The AI Doctor Is In: A Survey of Task-Oriented Dialogue Systems for Healthcare Applications

Mina Valizadeh and **Natalie Parde**
Natural Language Processing Laboratory
Department of Computer Science
University of Illinois at Chicago
{mvaliz2, parde}@uic.edu

Abstract

Task-oriented dialogue systems are increasingly prevalent in healthcare settings, and have been characterized by a diverse range of architectures and objectives. Although these systems have been surveyed in the medical community from a non-technical perspective, a systematic review from a rigorous computational perspective has to date remained noticeably absent. As a result, many important implementation details of healthcare-oriented dialogue systems remain limited or underspecified, slowing the pace of innovation in this area. To fill this gap, we investigated an initial pool of 4070 papers from well-known computer science, natural language processing, and artificial intelligence venues, identifying 70 papers discussing the system-level implementation of task-oriented dialogue systems for healthcare applications. We conducted a comprehensive technical review of these papers, and present our key findings including identified gaps and corresponding recommendations.

1 Introduction

Dialogue systems[1] have a daily presence in many individuals' lives, acting as virtual assistants (Hoy, 2018), customer service agents (Xu et al., 2017), or even companions (Zhou et al., 2020). While some systems are designed to conduct unstructured conversations in open domains (*chatbots*), others (*task-oriented dialogue systems*) help users to complete tasks in a specific domain (Jurafsky and Martin, 2009; Qin et al., 2019). Task-oriented dialogue systems can potentially play an important role in health and medical care (Laranjo et al., 2018), and they have been adopted by growing numbers of patients, caregivers, and clinicians (Kearns et al., 2019). Nonetheless, there remains a translational gap (Newman-Griffis et al., 2021) between cutting-edge, foundational work in dialogue systems and prototypical or deployed dialogue agents in healthcare settings. This limits the proliferation of scientific progress to real-world systems, constraining the potential benefits of fundamental research.

We move towards closing this gap by conducting a comprehensive, scientifically rigorous analysis of task-oriented healthcare dialogue systems. Our underlying objectives are to (a) explore how these systems have been employed to date, and (b) map out their characteristics, shortcomings, and subsequent opportunities for follow-up work. Importantly, we seek to address the limitations of prior systematic reviews by extensively investigating the included systems from a computational perspective. Our primary contributions are as follows:

1. We systematically search through 4070 papers from well-known technical venues and identify 70 papers fitting our inclusion criteria.[2]

2. We analyze these systems based on many factors, including system objective, language, architecture, modality, device type, and evaluation paradigm, among others.

3. We identify common limitations across systems, including an incomplete exploration of architecture, replicability concerns, ethical and privacy issues, and minimal investigation of usability or engagement. We offer practical suggestions for addressing these as an on-ramp for future work.

In the long term, we hope that the gaps and opportunities identified in this survey can stimulate more rapid advances in the design of task-oriented healthcare dialogue systems. We also hope that the survey provides a useful starting point and synthesis of prior work for NLP researchers and practi-

[1] We follow an inclusive definition of *dialogue systems*, encompassing any intelligent systems designed to converse with humans via natural language.

[2] A full listing of these papers is provided in the appendix.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6638 - 6660

tioners entering this critical yet surprisingly understudied application domain.

2 Related Work

Dialogue systems in healthcare have been the focus of several recent surveys conducted by the medical and clinical communities (Vaidyam et al., 2019; Laranjo et al., 2018; Kearns et al., 2019). These surveys have investigated the real-world utilization of deployed systems, rather than examining their design and implementation from a technical perspective. In contrast, studies examining these systems through the lens of AI and NLP research and practice have been limited. Zhang et al. (2020) and Chen et al. (2017) presented surveys of recent advances in general-domain task-oriented dialogue systems. Although they provide an excellent holistic portrait of the subfield, they do not delve into aspects of particular interest in healthcare settings (e.g., system objectives doubling as clinical goals), limiting their usefulness for this audience.

Vaidyam et al. (2019), Laranjo et al. (2018), and Kearns et al. (2019) conducted systematic reviews of dialogue systems deployed in mental health (Vaidyam et al., 2019) or general healthcare (Laranjo et al., 2018; Kearns et al., 2019) settings. Vaidyam et al. (2019) examined 10 articles, and Laranjo et al. (2018) and Kearns et al. (2019) examined 17 and 46 articles, respectively. All surveys were written for a medical audience and focused on healthcare issues and impact, covering few articles from AI, NLP, or general computer science venues.

Montenegro et al. (2019) and Tudor Car et al. (2020) recently reviewed 40 and 47 articles, respectively, covering conversational agents in the healthcare domain. These two surveys are the closest to ours, but differ in important ways. First, our focus is on a specific class of conversational agents: task-oriented dialogue systems. The surveys by Montenegro et al. (2019) and Tudor Car et al. (2020) used a wider search breading their ability to provide extensive technical depth. We also reviewed more papers (70 articles), which were then screened using a more thorough taxonomy as part of the analysis. Some aspects that we considered that differ from these prior surveys include the overall dialogue system architecture, the dialogue management architecture, the system evaluation methods, and the dataset(s) used when developing and/or evaluating the system.

Screening Process	**ACM**	**IEEE**	**ACL**	**AAAI**	**Total**
Initial Search	1050	1400	1020	600	4070
Title Screening	151	273	106	55	585
Abstract Screening	32	45	26	8	110
Final Screening	**21**	**31**	**16**	**2**	**70**

Table 1: The number of papers included from each database in each step of the paper screening process.

3 Search Criteria and Screening

We designed search criteria in concert with our goal of filling a translational information gap between fundamental dialogue systems research and applied systems in the healthcare domain. To do so, we retrieved articles from well-respected computer science, AI, and NLP databases and screened them for focus on task-oriented dialogue systems designed for healthcare settings. Our target databases were: (1) ACM,[3] (2) IEEE,[4] (3) the ACL Anthology,[5] and (4) the AAAI Digital Library.[6] ACM and IEEE are large databases of papers from prestigious conferences and journals across many CS fields, including but not limited to robotics, human-computer interaction, data mining, and multimedia systems. The ACL Anthology is the premier database of publications within NLP, hosting papers from major conferences and topic-specific venues (e.g., *SIGDIAL*, organized by the Special Interest Group on Discourse and Dialogue). The AAAI Digital Library hosts papers not only from the *AAAI Conference on Artificial Intelligence*, but also from other AI conferences, *AI Magazine*, and the *Journal of Artificial Intelligence Research*. We applied the following inclusion criteria when identifying papers:

- The main focus must be on the technical design or implementation of a task-oriented dialogue system.
- The system must be designed for health-related applications.
- The article must *not* be dedicated to one specific module of the system's architecture (e.g.,

[3] https://dl.acm.org/
[4] https://ieeexplore.ieee.org/
[5] https://www.aclweb.org/anthology/
[6] https://aaai.org/Library/library.php

the natural language understanding component of a health-related dialogue system).

Although a narrower scope—e.g., developing improved methods for slot-filling—is common when publishing in the dialogue systems community, these papers tend to place more emphasis on technical design irrespective of application context, offering less coverage of the system-level characteristics that are the target of this survey. We followed four steps in our screening process. First (*Initial Search)*, we applied a predefined search query to the databases to populate our initial list of papers. To generate the query, we used the keywords "task-oriented," "dialogue system," "conversational agent," "health," and "healthcare," and synonyms and abbreviations of these keywords. We shortlisted papers using these keywords individually as well as in combination with one another.

Next (*Title Screening*), we performed a preliminary screening through the initial list of papers by reading the titles, keeping those that satisfied the inclusion criteria. Then (*Abstract Screening*), we went through the list of papers remaining after the title screening and read the abstracts, keeping those that satisfied the inclusion criteria. Lastly (*Final Screening*), we read the body of the papers remaining after the abstract screening and kept those that satisfied the inclusion criteria.

These funnel filtering processes were conducted by a computer science graduate student (a fluent L2 English speaker) using predefined search and screening guidelines. Questions or uncertainties regarding a paper's compliance with inclusion criteria were forwarded along to the senior project lead (a computer science professor and fluent L1 English speaker with expertise in NLP) and final consensus was reached via discussion among the two parties. We detail the number of papers remaining after each screening step in Table 1. Overall, this screening process combined with our subsequent surveying methods spanned eight months, covering papers published prior to January 2021.

In total, 70 papers (21 from ACM, 31 from IEEE, 16 from ACL, and 2 from AAAI[7]) satisfied the inclusion criteria. We survey papers meeting our inclusion criteria according to a wide range of parameters, and present our findings in the following subsections, grouped into thematic categories: ontology (§4), system architecture (§5), system design (§6), dataset (§7), and system evaluation (§8).

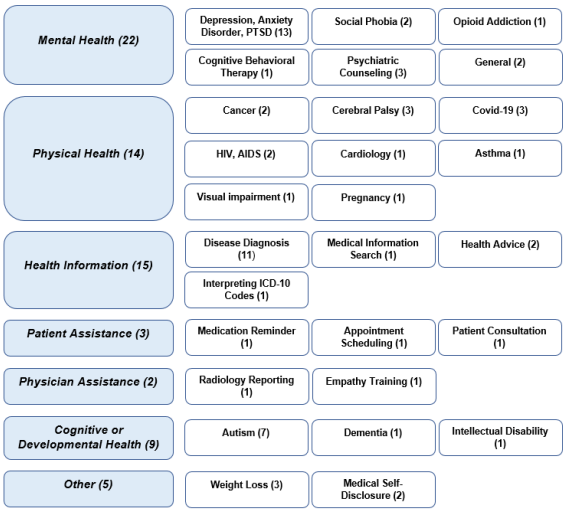

Figure 1: Research domains and corresponding subcategories for the included papers. Parentheses indicate the number of papers belonging to the (sub)category.

