studies. Three studies focused on improving the ASR for clinical conversations as the first step towards accurately extracting information from them. Eleven studies chose to manually transcribe the conversations and performed NLP tasks on the transcriptions. Five studies used input data representative of the input of an implemented digital scribe (ASR transcripts or chat dialogues).

Settings differed greatly between studies, as most did not define a specific specialty, while others were focused on primary care, home hemodialysis, orthopedic encounters, cardiology, family medicine, internal medicine, and patient-clinician dialogs via a telemedicine platform. Fifteen studies were performed by or in collaboration with a company.

All included studies focused on the technical validity of the digital scribe; only two studies investigated the clinical validity and usability by performing a qualitative evaluation with end-users. None of the studies investigated the clinical utility.

Automatic speech recognition (ASR)
In total, seven of 20 studies used ASR to automate clinical documentation, and one company provided data...
of these, two studies and the company presented a new ASR model\textsuperscript{19,21}, four used ASR to transcribe conversations as input for NLP models\textsuperscript{26,33,37,38}, one presented a model to correct ASR errors\textsuperscript{36}, and one compared the performance of existing ASR systems on clinical conversations\textsuperscript{23} (see Supplementary Table 3).

In all studies, the metric used to evaluate the ASR transcripts was the word error rate (WER, see Box 2). The lowest WER was 14.1\% according to the unpublished data provided by DeepScribe. This ASR system combines Google Video Model\textsuperscript{41}, IBM Watson\textsuperscript{42}, and a custom-made Kaldi model\textsuperscript{43}. The best performing published (as opposed to the unpublished data provided by DeepScribe) ASR system had a WER of 18\%\textsuperscript{19}. Four studies\textsuperscript{23,26,33,36} used existing ASR systems and found WERs between 38\% (IBM Watson) and 65\% (Mozilla DeepSpeech\textsuperscript{44}).

One study\textsuperscript{36} presented a postprocessing model to correct ASR errors. By using an attention-based neural network, WERs were improved from 41 to 35\% (Google Speech-to-Text\textsuperscript{35}) and 36 to 35\% (off-the-shelf open-source model\textsuperscript{46}).

**Natural language processing (NLP) tasks and models**

The NLP tasks that were performed could be split into three categories: entity extraction\textsuperscript{20,25–27,30,32,35,38}, classification\textsuperscript{22,24,30–35}, and summarization\textsuperscript{22,24,28,29,31,37} (see Fig. 2 and Supplementary Table 4). All except one study used word embeddings (see Box 3) as input to their model. This study did not use word embeddings as input but used a clustering model to create 2000 clusters\textsuperscript{24}. The model’s input consisted of the current words’ clusters, the number of words, and the previous words’ clusters.

**Entity extraction**

The eight studies using entity extraction focused on extracting symptoms\textsuperscript{20,25–27,30,32,35,38}, medication regimen\textsuperscript{20,26,27,32,35}, and conditions\textsuperscript{27}. However, the studies differed in the combination of entities and properties they extracted. Several studies examined the possibility of extracting symptoms and identifying whether a symptom was present or not\textsuperscript{20,27,38}, while only one study focused on all the other combinations (i.e., medication dosage, frequency, symptom properties). Almost all studies reported their results as F1 scores (see Box 2). The tasks of extracting the medication, medication dosage, and symptom resulted in the highest F1 scores and thus showed the best performance (see Fig. 3). All studies used neural networks, although the type of neural network differed. Some studies used general neural networks\textsuperscript{22,30,35}, but most used neural network-based sequence models with attention (see Box 3). In the studies that compared different types of models, the neural networks with attention layer achieved higher F1 scores than the neural networks without attention layer (see Fig. 3).

Three studies\textsuperscript{27,32,38} performed an error analysis of which one investigated the symptoms that were incorrectly labeled as “absent”. The authors reported that these symptoms were often discussed in multiple talk-turns. In the other study\textsuperscript{27}, ten human
Box 2: Explanation of metrics

WER: This metric counts the number of substitutions, deletions, and insertions in the automatic transcript, compared to the manual transcript. The lower the WER, the better the performance.

