Lessons Learned from the Usability Evaluation of a Simulated Patient Dialogue System

Leonardo Campillos-Llanos, Catherine Thomas, Éric Bilinski, Antoine Neuraz, Sophie Rosset, Pierre Zweigenbaum

▶ **To cite this version:**

Leonardo Campillos-Llanos, Catherine Thomas, Éric Bilinski, Antoine Neuraz, Sophie Rosset, et al.. Lessons Learned from the Usability Evaluation of a Simulated Patient Dialogue System. Journal of Medical Systems, 2021, 45 (7), 10.1007/s10916-021-01737-4 . hal-03452553

HAL Id: hal-03452553
https://hal.science/hal-03452553

Submitted on 27 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

AUTHOR'S PROOF

JrnlID 10916_ArtID 1737_Proof#1 - 19/04/2021

Journal of Medical Systems _####################_
https://doi.org/10.1007/s10916-021-01737-4

EDUCATION & TRAINING

Lessons Learned from the Usability Evaluation of a Simulated Patient Dialogue System

Leonardo Campillos-Llanos[1,2] · Catherine Thomas[3] · Éric Bilinski[3] · Antoine Neuraz[3] · Sophie Rosset[3] · Pierre Zweigenbaum[3]

Received: 23 December 2020 / Accepted: 5 April 2021
© Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Simulated consultations through virtual patients allow medical students to practice history-taking skills. Ideally, applications should provide interactions in natural language and be multi-case, multi-specialty. Nevertheless, few systems handle or are tested on a large variety of cases. We present a virtual patient dialogue system in which a medical trainer types new cases and these are processed without human intervention. To develop it, we designed a patient record model, a knowledge model for the history-taking task, and a termino-ontological model for term variation and out-of-vocabulary words. We evaluated whether this system provided quality dialogue across medical specialities (n = 18), and with unseen cases (n = 29) compared to the cases used for development (n = 6). Medical evaluators (students, residents, practitioners, and researchers) conducted simulated history-taking with the system and assessed its performance through Likert-scale questionnaires. We analysed interaction logs and evaluated system correctness. The mean user evaluation score for the 29 unseen cases was 4.06 out of 5 (very good). The evaluation of correctness determined that, on average, 74.3% (sd = 9.5) of replies were correct, 14.9% (sd = 6.3) incorrect, and in 10.7% the system behaved cautiously by deferring a reply. In the user evaluation, all aspects scored higher in the 29 unseen cases than in the 6 seen cases. Although such a multi-case system has its limits, the evaluation showed that creating it is feasible; that it performs adequately; and that it is judged usable. We discuss some lessons learned and pivotal design choices affecting its performance and the end-users, who are primarily medical students.

Keywords Medical history taking · Natural language processing · Education · Medical · Virtual patient · Artificial intelligence

UNCORRECTED PROOF

Introduction

Developing diagnosis and clinical reasoning skills is a key element of medical education. In addition to clinical practice, medical students and practitioners can enhance these abilities by means of mannequins, role games and simulation systems. These have shown beneficial results [1–7] and are currently integrated in virtual patients [8–15]. Virtual patients (VPs)[1] are software through which students can train themselves by emulating the roles of health providers [16].

Ideally, a VP simulation system should simulate a patient in all consultation stages. The patient's medical history taking (*anamnesis*) is an essential but difficult-to-master skill. Real consultations occur in time-restricted settings and there is a language-level gap in doctor-patient communication. Due to the health implications, doctors need to receive training to acquire these skills so that they assess patients' conditions and make a correct diagnosis.

This article is part of the Topical Collection on *Education & Training*

✉ Leonardo Campillos-Llanos
campillos@limsi.fr; leonardo.campillos@csic.es

Sophie Rosset
sophie.rosset@lisn.upsaclay.fr

Pierre Zweigenbaum
pz@lisn.upsaclay.fr

[1] Université Paris-Saclay, CNRS, LIMSI, Orsay, France

[2] Present address: Consejo Superior de Investigaciones Científicas (CSIC), Madrid, Spain

[3] Université Paris-Saclay, CNRS, LISN, Orsay, France

[1] We refer with this term to *virtual standardised patients*.