4 Ontology

We map each paper to its domain of research (§4.1), system objective (§4.2), target audience (§4.3), and language (§4.4), and present our findings.

4.1 Domain of Research

Task-oriented dialogue systems can potentially impact many facets of healthcare in society (Bickmore and Giorgino, 2004). We define a *domain of research* as the healthcare area in which the system operates. We identify both broad domains and more specific subcategories thereof based on the systems surveyed, outlined in Figure 1. Broad domain categories include *mental health*, *physical health*, *health information*, *patient assistance*, *physician assistance*, *cognitive or developmental health*, and *other* (comprising subcategories not easily classifiable to one of the broader domains).

Systems in the *mental health* domain supported individuals with mental or psychological health conditions, and systems in the *cognitive or developmental health* domain were a close analogue for individuals with conditions impacting memory, executive, or other cognitive function. Systems in the *physical health* domain were targeted towards individuals with specific physical health concerns, including infectious (e.g., Covid-19), non-infectious (e.g., cancer), and temporary (e.g., preg-

[7]Papers about task-oriented dialogue systems published at AAAI often focus on one specific component of the system from a technical perspective, rather than proposing a conversational agent as a whole. Therefore, only two papers from the AAAI Digital Library satisfied the inclusion criteria.

System Objective	# Papers
Diagnosis	7
Monitoring	8
Intervention	13
Counseling	5
Assistance	12
Multi-Objective	25

Table 2: Distribution of system objectives across the surveyed papers. Additional details regarding *multi-objective* papers are provided in the appendix.

nancy) conditions. Systems providing *health information* performed general-purpose actions such as offering advice or suggesting disease diagnoses. Finally, systems performing *patient assistance* or *physician assistance* supported specific patient- or physician-focused healthcare tasks. Dialogue systems designed for *mental health*, *physical health*, and *health information* were the most prevalent, covering 51 of the 70 included papers.

4.2 System Objective

Task-oriented dialogue systems define value relative to the goals of a target task. We define the *system objective* as the healthcare task for which a system is designed. Some system objectives may be closely aligned with a single domain, whereas others may occur in many different domains (e.g., *monitoring* mental, physical, or cognitive conditions). Thus, although the domain of research and system objective may frequently correlate, there is not by necessity a direct association.

Included systems were categorized as being designed to: *diagnose* a health condition (e.g., by predicting whether the user suffers from cognitive decline); *monitor* user states (e.g., by tracking their diets or periodically checking their mood); *intervene* by addressing users' health concerns or improving their states (e.g., by teaching children how to map facial expressions to emotions); *counsel* users without providing any direct intervention (e.g., by listening to users' concerns and empathizing with them); or *assist* users by providing information or guidance (e.g., by answering questions from users who are filling out forms). Many systems were also categorized as *multi-objective*, meaning that they were designed for more than one of those goals.

Table 2 shows the number of systems having each objective. Many systems (25/70) were designed for more than one target objective. Among *multi-objective* systems, those that were designed for both diagnosis and assistance had the highest frequency (7/25); we provide additional details regarding these systems in Table 8 of the appendix.

Target Audience	# Papers
Patients	59
Caregivers	3
Patients & Caregivers	2
Clinicians	11

Table 3: Distribution of the target audiences of the systems described in the surveyed papers.

Separately, we also considered the role of *engagement* as an objective of each system. We define this as a goal of engaging target users in interaction, irrespective of underlying health goals. Engagement may be of particular interest in healthcare settings since it can be critical in encouraging adoption or adherence with respect to healthcare outcomes (Montenegro et al., 2019). Surprisingly, almost 60% of the papers (41 of the 70 surveyed) did not mention any goals pertaining to engaging users in more interactions.

4.3 Target Audience

The final consumers of healthcare systems often fall into three groups: *patients*, *caregivers*, and *clinicians*. Table 3 shows the number of systems surveyed that focus on each category. We find that out of 70 task-oriented dialogue systems, 59 are designed specifically for patients.

4.4 Language

Most general-domain dialogue systems research has been conducted in English and other high-resource languages (Artetxe et al., 2020). Expanding language diversity may extend the benefits of health-related dialogue systems more globally. As shown in Figure 2, among the systems included in our review a majority (56%) are designed for English speakers. Encouragingly, several of the included systems did focus on lower-resource languages, including Telugu (Duggenpudi et al., 2019), Bengali (Rahman et al., 2019), and Setswana (Grover et al., 2009).

5 System Architecture

We investigate both the general architecture of the system (§5.1), and if applicable, the dialogue man-

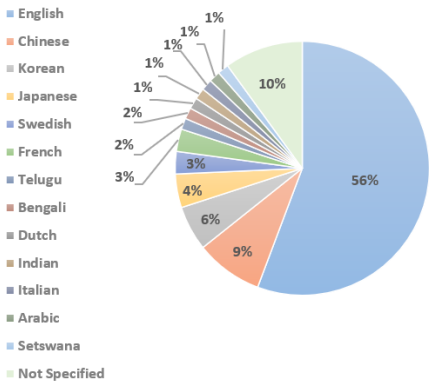

Figure 2: Language diversity across the surveyed systems. A small percentage (10%) of papers do not specify the system's language.

System Architecture	# Papers
Pipeline	58
End-to-End	2
Not Specified	10

Table 4: Distribution of papers describing systems with pipeline or end-to-end architectures, or that do not specify the architecture.

agement architecture specifically (§5.2).

5.1 General Architecture

Task-oriented dialogue systems are generally designed using *pipeline* or *end-to-end* architectures. Pipeline architectures typically consist of separate components for natural language understanding, dialogue state tracking, dialogue policy, and natural language generation. The ensemble of the dialogue state tracker and dialogue policy is the *dialogue manager* (Chen et al., 2017). End-to-end architectures train a single model to produce output for a given input, often interacting with structured external databases and requiring extensive training data (Chen et al., 2017). As shown in Table 4, only 2.85% of papers (2 of the 70 surveyed) implemented an end-to-end system; this is unsurprising given the limited training data available in most healthcare domains. We also found that 14% (10 papers) did not directly specify the architecture of their developed system.

5.2 Dialogue Management Architecture

Unlike other pipeline components that impact user experience and engagement but not fundamental decision-making, the dialogue manager is central to overall functionality (Zhao et al., 2019); thus, we afford it special attention. In *rule-based* approaches, the system interacts with users based on a predefined set of rules, with success conditioned upon coverage of all relevant cases (Siangchin and Samanchuen, 2019). *Intent-based* approaches seek to extract the user's intention from the dialogue, and then perform the relevant action (Jurafsky and Martin, 2009). In *hybrid* dialogue management architectures, the system leverages a combination of rule-based and intent-based approaches, and finally *corpus-based* approaches mine the dialogues of human-human conversations and produce responses using retrieval methods or generative methods (Jurafsky and Martin, 2009). As shown in Table 5, among papers reporting on dialogue management architecture, we observe a fairly even mix of rule-based, intent-based, and hybrid architectures.

Dialogue Management Architecture	# Papers
Rule-based	17
Intent-based	20
Hybrid Architecture	21
Corpus-based	0

Table 5: Distribution of dialogue management architectures across the surveyed papers. This table does not include papers describing end-to-end architectures ($n = 2$) or for which system architecture was not specified ($n = 10$).

6 System Design

6.1 Modality

Modality, the channel through which information is exchanged between a computer and a human (Karray et al., 2008), can play an important role in dialogue quality and user satisfaction (Bilici et al., 2000). *Unimodal* systems use a single modality for information exchange, whereas *multimodal* systems use multiple modalities (Karray et al., 2008). Systems reviewed in this survey operated using one or more of several modalities. In *text-based* or *spoken* interaction, users interact with the system by typing or speaking, respectively. In interaction via *graphical user interface (GUI)*, users interact with the system through the use of visual elements.

In general, multimodal dialogue systems can be flexible and robust, but especially challenging to implement in the medical domain (Sonntag et al., 2009). We find that 49 papers describe unimodal systems and 21 describe multimodal systems. Ta-

Unimodal		Multimodal	
Category	**# Papers**	**Category**	**# Papers**
Text	23	Spoken + Text	14
Spoken	25	Spoken + GUI	4
GUI	1	Text + GUI	3

Table 6: Distribution of modality type across the unimodal (49 total, left) and multimodal (21 total, right) systems surveyed.

Evaluation Type	**# Papers**
Human Evaluation	28
Automated Evaluation	7
Human & Automated Evaluation	9
Not Specified	26

Table 7: Distribution of evaluation methods across the surveyed papers.

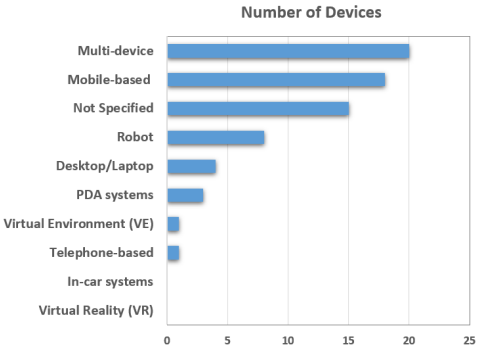

Figure 3: Distribution of device type across the surveyed papers.

ble 6 provides more details regarding their distribution across modalities.