F1 score: The F1 score is the harmonic mean between the precision (or positive predictive value) and the recall (or sensitivity).

ROUGE: This is a score that measures the similarity between the automatic summary and the gold standard summary, in unigrams (ROUGE-1), bigrams (ROUGE-2), or the longest common subsequence (ROUGE-L). The ROUGE-L score considers sentence-level structure, while the ROUGE-1 and ROUGE-2 scores only examine if a uni- or bigram occurs in both the automatic and gold standard summary.

Classification. Six studies performed a type of classification, which varied greatly: in which summary section it belonged, if a sentence was said by the patient or the physician, relevant diagnoses of the patient, if any abnormalities were found in the medical history (see Supplementary Table 4). A greater variety of models was used for classification than for entity extraction, although neural networks were used most often. The classification tasks resulting in the highest F1 scores were the classification of primary diagnosis, utterance type, and entity status (see Supplementary Table 4). In two of these tasks, support vector machines were used.

One study [33] tested their classification model on manually transcribed data and automatically transcribed data. The model performed better on the manually transcribed data, with a difference in F1 score ranging from 0.03 to 0.06, although they did not mention if the difference was significant.

One study [22] assessed possible disparities of their classification model towards disadvantaged groups. They formed 18 disadvantaged and advantaged groups based on gender, ethnicity, socioeconomic status, age, obesity, mental health, and location. In 7 of 90 cases, there was a statistically significant difference in favor of the advantaged group. The main reason for the disparity is a difference in the type of medical visit. For example, “blood” is a strong lexical cue to classify a sentence as important for the “Plan” section of the summary, but this word is said less often in conversations with Asian patients.

Summarization. Six studies [22,24,28,29,31,37] used NLP to summarize the conversation between patient and healthcare professional automatically. Four studies used pointer generator networks to create a hybrid extractive and abstractive summary. One of these studies approached the summarization problem as a machine translation problem, where the transcript has to be “translated” to a summary. This study compared the pointer generator network to three other attention-based models (see Supplementary Table 4).

The other two studies used extractive methods, where the output of the classification or entity extraction models was used to extract the most important utterances from the conversation. The combination of these utterances formed the summary. One of these studies did not compare their summaries to a gold standard; the other study asked physicians to extract the most important utterances as gold standard. The F1 score for the latter study was 0.61.

All studies using pointer generator networks reported their results as ROUGE-scores. However, one study only reported their results as ROUGE-L relative error rate reduction, limiting the comparability with the other studies.

The ROUGE-L scores in the other three studies were 0.42 [31], 0.55 [29], and 0.55 [30]. One study also presented a model that returned summaries with a ROUGE-L of 0.58, but this was based on manually extracting noteworthy utterances. When using the same model with automatically extracted noteworthy utterances, the performance dropped to 0.42.

The best performing model used a pretrained pointer generator network (see Box 3) fine-tuned on medical dialog summarization, with an added penalty for the generator distribution to forced the model to favor copying text from the transcript over generating new text. The other models were: a topic-aware pointer-generator network using embeddings (see Box 3), which takes the topic of the current segment into account when copying or generating the next word; an LSTM architecture with BERT embeddings to extract noteworthy utterances (see Box 3); a combination of a transformer and pointer generator network that creates a summary per summary section (see Box 3). Two studies included physicians to evaluate their summaries. One study examined physicians’ ability to answer questions about patient care based on the automatic summary. They did not find any significant difference in physicians’ answers using the human-made summaries compared to the automatic summaries. Another study asked physicians to rate the amount of relevant information in the summaries. Physicians found that 80% of the summaries included “all” or “most” relevant facts. The study did not specify which parts were deemed relevant or not.
Box 3: Neural network-based sequence models with attention and word embeddings

Attention-based neural networks: These models specifically take the sequence of the words into account, and have an attention layer. This layer acts as a filter, only passing the relevant subset of the input to the next layer.