AUTHOR'S PROOF

Natural language dialogue systems (*chatbots* or *conversational agents*) have been integrated in healthcare applications [17–19] and VP simulation environments. Interaction modules allow trainees to simulate history taking, mostly through constrained input—e.g. lists of questions and answers prepared for a specific case [11, 20–25]. Other methods for processing user input use rules, ontologies and knowledge bases [26, 27], statistical language models [28], machine-learning classifiers [29], crowd-sourcing data [22] and preliminary neural approaches [30, 31]. Some systems feature automatic speech recognition [32–34]. However, very few virtual patients feature dialogue through natural language [34] (humans' inherent mode of communication), which might result in more natural interaction with a conversational agent [35, 36].

A successful interaction relies both on the type of technology and the degree to which the VP helps users to acquire clinical reasoning and history-taking skills. To do so, interacting with a wide range of cases is beneficial [36]. Accordingly, a VP system should provide simulations with a variety of clinical specialities. Most systems, nonetheless, only deal with one or a few conditions [33, 34, 37–43]. Very few systems cope with diverse pathologies [22, 44].

Objectives

Our objective was to overcome the limitation of the scarce number of simulated cases by designing a dialogue-enabled VP system that can cope with a variety of clinical conditions. We hypothesise that a multi-case VP can be achieved if medical trainers can create VPs easily, through a graphical interface (Fig. 6, Appendix), without programming anything nor the development team's intervention. The description of the clinical case, in the form of a semi-structured record, is typed offline in natural language; next, the dialogue system embodies a patient with each clinical case.

Accordingly, a first requirement of the system is to cope with new contents across medical specialities. The second requirement is to provide unconstrained input, because the system aims at improving medical students' history-taking skills through the interaction with the VP. Figure 1 is a sample dialogue and illustrates natural dialogue phenomena. The system is integrated in a serious game developed with partner companies and a medical team [45]. The software features an animated avatar with text-to-speech, lip-synch and minor gestures.

To make the system able to handle plenty of cases, we gave it extensive conceptual and terminological coverage of the domain [27, 46]. The system can also adapt to new records dynamically. We provided it with components to detect out-of-vocabulary words (OOV) and predict morphological information of missing words. The system with adaptation modules is available in French;[2] English and Spanish versions are available but not well-supported.

This article reports a usability evaluation of the French system, where we assessed, in a simulated history-taking setting:

- Q1 Whether a multi-case system can provide quality dialogue (with regard to grammar and on-topic and realistic replies) through natural language across clinical cases.
- Q2 Whether quality dialogue is maintained when processing unseen records across medical specialities.

We evaluated these aspects through user experiments in a real context. Study participants (n = 39) interacted in French language with the dialogue system, then performed a user evaluation of their dialogue.

Material and methods

Dialogue system architecture

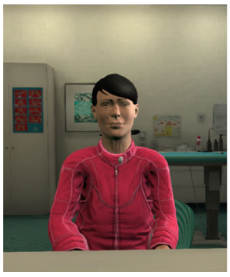
To tackle the task, we first designed a patient record model, which defines a virtual patient's health state in a semi-structured format. Table 9 (Appendix) shows an example. Second, we conceived a knowledge model for the task, i.e. a scheme of question types, dialogue acts and entity types concerning the anamnesis. Third, we created a termino-ontological model, which hosts structured thesauri for managing the variation of terms [46, 47]. Figure 2 is a schema of the different stages (which occur asynchronously): case creation by an instructor (1), comparison and analysis of a new record (2), and dialogue by a student (3).

We designed the system following a knowledge-based and rule-frame-based approach [27]. The user—typically a medical student or resident—types text. A natural language understanding (NLU) module performs the linguistic and semantic processing (e.g. *pain* is a SYMPTOM). A semantic frame is fed to a dialogue manager, which keeps track of the dialogue state and context information, queries the record, selects the information and replies through a template-based generation module (Fig. 3).