6.2 Device

Dialogue systems may facilitate interaction using a variety of devices (Arora et al., 2013), ranging from telephones (Garvey and Sankaranarayanan, 2012) to computers (McTear, 2010) to any other technology that allows interaction (e.g., VR-based avatars (Brinkman et al., 2012b; McTear, 2010)). We categorized the included systems as *mobile*, *telephone*, *desktop/laptop*, *in-car*, *PDA*, *robot*, *virtual environment*, or *virtual reality* (including virtual agents and avatars) systems, considering systems as *multi-device* if they leveraged multiple devices for interaction. As shown in Figure 3, we found that multi-device and mobile-based dialogue systems were most popular. Table 9 in the appendix provides additional details regarding multi-device systems.

7 Dataset

Data is crucial for effective system development (Serban et al., 2015), but many datasets for training dialogue systems are smaller than those used for other NLP tasks (Lowe et al., 2017). This is even more pronounced in the healthcare domain, in part due to the risk of data misuse by others or the lack of data sharing incentives (Lee and Yoon, 2017).

We reviewed each paper for information regarding the data used during system development, focusing on dataset size, availability, and privacy-preserving measures. Only 20 papers provide details about the data used (two papers provided a link to the dataset, and the remaining 18 discussed the dataset size). Unfortunately, the remaining papers did not provide rationale for their lack of data or other replicability information. Our assumption is that often the data contained sensitive information, preventing authors from releasing specific details, but only 19 of the 70 included papers provided information about data-related privacy or ethical considerations. Only 10 mentioned Institutional Review Board (IRB) approval for their dataset and/or task, despite IRB (or equivalent) review being a crucial step towards ensuring that research is conducted ethically and in such a way that protects human subjects to the extent possible (Amdur and Biddle, 1997).

8 System Evaluation

We examined the means through which systems were evaluated both qualitatively and quantitatively (Deriu et al., 2019; Hastie, 2012). We defined *human evaluation*, often implemented in prior work through questionnaires (Grover et al., 2009; Holmes et al., 2019; Parde and Nielsen, 2019; Wang et al., 2020) or direct feedback from real-world users (Deriu et al., 2019), as an evaluation that relies on subjective, first-hand, human user experience. In contrast, *automated evaluation* provides an objective, quantitative measurement of one or more dimensions of the system from a mathematical perspective (Finch and Choi, 2020). Some metrics used for automated evaluation of the reviewed systems include measures of task performance (Ali et al., 2020) and completion rates (Holmes et al., 2019), response correctness (Rosruen and Samanchuen, 2018), and response time (Grover et al., 2009).

In Table 7, we observe that nearly half of the papers conducted human evaluations; however, a large percentage (37%) also did not discuss evaluation at all. We further analyzed papers conducting human evaluations and found that they included an average of 26 (mode = 12) participants. More details regarding the human and automated evaluations are provided in Tables 10, 11, and 12 of the appendix. In a follow-up analysis of *system usability*, defined as the degree to which users are able to engage with a system safely, effectively, efficiently, and enjoyably (Lee et al., 2019), we observed that 33 papers explicitly evaluated the usability of their system.

9 Discussion

We identify common limitations across many surveyed systems, accompanied by recommendations for addressing them in future work.

9.1 Incomplete Exploration of System Design

We observed little system-level architectural diversity across the surveyed systems, with most (83%) having a pipeline architecture. This architectural homogeneity limits our understanding of good design practice within this domain. Recent studies demonstrate that end-to-end architectures for task-oriented dialogue systems could compete with pipeline architectures given sufficient high-quality data (Hosseini-Asl et al., 2020; Ham et al., 2020; Bordes et al., 2017; Wen et al., 2016). However, the external knowledge sources often leveraged in end-to-end systems are notoriously complex in many healthcare sub-domains (Campillos-Llanos et al., 2020). Additionally, for healthcare applications interpretability is highly desired (Ham et al., 2020), but explanations are often obfuscated in end-to-end systems (Ham et al., 2020; Wen et al., 2016). Finally, users of these systems may seek guidance on sensitive topics, which can exacerbate privacy concerns (Xu et al., 2021). Any system trained on large, weakly curated datasets may also learn unpleasant behaviors and amplify biases in the training data, in turn producing harmful consequences (Dinan et al., 2021; Bender et al., 2021). We recommend further experimentation with architectural design, in parallel with work towards developing high-quality healthcare dialogue datasets, which to date remain scarce (Farzana et al., 2020).

We noticed that a considerable number of the systems (33%) allowed only text-based interaction. However, it is well-established that individuals from certain demographic groups are more comfortable conversing with dialogue systems via speech (Tudor Car et al., 2020). Text-based systems may also be more likely to violate privacy considerations (Tudor Car et al., 2020). Thus, we recommend that researchers engage in further exploration of multimodal or spoken dialogue systems when applicable and appropriate.

Many of the surveyed systems were also implemented on mobile phones. Although an advantage of mobile-based systems is that they are readily available using a technology familiar to most users, Lee et al. (2018) found that users significantly reduced their usage over time when engaging long-term with mobile health applications. Tudor Car et al. (2020) suggest that one way to overcome this limitation in mobile-based systems is by directly embedding them in applications or platforms with which users already engage habitually (e.g., Facebook Messenger). This more ambient dissemination approach may facilitate easier and more lasting integration of system use in individuals' daily lives.

Finally, we identified that most systems (84%) target only patients, with research on systems targeted towards clinicians and caregivers remaining limited. We recommend further exploration of systems targeted towards these critical audiences. This may offer broad, high-impact support in understanding, diagnosing, and treating patients' health issues (Valizadeh et al., 2021; Kaelin et al., 2021).

9.2 Replicability Concerns

Data accessibility restrictions reduce the capacity of public health research (Strongman et al., 2019), and these limitations may be partially responsible for the imbalance of pipeline versus end-to-end architectures (§9.1). Only a small percentage of papers surveyed (29%) ventured to discuss the quantity or characteristics of the data used during system development in any way. A lack of data transparency hinders scientific progress and severely impedes replicability. We call upon researchers to publish data when permissible by governing protocol, and descriptive statistics to the extent allowable when circumstances prevent data release. We also view the development of high-quality, publicly available datasets as an important frontier in translational dialogue systems research (§9.1).

Many of the surveyed papers also lack important implementation details, such as evaluation meth-

ods (34%). This prevents the research community from replicating developed systems and generalizing study findings more broadly (Walker et al., 2018). Well-established guidelines exist and are being increasingly enforced within the NLP community to prevent reproducibility issues (Dodge et al., 2019). The disregard of reproducibility best practices observed with many healthcare dialogue systems may be partially attributed to the most common target venues for this work, which may place less emphasis on replication. This validates a central motivator for publishing this survey—without adequate inclusion of target domain *and* technical stakeholders in interdisciplinary, translational research, progress will remain constrained. We strongly urge researchers in this domain to provide implementation details in their publications.

9.3 Potential Ethical and Privacy Issues

Real-world medical data facilitates the development of high-quality healthcare applications (Bertino et al., 2005; Di Palo and Parde, 2019; Farzana et al., 2020), but protecting the rights and privacy of contributors to the data is critical for ensuring ethical research conduct (Institute of Medicine, 2009), as is proper treatment of copyright protections. We screened all included papers for coverage of privacy and ethical concerns, and observed that only 27% of the surveyed papers considered participant or patient privacy in the design of their system. Moreover, only 14% of the surveyed papers documented any evidence of Institutional Review Board (or IRB-equivalent) approval.

Research involving healthcare dialogue systems is unquestionably human-centered, and as such the absence of ethical oversight in the design of such systems is a grave concern. Although technical researchers entering this space may be unfamiliar with human subjects research and protocol, we urge all dialogue systems researchers to submit their experimental design and protocol for review by an appropriate external review board. We also ask that researchers consider the potential harms from use or misuse of their systems, following guidelines established by the ACM Code of Ethics.[8]

9.4 Room for Increased Language Diversity

We observed that most systems (56%) targeted English speakers. Developing multilingual dialogue systems or systems for speakers of low-resource languages brings up various challenges (López-Cózar Delgado and Araki, 2005), but solving this problem could have have tremendous benefit for individuals in non-English speaking communities with minimal or unreliable healthcare access. The systems developed by Duggenpudi et al. (2019), Rahman et al. (2019), and Grover et al. (2009) provide case examples for how such systems may be implemented. We also note that while troubling, a 56% share of systems targeted towards English speakers is consistent with linguistic homogeneity in the field in general, and actually slightly low relative to many other NLP tasks (Mielke, 2016; Bender, 2009). Healthcare dialogue systems may on some level offer a case example for how applications originally designed for high-resource (i.e., English-language) settings can be adapted and re-engineered to provide better coverage of the diverse, real-world potential user base.

9.5 Minimal Investigation of Usability or User Engagement

Finally, more than 50% (37/70) of the included papers did not evaluate system usability or general user experience. Usability testing can improve productivity and safeguard against errors (Rogers et al., 2005), both of which are critical in healthcare tasks. Therefore, we urge the research community to consider and assess usability when designing for this domain. The systems among those surveyed that do this already (e.g., those developed by Wang et al. (2020), Lee et al. (2020b), Wei et al. (2018), or Demasi et al. (2020)) provide case examples for how it might be done.