- Sequence2sequence (seq2seq)\(^{13}\): the seq2seq model uses a bidirectional encoder LSTM to include context, and has an attention mechanism to focus on the relevant parts of the input.
- Span-attribute tagging model (SAT)\(^{19}\): the sat model extracts symptoms and classifies them as present or not. It first identifies the relevant parts of the text and then classifies those relevant parts into symptoms that are or are not present. The relation-span-attribute tagging model (R-SAT) is a variant of the SAT that focuses on relations between attributes.
- Pointer generator network (PGNet)\(^{59}\): PG Nets are based on the seq2seq architecture. The added value of a PGNet is that it has the ability to generate new words or copy words from the text, increasing the summary's accuracy. Word embeddings: Word embeddings are used to numerically represent words in a way that similar words have similar representations. For example, the words “physician,” “clinician,” and “doctor” will have similar representations. There are different types of word embeddings, but the most important distinction for this review is between context-sensitive and context-insensitive embeddings. Context-sensitive embeddings have different representation for words that have multiple meanings. For example, the word “bank” can mean a riverbank, or a financial institution. Some word embeddings, like word2vec\(^{50}\), allow only one representation per word, whereas context-sensitive embeddings like ELMo\(^{61}\) and BERT\(^{62}\) can distinguish the different meanings of the word “bank.”

if the model missed specific information.

DeepScribe did not provide information on the models used for summarization but included how often a summary needed to be adjusted in practice. They report that 77% of their summaries do not need modification by a medical scribe before being sent to the physician. Furthermore, 74% of their summaries do not need modification from a medical scribe or a physician before being accepted as part of the patient’s record, saving time on administrative tasks.

DISCUSSION

This scoping review provides an overview of the current state of the development, validation, and implementation of digital scribes. Although the digital scribe is still in an early research phase, there appears to be a substantial research body testing various techniques in different settings. The first results are promising: state-of-the-art models are trained on vast corpora of clinical conversations. However, these aspects were not reported by any of the studies investigating summarization of the clinical conversation used both extractive and abstractive summarization techniques. However, the extractive techniques resulted in a list of the most important utterances instead of a new, full summary. Therefore, the studies performing abstractive summarization are more interesting to discuss. All four studies used the same model, the pointer generator network\(^{28,29,31,37}\). This network’s advantage, especially with the studies’ additions, makes sure it copies more words than it generates, keeping the summary as close to the conversation as possible. Two studies also included a quality check by physicians, which gives more insight into the possibility of implementation\(^{24,28}\). However, it would have been interesting to include error analyses to investigate the models’ blind spots.

When comparing the different NLP tasks, the diverseness in both tasks and underlying models was large. The classification models focused mainly on extracting metadata, such as relevance or structure induction of an utterance, and used various models ranging from logistic regression to neural networks. The entity extraction models were more homogeneous in models but extracted many different entities, complicating the comparison, whereas the summarization task was mostly uniform, both in models and in metrics. One notable aspect of the NLP tasks overall is the use of word embeddings. Only one study did not use word embeddings, but this was a study from 2006 when context-sensitive word embeddings were not yet available. All the other studies were published after 2019 and used various word embeddings as input. The introduction of context-sensitive word embeddings has been essential for extracting entities and summarizing clinical conversations.

In the entity extraction task, the specific tasks, such as extracting symptoms, led to better performance than more general tasks, such as extracting symptoms and their properties. An explanation for this is the heterogeneity in, for example, symptom properties, which entail the location, severity, duration, and other characteristics of a symptom. These properties can be phrased in various ways, in contrast to medication or frequency, which will be much more homogeneous in phrasing. Therefore, this homogeneity leads to many more annotations per entity, increasing performance.

The same pattern was observed in the models, where the addition of an attention layer increased performance. This finding is in line with previous studies on neural attention\(^{46,49}\), which describe the decrease in neural networks’ performance with increased input length. By adding weights to the input text, the model knows which parts of the text are important for its task. Adding attention not only improves performance; it also decreases the amount of training data needed, which is useful in a field such as healthcare, where gathering large datasets can be challenging.