The termino-ontological model contains lexical resources for processing linguistic variation: inflection (e.g. *lung* ↔ *lungs*), derivation (e.g. *face* ↔ *facial*), synonymy (e.g. *operation* ↔ *surgery*) and mapping between full words and affixes/roots (e.g. *heart* ↔ *cardio-*). The model also defines domain relations and concepts for processing and normalising the variety of terms in a case: e.g. *pain* and

[2] http://vps-9069f76a.vps.ovh.net

UNCORRECTED PROOF



D: What brings you here?
P: I am seeing you about persistent backache
D: When did it start?
P: I don't remember
D: Where is the pain exactly?
P: I have pain in the lower back, the pain is located in the lower back and sometimes down the legs
D: Can you sleep well?
P: The pain often makes it difficult to fall asleep
D: Do you suffer from any disease?
P: Chronic stomac ulcers
D: Do you follow any treatment for that?
P: Omeprazole for the chronic stomach ulcers. Maybe there is more information in my file

Fig. 1 Sample of an actual dialogue of a medical student (D for Doctor) with a virtual patient (P)—the transcript comes from a session with the English version of our system

ache refer to the same concept. These resources support a key feature of the system: its ability to map *doctor's language* to *patient's language* to better simulate a real patient. We populated this model with large general and domain resources (e.g., the Unified Medical Language System® [48]). Our lexicons contain domain lists (over 161,000 terms in French, 116,000 in English, and 103,000 in Spanish) and dictionaries (over 959,000 word/concept entries in French, 1,886,000 in English, and 1,428,000 in Spanish).

Although these resources allow the system to handle plenty of cases, the medical jargon evolves continually with neologisms. Not knowing out-of-vocabulary words (OOVs) might cause incorrectly generated replies, because the system lacks the linguistic information for morphological agreement of OOVs. We thus developed methods to predict the Part-of-Speech (PoS) and gender/number of OOVs (see Table 9 in the Appendix). Multiple approaches are run in parallel: dictionary-based, and inference from linguistic context or from the base form/affixes (Fig. 7 in the Appendix). They are combined using heuristic weights set during development. This prediction is executed offline whenever an instructor creates or modifies a case. Figure 8 (Appendix) gives more technical details of the system components.

Evaluation design

To assess whether the system provides quality dialogue across clinical cases (Q1), potential end-users (n = 39) tested 35 different VPs. Medical students, interns and expert practitioners conducted medical history-taking in French language with a VP and evaluated the system performance in different evaluation rounds in two types of conditions (Table 1). Some sessions used *unseen cases* that were just created; we did not modify the system between creation and use. Other sessions used already *seen cases*, created earlier, for which we had fine-tuned the system manually. The system evolved over evaluation rounds and improved gradually by correcting the errors in interaction logs.

The medical evaluators had varied profiles (Table 2) and some participated in multiple evaluation rounds. Medical instructors created the content of 6 seen and 23 unseen cases. A co-author of this paper (LC) input the records of