Almost 60% of the surveyed systems were not explicitly designed to engage users, despite this being a common objective in the general domain (Ghazarian et al., 2019). Healthcare dialogue systems may stand to benefit particularly well from such measures (Parde, 2018), since patient engagement is predictive of adoption and adherence to healthcare outcomes (Montenegro et al., 2019). To increase user satisfaction and system performance, we recommend that the research community more purposefully consider engagement when designing their healthcare-oriented dialogue systems.

10 Conclusion

In this work, we conducted a systematic technical survey of task-oriented dialogue systems used for health-related purposes, providing much-needed

[8] https://www.acm.org/code-of-ethics

analyses from a computational perspective and narrowing the translational gap between basic and applied dialogue systems research. We comprehensively searched through 4070 papers in computer science, NLP, and AI databases, finding 70 papers that satisfied our inclusion criteria. We analyzed these papers based on numerous technical factors including the domain of research, system objective, target audience, language, system architecture, system design, training dataset, and evaluation methods. Following this, we identified and summarized gaps in this existing body of work, including an incomplete exploration of system design, replicability concerns, potential ethical and privacy issues, room for increased language diversity, and minimal investigation of usability or user engagement. Finally, we presented evidence-based recommendations stemming from our findings as a launching point for future work. It is our hope that interested researchers find the information provided in this survey to be a unique and helpful resource for developing task-oriented dialogue systems for healthcare applications.

11 Ethical Considerations

Beyond the concrete changes suggested during the discussion, it is important to consider the broader ethical implications of task-oriented dialogue systems in healthcare settings. Although the goal of such systems may not be to replace human healthcare providers, it is likely that deployed systems would support clinicians, defraying workload for overburdened individuals. In doing so, these systems may have significant impact on healthcare decision-making. Machines are imperfect, and thus a possible harm is that these systems may misinterpret user input or make incorrect predictions—a mistake that in high-stakes healthcare settings could prove detrimental or even dangerous. Researchers and developers should be cognizant of possible harms stemming from the use and misuse of task-oriented dialogue systems for healthcare settings, and should implement both automated (e.g., strict thresholds for diagnostic suggestions) and human (e.g., training to ensure staff awareness of potential system fallibilities) safeguards.

Moreover, a potential benefit of these systems is their potential to meaningfully and beneficially extend healthcare access to underserved populations. As such, it is important to ensure that automated systems do not fall prey to the same biases often observed among human healthcare providers (FitzGerald and Hurst, 2017). Systems trained to perform healthcare tasks using datasets that are not representative of the target population may exhibit poorer performance with users who already experience marginalization or are otherwise vulnerable, impeding or even reversing benefits. We call upon researchers to examine, debias, and curate their training data such that task-oriented dialogue systems for healthcare applications elevate, rather than diminish, outcomes for the historically underserved users which they are best poised to benefit.

12 Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 2125411, and by a start-up grant from the University of Illinois at Chicago. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank the anonymous reviewers for their insightful suggestions, which further strengthened this work.

References

Parham Aarabi. 2013. Virtual cardiologist — a conversational system for medical diagnosis. In *2013 26th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1–4.

Yuna Ahn, Yilin Zhang, Yujin Park, and Joonhwan Lee. 2020. A chatbot solution to chat app problems: Envisioning a chatbot counseling system for teenage victims of online sexual exploitation. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–7, New York, NY, USA. Association for Computing Machinery.

Mohammad Rafayet Ali, Seyedeh Zahra Razavi, Raina Langevin, Abdullah Al Mamun, Benjamin Kane, Reza Rawassizadeh, Lenhart K. Schubert, and Ehsan Hoque. 2020. A virtual conversational agent for teens with autism spectrum disorder: Experimental results and design lessons. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20, New York, NY, USA. Association for Computing Machinery.

Mohammad Rafayet Ali, Taylan Sen, Benjamin Kane, Shagun Bose, Thomas Carroll, Ronald Epstein, Lenhart K. Schubert, and Ehsan Hoque. 2021. Novel computational linguistic measures, dialogue system and the development of sophie: Standardized online patient for healthcare interaction education. *IEEE Transactions on Affective Computing*, pages 1–1.

Robert J. Amdur and Chuck Biddle. 1997. Institutional Review Board Approval and Publication of Human Research Results. *JAMA*, 277(11):909–914.

Masahiro Araki, Kana Shibahara, and Yuko Mizukami. 2011. Spoken dialogue system for learning braille. In *2011 IEEE 35th Annual Computer Software and Applications Conference*, pages 152–156.

Suket Arora, Kamaljeet Batra, and Sarabjit Singh. 2013. Dialogue system: A brief review. *CoRR*, abs/1306.4134.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388, Online. Association for Computational Linguistics.

Lekha Athota, Vinod Kumar Shukla, Nitin Pandey, and Ajay Rana. 2020. Chatbot for healthcare system using artificial intelligence. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pages 619–622.

Saminda Sundeepa Balasuriya, Laurianne Sitbon, Andrew A. Bayor, Maria Hoogstrate, and Margot Brereton. 2018. Use of voice activated interfaces by people with intellectual disability. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction*, OzCHI '18, page 102–112, New York, NY, USA. Association for Computing Machinery.

R. V. Belfin, A. J. Shobana, Megha Manilal, Ashly Ann Mathew, and Blessy Babu. 2019. A graph based chatbot for cancer patients. In *2019 5th International Conference on Advanced Computing Communication Systems (ICACCS)*, pages 717–721.

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

E. Bertino, B.C. Ooi, Y. Yang, and R.H. Deng. 2005. Privacy and ownership preserving of outsourced medical data. In *21st International Conference on Data Engineering (ICDE'05)*, pages 521–532.

Timothy Bickmore and Toni Giorgino. 2004. Some novel aspects of health communication from a dialogue systems perspective. *AAAI Fall Symposium - Technical Report*.

Vildan Bilici, Emiel Krahmer, Saskia te Riele, and Raymond Veldhuis. 2000. Preferred modalities in dialogue systems. In *Sixth International Conference on Spoken Language Processing*.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog.

Willem-Paul Brinkman, Dwi Hartanto, Ni Kang, Daniel de Vliegher, Isabel L. Kampmann, Nexhmedin Morina, Paul G.M. Emmelkamp, and Mark Neerincx. 2012a. A virtual reality dialogue system for the treatment of social phobia. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '12, page 1099–1102, New York, NY, USA. Association for Computing Machinery.

Willem-Paul Brinkman, Dwi Hartanto, Ni Kang, Daniel Vliegher, Isabel Kampmann, Nexhmedin Morina, Paul Emmelkamp, and Mark Neerincx. 2012b. A virtual reality dialogue system for the treatment of social phobia. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 1099–1102.

Jacqueline Brixey, Rens Hoegen, Wei Lan, Joshua Rusow, Karan Singla, Xusen Yin, Ron Artstein, and Anton Leuski. 2017. SHIHbot: A Facebook chatbot for sexual health information on HIV/AIDS. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 370–373, Saarbrücken, Germany. Association for Computational Linguistics.

Leonardo Campillos Llanos, Dhouha Bouamor, Éric Bilinski, Anne-Laure Ligozat, Pierre Zweigenbaum, and Sophie Rosset. 2015. Description of the Patient-Genesys dialogue system. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 438–440, Prague, Czech Republic. Association for Computational Linguistics.

Leonardo Campillos-Llanos, Catherine Thomas, Éric Bilinski, Pierre Zweigenbaum, and Sophie Rosset. 2020. Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation. *Natural Language Engineering*, 26(2):183–220.

Bo-Wei Chen, Po-Yi Shih, Karunanithi Bharanitharan, Po-Chuan Lin, Jhing-Fa Wang, and Chia-Ming Chen. 2013. Customizable cloud-healthcare dialogue system based on lvcsr with prosodic-contextual post-processing. In *2013 1st International Conference on Orange Technologies (ICOT)*, pages 246–249.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *CoRR*, abs/1711.01731.

Ching-Hua Chuan and Susan Morgan. 2021. Creating and evaluating chatbots as eligibility assistants

for clinical trials: An active deep learning approach towards user-centered classification. *ACM Trans. Comput. Healthcare*, 2(1).

Karl Daher, Jacky Casas, Omar Abou Khaled, and Elena Mugellini. 2020. Empathic chatbot response for medical assistance. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20, New York, NY, USA. Association for Computing Machinery.

Prathyusha Danda, Brij Mohan Lal Srivastava, and Manish Shrivastava. 2016. Vaidya: A spoken dialog system for health domain. In *Proceedings of the 13th International Conference on Natural Language Processing*, pages 161–166, Varanasi, India. NLP Association of India.

Johan Oswin De Nieva, Jose Andres Joaquin, Chaste Bernard Tan, Ruzel Khyvin Marc Te, and Ethel Ong. 2020. Investigating students' use of a mental health chatbot to alleviate academic stress. In *6th International ACM In-Cooperation HCI and UX Conference*, CHIuXiD '20, page 1–10, New York, NY, USA. Association for Computing Machinery.

Orianna Demasi, Yu Li, and Zhou Yu. 2020. A multi-persona chatbot for hotline counselor training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3623–3636, Online. Association for Computational Linguistics.