In the studies performing the entity extraction task, the error analyses showed that often, symptoms, medications, or properties are hard to interpret even by human annotators. This result is in line with the concerns discussed in the introduction, questioning if a model would accurately extract all relevant information from a non-linear, fragmented conversation. However, this takes the concern one step further, namely how the “gold standard” will be determined if there is ambiguity between human annotators. More research is needed to define methods for developing gold standards. Shafraan et al.\(^{57}\) have taken an exciting first step towards such a method by publishing an article about the development of their corpus, including how they dealt with ambiguity and labeling errors.

The studies investigating summarization of the clinical conversation used both extractive and abstractive summarization techniques. However, the extractive techniques resulted in a list of the most important utterances instead of a new, full summary. Therefore, the studies performing abstractive summarization are more interesting to discuss. All four studies used the same model, the pointer generator network\(^{28,29,31,37}\). This network’s advantage, especially with the studies’ additions, makes sure it copies more words than it generates, keeping the summary as close to the conversation as possible. Two studies also included a quality check by physicians, which gives more insight into the possibility of implementation\(^{24,28}\). However, it would have been interesting to include error analyses to investigate the models’ blind spots.

Future work

First of all, we believe it is vital to improve the ASR for clinical conversations further and use them as input for NLP models. A remarkable finding was that most studies used manually transcribed conversations as input to their NLP model. These manual transcripts may outperform automatically transcribed conversations regarding...
data quality, leading to an overestimation of the results. NLP models that require manual transcription may increase administrative burden when implemented in clinical practice.

Secondly, the current body of research is mostly focused on improving the performance of different models. Although some studies performed error analyses and qualitative analyses of the model’s output, most did not. Moreover, most studies did not fully cover the technical validity phase because of insufficient reporting on the setting, data, and situations in which the model succeeded and failed. This information is essential to describe for a model that could potentially be implemented in clinical practice. The proposed models might contain bias or lead to unintended results, as Ferracane and Konan[14,15] show. This study is an inspirational example of how researchers can investigate the strengths and weaknesses of their model. A recent paper by Hernandez-Bousset et al.50 proposes reporting standards for AI in healthcare, which should be the basis for reporting on digital scribes as well.

Although most studies are in an early development phase, including qualitative analyses of the model’s output is necessary to know if the solution researchers or developers are working on is applicable in practice. The lack of implementation following the development of an AI model is common in healthcare51, which can be limited by investigating clinical validity and usability while working on technical validity. A good example is the study by Joshi et al.28, where physicians qualitatively analyze the model’s output. These results lead to new insights for improving technical validity. Studying these two research phases iteratively leads to a solution that is well-suited for clinical practice.

Most of the presented models need to be technically and clinically validated before moving on to the clinical utility phase. However, the companies already offering digital scribes seem to have skipped all four research phases, including clinical utility. We urge these companies to publish data on their digital scribes’ technical validity, clinical validity and usability, and clinical utility. Not only is transparency in the model and its performance crucial for clinical practice, but it also helps the community better understand the models and enables researchers to build on past work52.

The suitability phase falls outside the scope of this review but is nevertheless vital for developing and implementing the digital scribe. One research group has published several studies investigating which parts of a clinical conversation are relevant for creating a summary and how physicians see the potential role of a digital scribe53,54. These studies should be the starting point for researchers and developers working on a digital scribe.

Strengths and limitations
The current work is the first effort to review all available literature on developing a digital scribe. We believe our search strategy was complete, leading to a comprehensive and focused scope of the digital scribe’s current research body. By adding the company’s data, we create a broader overview than just the digital scribe’s scientific status. However, this data is unpublished, which means we have to trust the company in providing us with legitimate data. We hope this review is an encouragement for other companies to study their digital scribes scientifically.

One limitation is the small amount of journal papers included in this review, as opposed to the amount of Arxiv preprints and workshop proceedings. These types of papers are often referred very loosely. However, only including journal papers would not lead to a complete scope of this quickly evolving field.