Jan Deriu, Álvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2019. Survey on evaluation methods for dialogue systems. *CoRR*, abs/1905.04071.

David DeVault, Kallirroi Georgila, Ron Artstein, Fabrizio Morbini, David Traum, Stefan Scherer, Albert Skip Rizzo, and Louis-Philippe Morency. 2013. Verbal indicators of psychological distress in interactive dialogue with a virtual human. In *Proceedings of the SIGDIAL 2013 Conference*, pages 193–202, Metz, France. Association for Computational Linguistics.

Alessandro Di Nuovo, Josh Bamforth, Daniela Conti, Karen Sage, Rachel Ibbotson, Judy Clegg, Anna Westaway, and Karen Arnold. 2020. An explorative study on robotics for supporting children with autism spectrum disorder during clinical procedures. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, page 189–191, New York, NY, USA. Association for Computing Machinery.

Flavio Di Palo and Natalie Parde. 2019. Enriching neural models with targeted features for dementia detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 302–308, Florence, Italy. Association for Computational Linguistics.

Emily Dinan, Gavin Abercrombie, A. Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. Anticipating safety issues in e2e conversational ai: Framework and tooling.

Francesca Dino, Rohola Zandie, Hojjat Abdollahi, Sarah Schoeder, and Mohammad H. Mahoor. 2019. Delivering cognitive behavioral therapy using a conversational social robot. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2089–2095.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Suma Reddy Duggenpudi, Kusampudi Siva Subrahamanyam Varma, and Radhika Mamidi. 2019. Samvaadhana: A Telugu dialogue system in hospital domain. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 234–242, Hong Kong, China. Association for Computational Linguistics.

Wilmer Stalin Erazo, Germán Patricio Guerrero, Carlos Carrión Betancourt, and Iván Sánchez Salazar. 2020. Chatbot implementation to collect data on possible covid-19 cases and release the pressure on the primary health care system. In *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0302–0307.

Ahmed Fadhil and Ahmed Ghassan Tawfiq AbuRa'ed. 2019. Ollobot - towards a text-based arabic health conversational agent: Evaluation and results. In *RANLP*.

Shahla Farzana, Mina Valizadeh, and Natalie Parde. 2020. Modeling dialogue in conversational cognitive health screening interviews. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1167–1177, Marseille, France. European Language Resources Association.

Sarah E. Finch and Jinho D. Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting. Association for Computational Linguistics.

Chloë FitzGerald and Samia Hurst. 2017. Implicit bias in healthcare professionals: a systematic review. *BMC medical ethics*, 18(1):1–18.

Floyd Garvey and Suresh Sankaranarayanan. 2012. Intelligent agent based flight search and booking system. *International Journal of Advanced Research in Artificial Intelligence*, 1(4).

Sarik Ghazarian, Ralph M. Weischedel, Aram Galstyan, and Nanyun Peng. 2019. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. *CoRR*, abs/1911.01456.

Nancy Green, William Lawton, and Boyd Davis. 2004. An assistive conversation skills training system for caregivers of persons with alzheimer's disease. In *Proceedings of the AAAI 2004 Fall Symposium on Dialogue Systems for Health Communication*.

Aditi Sharma Grover, Madelaine Plauché, Etienne Barnard, and Christiaan Kuun. 2009. Hiv health information access using spoken dialogue systems: Touchtone vs. speech. In *Proceedings of the 3rd International Conference on Information and Communication Technologies and Development*, ICTD'09, page 95–107. IEEE Press.

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 583–592, Online. Association for Computational Linguistics.

Helen Hastie. 2012. Metrics and evaluation of spoken dialogue systems. In *Data-driven methods for adaptive spoken dialogue systems*, pages 131–150. Springer.

Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael Mctear. 2019. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In *Proceedings of the 31st European Conference on Cognitive Ergonomics*, ECCE 2019, page 207–214, New York, NY, USA. Association for Computing Machinery.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *CoRR*, abs/2005.00796.

Matthew B. Hoy. 2018. Alexa, siri, cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly*, 37(1):81–88. PMID: 29327988.

Chin-Yuan Huang, Ming-Chin Yang, Chin-Yu Huang, Yu-Jui Chen, Meng-Lin Wu, and Kai-Wen Chen. 2018. A chatbot-supported smart wireless interactive healthcare system for weight control and health promotion. In *2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 1791–1795.

Tae-Ho Hwang, JuHui Lee, Se-Min Hyun, and KangYoon Lee. 2020. Implementation of interactive healthcare advisor model using chatbot and visualization. In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 452–455.

Koji Inoue, Pierrick Milhorat, Divesh Lala, Tianyu Zhao, and Tatsuya Kawahara. 2016. Talking with ERICA, an autonomous android. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 212–215, Los Angeles. Association for Computational Linguistics.

Institute of Medicine. 2009. *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. The National Academies Press, Washington, DC.

Hifza Javed, Myounghoon Jeon, Ayanna Howard, and Chung Hyuk Park. 2018. Robot-assisted socio-emotional intervention framework for children with autism spectrum disorder. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, page 131–132, New York, NY, USA. Association for Computing Machinery.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., USA.

Dipesh Kadariya, Revathy Venkataramanan, Hong Yung Yip, Maninder Kalra, Krishnaprasad Thirunarayanan, and Amit Sheth. 2019. kbot: Knowledge-enabled personalized chatbot for asthma self-management. In *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 138–143.

Vera C Kaelin, Mina Valizadeh, Zurisadai Salgado, Natalie Parde, and Mary A Khetani. 2021. Artificial intelligence in rehabilitation targeting the participation of children and youth with disabilities: Scoping review. *J Med Internet Res*, 23(11):e25745.

Takeshi Kamita, Atsuko Matsumoto, Boyu Sun, and Tomoo Inoue. 2020. Promotion of continuous use of a self-guided mental healthcare system by a chatbot. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, CSCW '20 Companion, page 293–298, New York, NY, USA. Association for Computing Machinery.

B. Amir H. Kargar and Mohammad H. Mahoor. 2017. A pilot study on the ebear socially assistive robot: Implication for interacting with elderly people with moderate depression. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pages 756–762.

Fakhri Karray, Milad Alemzadeh, Jamil Saleh, and Mo Nours Arab. 2008. Human-computer interaction: Overview on state of the art. *International Journal on Smart Sensing and Intelligent Systems*, 1:137–159.

William Kearns, Nai-Ching Chi, Yong Choi, Shih-Yin Lin, Hilaire Thompson, and George Demiris. 2019. A systematic review of health dialog systems. *Methods of Information in Medicine*, 58:179–193.

Liliana Laranjo, Adam Dunn, Huong Ly Tong, A. Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Lau, and Enrico Coiera. 2018. Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*, 0.

Choong Lee and Hyung-Jin Yoon. 2017. Medical big data: promise and challenges. *Kidney Research and Clinical Practice*, 36:3–11.

Dongkeon Lee, Kyo-Joong Oh, and Ho-Jin Choi. 2017. The chatbot feels you - a counseling service using emotional response generation. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 437–440.

Ju Yeon Lee, Ju Young Kim, Seung Ju You, You Soo Kim, Hye Yeon Koo, Jeong Hyun Kim, Sohye Kim, Jung Ha Park, Jong Soo Han, Siye Kil, Hyerim Kim, Ye Seul Yang, and Kyung Min Lee. 2019. Development and usability of a life-logging behavior monitoring application for obese patients. *Journal of Obesity and Metabolic Syndrome*, 28(3):194–202. Publisher Copyright: Copyright © 2019 Korean Society for the Study of Obesity.

Kyunghee Lee, Hyeyon Kwon, Byungtae Lee, Guna Lee, Jae Ho Lee, Yu Rang Park, and Soo-Yong Shin. 2018. Effect of self-monitoring on long-term patient engagement with mobile health applications. *PLOS ONE*, 13(7):1–12.

Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. 2020a. Designing a chatbot as a mediator for promoting deep self-disclosure to a real mental health professional. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).

Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020b. "i hear you, i feel you": Encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA. Association for Computing Machinery.

Peter Ljunglöf, Britt Claesson, Ingrid Mattsson Müller, Stina Ericsson, Cajsa Ottesjö, Alexander Berman, and Fredrik Kronlid. 2011. Lekbot: A talking and playing robot for children with disabilities. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 110–119, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Peter Ljunglöf, Staffan Larsson, Katarina Heimann Mühlenbock, and Gunilla Thunberg. 2009. TRIK: A talking and drawing robot for children with communication disabilities. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pages 275–278, Odense, Denmark. Northern European Association for Language Technology (NEALT).

A. Loisel, N. Chaignaud, and J-Ph. Kotowicz. 2007. Designing a human-computer dialog system for medical information search. In *2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops*, pages 350–353.

Ramón López-Cózar Delgado and Masahiro Araki. 2005. *Spoken, Multilingual and Multimodal Dialogue Systems: Development and Assessment*. Wiley, Chichester, UK.

Ryan Lowe, Nissan Pow, Iulian Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. 2017. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue and Discourse*, 8:31–65.

Raju Maharjan, Per Bækgaard, and Jakob E. Bardram. 2019. "hear me out": Smart speaker based conversational agent to monitor symptoms in mental health. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '19 Adjunct, page 929–933, New York, NY, USA. Association for Computing Machinery.

Rohit Binu Mathew, Sandra Varghese, Sera Elsa Joy, and Swanthana Susan Alex. 2019. Chatbot for disease prediction and treatment recommendation using machine learning. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 851–856.

Michael McTear. 2010. Chapter 9 - the role of spoken dialogue in user–environment interaction. In Hamid Aghajan, Ramón López-Cózar Delgado, and Juan Carlos Augusto, editors, *Human-Centric Interfaces for Ambient Intelligence*, pages 225–254. Academic Press, Oxford.

Sabrina J. Mielke. 2016. Language diversity in ACL 2004 - 2016.

Mahdi Naser Moghadasi, Yu Zhuang, and Hashim Gellban. 2020. Robo: A counselor chatbot for opioid addicted patients. In *2020 2nd Symposium on Signal Processing Systems*, SSPS 2020, page 91–95, New York, NY, USA. Association for Computing Machinery.

Joao Luis Zeni Montenegro, Cristiano André da Costa, and Rodrigo da Rosa Righi. 2019. Survey of conversational agents in health. *Expert Systems with Applications*, 129:56–67.

Fabrizio Morbini, David DeVault, Kallirroi Georgila, Ron Artstein, David Traum, and Louis-Philippe Morency. 2014. A demonstration of dialogue processing in SimSensei kiosk. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 254–256, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Fabrizio Morbini, Eric Forbell, David DeVault, Kenji Sagae, David Traum, and Albert Rizzo. 2012. A mixed-initiative conversational dialogue system for healthcare. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–139, Seoul, South Korea. Association for Computational Linguistics.

Denis Newman-Griffis, Jill Fain Lehman, Carolyn Rosé, and Harry Hochheiser. 2021. Translational NLP: A new paradigm and general principles for natural language processing research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4125–4138, Online. Association for Computational Linguistics.

Kyo-Joong Oh, Dongkun Lee, Byungsoo Ko, and Ho-Jin Choi. 2017. A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. In *2017 18th IEEE International Conference on Mobile Data Management (MDM)*, pages 371–375.

Alexandros Papangelis, Robert Gatchel, Vangelis Metsis, and Fillia Makedon. 2013. An adaptive dialogue system for assessing post traumatic stress disorder. In *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '13, New York, NY, USA. Association for Computing Machinery.

Natalie Parde. 2018. Reading with robots: Towards a human-robot book discussion system for elderly adults. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Natalie Parde and Rodney D. Nielsen. 2019. Ai meets austen: Towards human-robot discussions of literary metaphor. In *Artificial Intelligence in Education*, pages 213–219, Cham. Springer International Publishing.

Falguni Patel, Riya Thakore, Ishita Nandwani, and Santosh Kumar Bharti. 2019. Combating depression in students using an intelligent chatbot: A cognitive behavioral therapy. In *2019 IEEE 16th India Council International Conference (INDICON)*, pages 1–4.

Frano Petric, Damjan Miklic, and Zdenko Kovacic. 2017. Robot-assisted autism spectrum disorder diagnostics using pomdps. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, page 369–370, New York, NY, USA. Association for Computing Machinery.

Marco Polignano, Fedelucio Narducci, Andrea Iovine, Cataldo Musto, Marco De Gemmis, and Giovanni Semeraro. 2020. Healthassistantbot: A personal health assistant for the italian language. *IEEE Access*, 8:107479–107497.

A. Prange, Margarita Chikobava, P. Poller, Michael Barz, and D. Sonntag. 2017. A multimodal dialogue system for medical decision support inside virtual reality. In *SIGDIAL Conference*.

Libo Qin, Yijia Liu, Wanxiang Che, Haoyang Wen, Yangming Li, and Ting Liu. 2019. Entity-consistent end-to-end task-oriented dialogue system with kb retriever.

Juan C. Quiroz, Tristan Bongolan, and Kiran Ijaz. 2020. Alexa depression and anxiety self-tests: A preliminary analysis of user experience and trust. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, UbiComp-ISWC '20, page 494–496, New York, NY, USA. Association for Computing Machinery.

Md. Moshiur Rahman, Ruhul Amin, Md Nazmul Khan Liton, and Nahid Hossain. 2019. Disha: An implementation of machine learning based bangla healthcare chatbot. In *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, pages 1–6.

Michelle L Rogers, Emily Patterson, Roger Chapman, and Marta Render. 2005. Usability testing and the relation of clinical information systems to patient safety. In *Advances in Patient Safety: From Research to Implementation (Volume 2: Concepts and Methodology)*. Agency for Healthcare Research and Quality (US).

Nudtaporn Rosruen and Taweesak Samanchuen. 2018. Chatbot utilization for medical consultant system. In *2018 3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON)*, pages 1–5.

Sanket Sanjay Sadavarte and Eliane Bodanese. 2019. Pregnancy companion chatbot using alexa and amazon web services. In *2019 IEEE Pune Section International Conference (PuneCon)*, pages 1–5.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *CoRR*, abs/1512.05742.

Bhuvan Sharma, Harshita Puri, and Deepika Rawat. 2018. Digital psychiatry - curbing depression using therapy chatbot and depression analysis. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 627–631.

Tianhao She, Xin Kang, Shun Nishide, and Fuji Ren. 2018. Improving leo robot conversational ability via deep learning algorithms for children with autism. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pages 416–420.

Naohiro Shoji, Takayo Namba, and Keiichi Abe. 2020. Proposal of spoken interactive home doctor system for elderly people. In *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*, pages 421–423.

Noppon Siangchin and Taweesak Samanchuen. 2019. Chatbot implementation for icd-10 recommendation system. In *2019 International Conference on Engineering, Science, and Industrial Applications (ICESI)*, pages 1–6.

Daneil Sonntag and Manuel Moller. 2010. Prototyping semantic dialogue systems for radiologists. In *2010 Sixth International Conference on Intelligent Environments*, pages 84–89.

Daniel Sonntag, Gerhard Sonnenberg, Robert Neßelrath, and Gerd Herzog. 2009. Supporting a rapid dialogue engineering process. In *Proceedings of the First International Workshop On Spoken Dialogue Systems Technology (IWSDS)*.

Prakhar Srivastava and Nishant Singh. 2020. Automatized medical chatbot (medibot). In *2020 International Conference on Power Electronics IoT Applications in Renewable Energy and its Control (PARC)*, pages 351–354.

H. Strongman, R. Williams, W. Meeraus, T. Murray-Thomas, J. Campbell, L. Carty, D. Dedman, A. Gallagher, J. Oyinlola, A. Kousoulis, and J. Valentine. 2019. Limitations for health research with restricted data collection from uk primary care. *Pharmacoepidemiol Drug Saf*.

Bo-Hao Su, Shih-Pang Tseng, Yu-Shan Lin, and Jhing-Fa Wang. 2018. Health care spoken dialogue system for diagnostic reasoning and medical product recommendation. In *2018 International Conference on Orange Technologies (ICOT)*, pages 1–4.

Konstantinos Tsiakas, Lynette Watts, Cyril Lutterodt, Theodoros Giannakopoulos, Alexandros Papangelis, Robert Gatchel, Vangelis Karkaletsis, and Fillia Makedon. 2015. A multimodal adaptive dialogue manager for depressive and anxiety disorder screening: A wizard-of-oz experiment. In *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '15, New York, NY, USA. Association for Computing Machinery.

Lorainne Tudor Car, Dhakshenya Ardhithy Dhinagaran, Bhone Myint Kyaw, Tobias Kowatsch, Shafiq Joty, Yin-Leng Theng, and Rifat Atun. 2020. Conversational agents in health care: Scoping review and conceptual analysis. *J Med Internet Res*, 22(8):e17158.

A. Vaidyam, Hannah Wisniewski, J. Halamka, M. S. Kashavan, and J. Torous. 2019. Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64:456 – 464.

Mina Valizadeh, Pardis Ranjbar-Noiey, Cornelia Caragea, and Natalie Parde. 2021. Identifying medical self-disclosure in online communities. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4398–4408, Online. Association for Computational Linguistics.

Richard M Walker, Gene A Brewer, M Jin Lee, Nicolai Petrovsky, and Arjen van Witteloostuijn. 2018. Best Practice Recommendations for Replicating Experiments in Public Administration. *Journal of Public Administration Research and Theory*, 29(4):609–626.

Jinping Wang, Hyun Yang, Ruosi Shao, Saeed Abdullah, and S. Shyam Sundar. 2020. Alexa as coach: Leveraging smart speakers to build social agents that reduce public speaking anxiety. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

J. V. Waterschoot, Iris Hendrickx, Arif Khan, E. Klabbers, M. D. Korte, H. Strik, C. Cucchiarini, and M. Theune. 2020. Bliss: An agent for collecting spoken dialogue data about health and well-being. In *LREC*.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207, Melbourne, Australia. Association for Computational Linguistics.

Charles Welch, Allison Lahnala, Veronica Perez-Rosas, Siqi Shen, Sarah Seraj, Larry An, Kenneth Resnicow, James Pennebaker, and Rada Mihalcea. 2020. Expressive interviewing: A conversational system for coping with COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina Maria Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve J. Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *CoRR*, abs/1604.04562.

Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3506–3510.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Recipes for safety in open-domain chatbots.

L. Xu, Q. Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. *ArXiv*, abs/1901.10623.

Keigo Yabuki and Kaoru Sumi. 2018. Learning support system for effectively conversing with individuals with autism using a humanoid robot. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 4266–4270.