Contacting various digital scribe companies was a first step towards gaining insight into implemented digital scribes and their performance on the different ASR and NLP tasks. Although only one company replied, we believe it is a valuable addition to this review. It indicates that their implemented digital scribe does not differ significantly in techniques or performance from the included studies’ models while already saving physicians’ time. Nevertheless, it highlights the gap between research and practice. The studies published by companies all describe techniques that are not part of a fully functional digital scribe (yet). However, none of the companies offering digital scribes have published about the technical validity, clinical validity and usability, or clinical utility of their systems.

CONCLUSION
Although the digital scribe field has only recently started to accelerate, the presented techniques achieve promising results. The most promising models use context-sensitive word embeddings in combination with attention-based neural networks. However, the studies on digital scribes only focus on technical validity, while companies offering digital scribes do not publish on any of the research phases. Future research should focus on more extensive reporting, iteratively studying technical validity and clinical validity and usability, and investigating the clinical utility of digital scribes.

DATA Availability
Any data generated or analyzed are included in this article and the Supplementary Information files. Aggregate data analyzed in this study are available from the corresponding author on reasonable request.

Received: 8 September 2020; Accepted: 25 February 2021; Published online: 26 March 2021

REFERENCES
1. Shanafelt, T. D. et al. Changes in burnout and satisfaction with work-life integration in physicians and the general US working population between 2011 and 2017. Mayo Clin. Proc. 94, 1681–1694 (2019).
2. National Academies of Sciences, Engineering, and Medicine. Taking Action Against Clinician Burnout: A Systems Approach to Professional Well-Being (The National Academies Press, 2019).
3. Arndt, B. G. et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann. Fam. Med. 15, 419–426 (2017).
4. Sinsky, C. et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. Ann. Intern. Med. 165, 753–760 (2016).
5. Tai-Seale, M. et al. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. Health Aff. 36, 655–662 (2017).
6. Rao, S. K. et al. The impact of administrative burden on academic physicians. Acad. Med. 92, 237–243 (2017).
7. Pellow, K. D., Baier, R. R. & Gardner, R. L. “It’s like texting at the dinner table”: a qualitative analysis of the impact of electronic health records on patient-physician interaction in hospitals. J. Innov. Health Inform. 24, 216–223 (2017).
8. Robertson, S. L., Robinson, M. D. & Reid, A. Electronic health record effects on work-life balance and burnout within the I3 population collaborative. J. Grad. Med. Educ. 9, 479–484 (2017).
9. Gardner, R. L. et al. Physician stress and burnout: the impact of health information technology. J. Am. Med. Inform. Assoc. 26, 106–114 (2019).
10. Gidwani, R. et al. Impact of scribes on physician satisfaction, patient satisfaction, and charting efficiency: a randomized controlled trial. Ann. Fam. Med. 15, 427–433 (2017).
11. Mishra, P., Kiang, J. C. & Grant, R. W. Association of medical scribes in primary care with physician workflow and patient experience. JAMA Intern. Med. 178, 1467 (2018).
12. Coiera, E., Kocaballi, B., Halamka, J. & Laranjo, L. The digital scribe. Npj Digital Med. 1, 1–5 (2018).
13. Quiroz, J. C. et al. Challenges of developing a digital scribe to reduce clinical documentation burden. Npj Digital Med. 2, 1–6 (2019).
14. Ambient clinical intelligence: the exam of the future has arrived. Nuance Communications (2019). Available at: https://www.nuance.com/healthcare/ambient-clinical-intelligence.html. (Accessed: 18th February 2021).
15. Amazon comprehend medical. Amazon Web Services, Inc (2018). Available at: https://aws.amazon.com/comprehend/medical/. (Accessed: 18th February 2021).
conception, design, and critical revision of the manuscript. S.A.C. contributed to conception, literature search, and data synthesis. I.M.J.K. contributed to design and critical revision of the manuscript. All authors gave their final approval and accepted accountability for all aspects of the work.

COMPETING INTERESTS
The authors declare no competing interests.

ADDITIONAL INFORMATION
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-021-00432-5.

Correspondence and requests for materials should be addressed to M.M.v.B.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021