Akihiro Yorita, Simon Egerton, Carina Chan, and Naoyuki Kubota. 2020. Chatbot for peer support realization based on mutual care. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1601–1606.

Zheng Zhang, Ryuichi Takanobu, Minlie Huang, and Xiaoyan Zhu. 2020. Recent advances and challenges in task-oriented dialog system. *CoRR*, abs/2003.07490.

Yin Jiang Zhao, Yan Ling Li, and Min Lin. 2019. A review of the research on dialogue management of task-oriented systems. In *Journal of Physics: Conference Series*, volume 1267. IOP Publishing.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

A Multi-Objective Systems

Multi-Objective System	# Papers
Diagnosis + Assistance	7
Diagnosis + Intervention	2
Diagnosis + Monitoring	1
Diagnosis + Counseling	1
Intervention + Monitoring	2
Intervention + Assistance	1
Assistance + Counseling	2
Intervention + Monitoring + Diagnosis	2
Intervention + Monitoring + Assistance	2
Intervention + Monitoring + Counseling	1
Diagnosis + Monitoring + Counseling	1
Diagnosis + Assistance + Intervention	2
Diagnosis + Intervention + Monitoring + Assistance	1

Table 8: Distribution of varying combinations of multiple system objectives across the surveyed papers.

Conversational agents seek to generate dialogues that have value to their end-users. We categorized included articles as having one or more of the following objectives: diagnosis, monitoring, intervention, counseling, or assistance. We found that 25 out of 70 surveyed systems were designed for more than one target objective, and provide additional details describing these multi-objective systems in Table 8.

Multi-Device Category	# Papers
Desktop/Laptop + Mobile-based	8
Desktop/Laptop + VE	5
Desktop/Laptop + Robot	2
Mobile-based + PDA systems	2
Desktop/Laptop + GUI	1
Desktop/Laptop + PDA systems	1
Mobile-based + VE	1

Table 9: Details regarding the distribution of multi-device systems across the surveyed papers (20 total).

User Population	# Papers
Lab Experiments	15
Field Experiments	17
Crowdsourcing	1
Not Specified	4

Table 10: Distribution of user populations across the surveyed papers that conducted a human evaluation.

B Multi-Device Systems

Many of the surveyed systems functioned using multiple device types. Table 9 shows the distribution of included devices across all multi-device systems. We found that the most common multi-device pairing was systems operating using computers and mobile devices.

C Additional Evaluation Details

From among the surveyed systems that conducted system and/or human evaluations, we further examined the types of evaluations conducted. Table 10 describes the populations leveraged for human evaluation across the surveyed systems, and Table 11 presents broad categories of the types of human evaluations conducted. We found that most human evaluations were conducted in a laboratory or field setting, and often included opportunities for participants to both interact with the system directly, and rate the quality of the dialogue. Table 12 details

Human Evaluation Type	# Papers
Interact with the System	8
Rate a Dialogue	1
Both	28

Table 11: Distribution of evaluation types across the surveyed papers that conducted a human evaluation.

Type of System Evaluation	# Papers
Task Completion	4
Task Performance	9
Response Correctness	5
Naturalness	2
Response Time	3
Routing Time	1

Table 12: Type of system evaluation across the surveyed papers.

the various types of system evaluations conducted across the surveyed systems. We found that the most common assessment item in system evaluations was the system's overall task performance.

D Included Papers

In this systematic review, we investigated 4070 papers involving dialogue systems for healthcare applications, identifying 70 papers that satisfied our defined inclusion criteria. We comprehensively analyzed these papers on the basis of numerous technical factors. We provide aggregated statistics for each of these categories in the main body of the paper. In Table 13 beginning on the following page, we provide a listing of each included paper and its categorization across all included classes. Full references for each included paper can be found in the bibliography.

Table 13: All papers included in the survey, with their categorizations for each class.

Paper	Dialogue System Architecture	Dialogue Manager	Modality	Device	System Objective	Engagement	Domain of Research	Target Audience	Language	Evaluation Method	Dataset Size
Papangelis et al. (2013)	Pipeline	Intent-based	Multi-Modal	Desktop or Laptop	Monitoring, Intervention, Diagnosis	Yes	PTSD	Patients	English	Not Specified	Not Specified
Brinkman et al. (2012a)	Pipeline	Rule-based	Speech	Virtual Environment	Monitoring, Diagnosis	No	Social Phobia	Clinicians	English	Human Evaluation	Not Specified
Ali et al. (2020)	Pipeline	Intent-based	Speech	Desktop or Laptop	Monitoring, Assistance, Intervention	Yes	Autism Spectrum Disorder	Patients	English	Human Evaluation	46 Videos
Tsiakas et al. (2015)	Pipeline	Intent-based	Multi-Modal	Desktop or Laptop, Virtual Environment	Diagnosis, Assistance	Yes	Anxiety Disorders, Depression, PTSD	Patients	English	Human Evaluation	90 Speech Segments
Wang et al. (2020)	Pipeline	Hybrid	Speech	PDA	Intervention	Yes	Social Phobia	Patients	English	Human Evaluation	Not Specified
Balasuriya et al. (2018)	Pipeline	Hybrid	Speech, GUI	PDA	Monitoring	Yes	Intellectual Disability	Patients	English	Human Evaluation	Not Specified
Chuan and Morgan (2021)	Pipeline	Intent-based	Speech	Desktop or Laptop	Assistance	No	Clinical Application	Patients	English	Human Evaluation	Not Specified
Grover et al. (2009)	Pipeline	Rule-based	Speech	Telephone	Assistance	No	HIV	Clinicians	Setswana	Human & Automated Evaluation	Not Specified
Petric et al. (2017)	Pipeline	Intent-based	Speech	Robot	Diagnosis	No	Autism Spectrum Disorder	Clinicians	English	Human Evaluation	Not Specified
Javed et al. (2018)	Not Specified	Not Specified	Speech, GUI	Robot	Monitoring	Yes	Autism Spectrum Disorder	Patients	English	Human Evaluation	Not Specified
Di Nuovo et al. (2020)	Not Specified	Not Specified	Speech	Robot	Monitoring	Yes	Autism Spectrum Disorder	Patients, Caregivers	English	Human Evaluation	Not Specified
Quiroz et al. (2020)	Pipeline	Hybrid	Speech	PDA, Mobile	Diagnosis, Intervention	Yes	Depression, Anxiety	Patients	English	Human Evaluation	Not Specified

Paper	Dialogue System Architecture	Dialogue Manager	Modality	Device	System Objective	Engage-ment	Domain of Research	Target Audience	Lan-guage	Evaluation Method	Dataset Size
Maharjan et al. (2019)	Pipeline	Hybrid	Speech	PDA, Mobile	Monitoring	No	Mental Health	Patients	English	Not Specified	Not Specified
Ahn et al. (2020)	Not Specified	Not Specified	Text	Mobile	Intervention, Assistance	Yes	Online Sexual Exploitation, PTSD	Patients	Korean	Not Specified	Not Specified
Kamita et al. (2020)	Not Specified	Not Specified	Text	Mobile	Intervention	Yes	Cognitive Behavioral Therapy, Stress Reduction	Patients	Japanese	Human Evaluation	Not Specified
Lee et al. (2020b)	Pipeline	Hybrid	Speech	Mobile	Monitoring	Yes	Health-Related Self-Disclosure	Patients	English	Human Evaluation	Not Specified
Moghadasi et al. (2020)	Pipeline	Hybrid	Text	Desktop or Laptop, Mobile	Assistance, Counseling	No	Opioid Addiction	Patients	English	Not Specified	20,494 Records
De Nieva et al. (2020)	Pipeline	Hybrid	Text	Mobile	Monitoring, Intervention, Counseling	Yes	Anxiety, Depression	Patients	English	Human & Automated Evaluation	Not Specified
Lee et al. (2020a)	Pipeline	Hybrid	Text	Mobile	Monitoring	Yes	Health-Related Self-Disclosure	Patients	English	Human Evaluation	Not Specified
Daher et al. (2020)	Pipeline	Rule-based	GUI	Not Specified	Monitoring	No	Empathy for Medical Assistance	Patients	English	Human Evaluation	Not Specified
Holmes et al. (2019)	Pipeline	Hybrid	Multi-Modal	Mobile	Assistance	Yes	Weight Loss	Patients	English	Human & Automated Evaluation	Not Specified
Oh et al. (2017)	Pipeline	Intent-based	Multi-Modal	Mobile	Diagnosis, Monitoring, Intervention	Yes	Psychiatric Counseling	Patients	Korean	Not Specified	49,846,477 Records
Dino et al. (2019)	Pipeline	Rule-based	Speech	Robot	Intervention	Yes	Depression	Patients	English	Human Evaluation	Not Specified

Paper	Dialogue System Architecture	Dialogue Manager	Modality	Device	System Objective	Engage-ment	Domain of Research	Target Audience	Lan-guage	Evaluation Method	Dataset Size
Patel et al. (2019)	Not Specified	Not Specified	Text	Not Specified	Diagnosis	No	Stress, Depression	Patients	English	Not Specified	7,652 Records, ISEAR Dataset
Sharma et al. (2018)	Not Specified	Not Specified	Text	Mobile	Diagnosis, Intervention, Assistance	No	Depression	Patients	Not Specified	Not Specified	Not Specified
Belfin et al. (2019)	Pipeline	Intent-based	Multi-Modal	Desktop or Laptop, Mobile	Assistance	No	Cancer	Patients	English	Not Specified	Not Specified
Yorita et al. (2020)	Pipeline	Rule-based	Multi-Modal	Mobile	Diagnosis, Counseling	No	Stress Management	Clinicians	English	Not Specified	Not Specified
Kargar and Mahoor (2017)	Pipeline	Rule-based	Speech	Robot	Intervention	Yes	Depression	Patients	English	Human Evaluation	Not Specified
Hwang et al. (2020)	Pipeline	Rule-based	Text	Not Specified	Diagnosis, Intervention	No	Medical Assistance	Patients	Korean	Not Specified	Not Specified
Srivastava and Singh (2020)	Pipeline	Rule-based	Text	Not Specified	Diagnosis, Assistance	Yes	Disease Diagnosis	Patients	English	Human Evaluation	Not Specified
Mathew et al. (2019)	Pipeline	Rule-based	Text	Mobile	Diagnosis, Assistance	Yes	Disease Diagnosis	Patients	English	Human Evaluation	Not Specified
Athota et al. (2020)	Pipeline	Rule-based	Multi-Modal	Mobile	Diagnosis, Assistance	No	Disease Diagnosis	Patients	English	Not Specified	Not Specified
Sadavarte and Bodanese (2019)	Pipeline	Hybrid	Multi-Modal	PDA	Assistance	No	Pregnancy	Patients	English	Human Evaluation	Not Specified
Lee et al. (2017)	Pipeline	Hybrid	Text	Mobile	Counseling	Yes	Psychiatric Counseling	Patients	Korean	Not Specified	Not Specified
Rahman et al. (2019)	Pipeline	Hybrid	Text	Not Specified	Diagnosis, Monitoring, Counseling	No	Medical Assistance	Patients	Bengali	Automated Evaluation	4,961 records
Yabuki and Sumi (2018)	Not Specified	Not Specified	Speech	Robot	Intervention	No	Autism Spectrum Disorder	Caregivers	English	Not Specified	Not Specified

Paper	Dialogue System Architecture	Dialogue Manager	Modality	Device	System Objective	Engage-ment	Domain of Research	Target Audience	Lan-guage	Evaluation Method	Dataset Size
Su et al. (2018)	Pipeline	Intent-based	Speech	Not Specified	Diagnosis, Assistance	No	Disease Diagnosis	Patients	Chinese	Automated Evaluation	Not Specified
Shoji et al. (2020)	Not Specified	Not Specified	Speech	Desktop or Laptop, PDA	Diagnosis	No	Pneumonia	Patients	Not Specified	Automated Evaluation	Not Specified
Polignano et al. (2020)	Pipeline	Hybrid	Multi-Modal	Mobile	Diagnosis, Intervention, Assistance, Monitoring	No	Medical Assistance	Patients	Italian	Human & Automated Evaluation	1,865,700 Records
Ali et al. (2021)	Pipeline	Hybrid	Speech	Desktop or Laptop, Virtual Environment	Intervention	No	Cancer	Clinicians	English	Automated Evaluation	382 Conversation Transcripts
Aarabi (2013)	Pipeline	Intent-based	Text	Not Specified	Diagnosis	No	Cardiology	Patients	English	Not Specified	Not Specified
Loisel et al. (2007)	Pipeline	Hybrid	Text	Not Specified	Assistance	No	Medical Assistance	Patients	French	Not Specified	Not Specified
Rosruen and Samanchuen (2018)	Pipeline	Hybrid	Multi-Modal	Desktop or Laptop, Mobile	Assistance	No	Medical Assistance	Patients	Chinese	Automated Evaluation	Not Specified
Sonntag and Moller (2010)	Pipeline	Intent-based	Multi-Modal	Desktop or Laptop	Assistance	Yes	Radiology	Clinicians	Not Specified	Human & Automated Evaluation	Not Specified
Kadariya et al. (2019)	Pipeline	Hybrid	Multi-Modal	Mobile	Monitoring, Intervention	Yes	Asthma	Patients	English	Human & Automated Evaluation	Not Specified
Siangchin and Samanchuen (2019)	Pipeline	Hybrid	Text	Mobile	Assistance	No	Medical Assistance	Clinicians	Chinese	Human & Automated Evaluation	Not Specified
Erazo et al. (2020)	Pipeline	Rule-based	Text	Desktop or Laptop, Mobile	Diagnosis, Assistance	No	Covid-19	Patients	Not Specified	Human Evaluation	Not Specified
Huang et al. (2018)	Pipeline	Hybrid	Multi-Modal	Mobile	Monitoring, Intervention	Yes	Weight Loss	Patients	English, Chinese	Not Specified	Not Specified

Paper	Dialogue System Architecture	Dialogue Manager	Modality	Device	System Objective	Engagement	Domain of Research	Target Audience	Language	Evaluation Method	Dataset Size
Chen et al. (2013)	Pipeline	Rule-based	Speech	Desktop or Laptop, Mobile	Assistance	No	Medical Assistance	Patients, Caregivers	Chinese	Human Evaluation	MAT 400 Dataset
Araki et al. (2011)	Pipeline	Intent-based	Multi-Modal	Desktop or Laptop	Intervention	No	Visual Impairment	Patients	Japanese	Human Evaluation	Not Specified
She et al. (2018)	End-to-End	Not Applicable	Speech	Robot	Intervention	Yes	Autism Spectrum Disorder	Patients	English	Automated Evaluation	Tager-Flusberg, Nadig ASD English, and Rollins Corpus
Yabuki and Sumi (2018)	Not Specified	Not Specified	Speech	Robot	Intervention	Yes	Autism Spectrum Disorder	Caregivers	Japanese	Not Specified	Self-Constructed Dataset
Wei et al. (2018)	Pipeline	Intent-based	Text	Not Specified	Diagnosis	No	Medical Assistance	Clinicians	Chinese	Automated Evaluation	Self-Constructed Dataset
Fadhil and AbuRa'ed (2019)	Pipeline	Intent-based	Multi-Modal	Mobile	Monitoring, Assistance, Intervention	No	Medical Assistance	Patients	Arabic	Human Evaluation	Not Specified
Demasi et al. (2020)	Pipeline	Intent-based	Text	Not Specified	Counseling	No	Mental Health	Patients	English	Human Evaluation	Self-Constructed Dataset
Waterschoot et al. (2020)	Pipeline	Intent-based	Speech	Not Specified	Monitoring	No	Mental Health	Patients	Dutch	Not Specified	Self-Constructed Dataset
Danda et al. (2016)	Pipeline	Hybrid	Speech	Desktop or Laptop, Mobile	Diagnosing, Intervention, Assistance	No	Medical Assistance	Patients	Indian	Human & Automated Evaluation	CMU Arctic Dataset
Duggenpudi et al. (2019)	Pipeline	Rule-based	Text	Not Specified	Assistance	No	Medical Assistance	Patients	Telugu	Human Evaluation	Self-Constructed Dataset
Prange et al. (2017)	Pipeline	Rule-based	Multi-Modal	Mobile	Assistance	No	Medical Assistance	Clinicians	Not Specified	Not Specified	475 Records

Paper	Dialogue System Architecture	Dialogue Manager	Modality	Device	System Objective	Engagement	Domain of Research	Target Audience	Language	Evaluation Method	Dataset Size
Campillos Llanos et al. (2015)	Pipeline	Intent-based	Multi-Modal	Not Specified	Intervention	No	Medical Assistance	Clinicians	French	Not Specified	Not Specified
Welch et al. (2020)	Pipeline	Intent-based	Text	Not Specified	Counseling, Assistance	Yes	Mental Health	Patients	Not Specified	Human Evaluation	Not Specified
Ljunglöf et al. (2009)	Pipeline	Intent-based	Speech	Desktop or Laptop, Robot	Intervention	No	Communication Disorders	Patients	Swedish	Human Evaluation	Not Specified
Ljunglöf et al. (2011)	Pipeline	Intent-based	Speech	Desktop or Laptop, Robot	Intervention	Yes	Communication Disorders	Patients	Swedish	Human Evaluation	Not Specified
Brixey et al. (2017)	Pipeline	Hybrid	Text	Desktop or Laptop, Mobile	Assistance	No	HIV	Patients	English	Human Evaluation	Self-Constructed Dataset
Morbini et al. (2014)	Pipeline	Rule-based	Speech	Desktop or Laptop, Virtual Environment	Counseling	Yes	Mental Health	Patients	English	Not Specified	Not Specified
DeVault et al. (2013)	Not Specified	Not Specified	Speech	Desktop or Laptop, Virtual Environment	Diagnosis	No	Mental Health	Clinicians	English	Not Specified	Not Specified
Inoue et al. (2016)	Pipeline	Rule-based	Multi-Modal	Mobile, Virtual Environment	Counseling	Yes	Mental Health	Patients	Not Specified	Not Specified	Not Specified
Morbini et al. (2012)	Pipeline	Intent-based	Text	Desktop or Laptop, Mobile	Counseling	Yes	PTSD	Patients	English	Not Specified	Not Specified
Xu et al. (2019)	End-to-End	Not Applicable	Text	Not Specified	Diagnosis	No	Disease Diagnosis	Patients	Chinese	Human & Automated Evaluation	Self-Constructed Dataset
Green et al. (2004)	Pipeline	Rule-based	Speech	Desktop or Laptop	Intervention	No	Dementia	Caregivers	English	Human Evaluation	Not Specified