UNCORRECTED PROOF

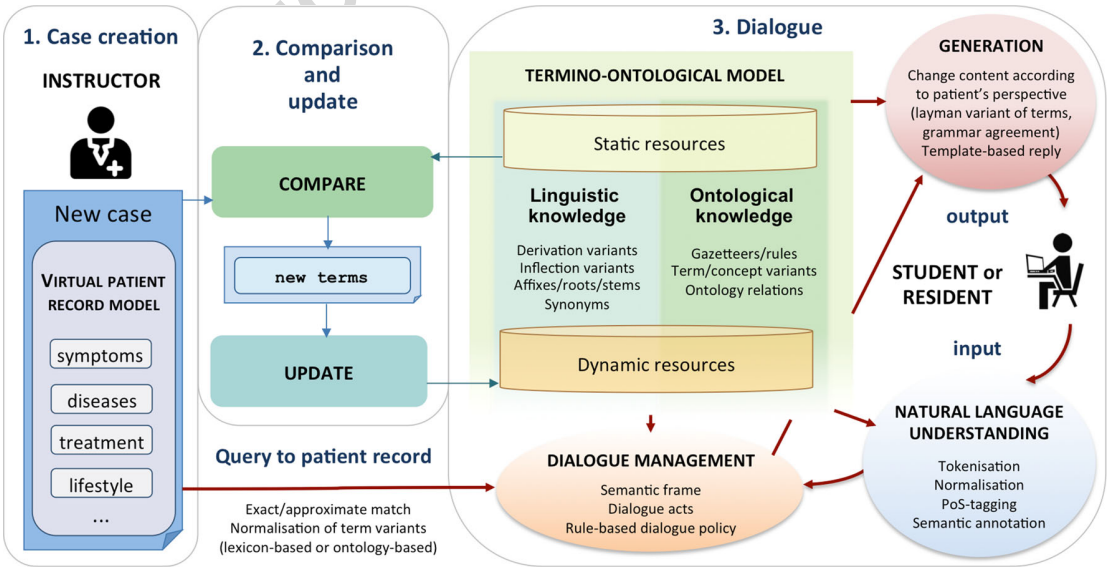


Fig. 2 Schema of the virtual patient dialogue system and update components

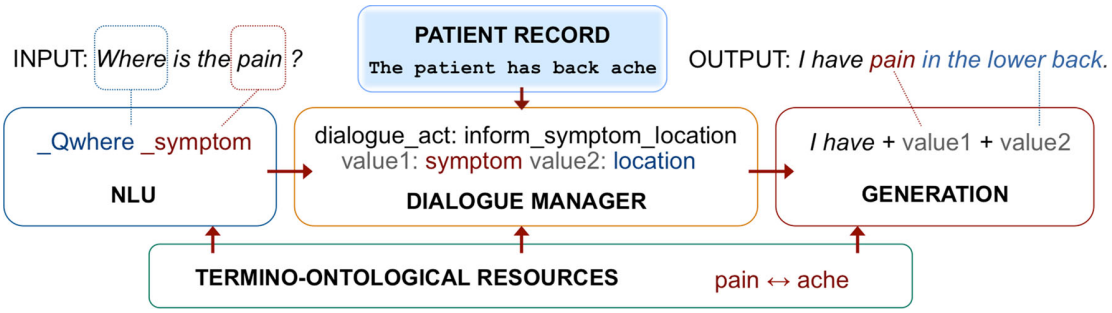


Fig. 3 Example of functioning of the dialogue system from input to output. The patient record is simplified; Table 9 shows a full example

6 unseen cases using the wordings of the clinical cases of French national classifying exams for medical students.[3] Tables 10 and 11 (Appendix) provide a brief description of each case.

We first conducted a user evaluation by means of 5-point Likert-scale questionnaires ranging from 1 (Very poor) to 5 (Very good). After each interaction, evaluators assessed the system on nine aspects (Table 3), which come from the evaluation framework of dialogue systems [49, 50]. Evaluators were given instructions on the types of utterances the system can process, and an online link to the questionnaire.

We also evaluated the dialogue system's correctness. We gathered data from the dialogues with all the 35 VP cases. We analysed dialogue logs and quantified the number of correct replies. We considered correct those replies giving a coherent answer (consistent according to the user input and correct regarding the data in the record). Table 6 (Appendix) describes some examples of correct, incorrect and deferred replies. An author of this paper (LC) annotated all data; another author (SR) checked the annotations of a subset of 84 (2%) turn-reply pairs that were unclear about how to classify; finally, a consensus was reached. We computed the kappa agreement between both annotators.

To evaluate whether quality dialogue is maintained with new cases (Q2), we compared the evaluation scores given to seen and unseen cases (Table 1). 26 of the 39 medical evaluators assessed 6 seen VP cases (50 questionnaires), and 23 of the 39 evaluators evaluated 29 unseen cases (67 questionnaires); some evaluators assessed both seen and unseen cases. We conducted two-tailed t-tests and Mann-Whitney tests, using the Prism 5 software, to determine if the differences in scores were statistically significant.

To measure the diversity of the unseen cases, we counted the word types (i.e. different word forms) appearing in only one record, and the types shared across different cases. The unseen cases belong to 14 specialities (Table 1).

[3] http://umvf.cerimes.fr/portail/ecn.php

We analysed how scores varied according to evaluators' profiles.

Results

Quality of natural language dialogue

Each case was tested by an average of 3.74 evaluators (±2.8; minimum number of evaluators per case = 1; maximum = 13). Panels A and B of Fig. 4 display the average evaluator scores for the seen and unseen cases respectively. Lower scores are placed to the left of each Y axis; neutral scores, in the middle; and higher scores, to the right. The bars show the cumulated percentages of evaluator scores that were Very good, Good, Neutral, Poor and Very poor. For example, in the seen cases, performance was assessed as Very good by 6% of the evaluators, as Good by an additional 52% of evaluators, as Neutral by 28% of them, and as Poor by the remaining 14%. The overall average score, obtained by averaging the mean scores given to the 9 evaluated aspects, was of 3.84 out of 5 for seen cases, and of 4.05 for unseen cases. This is above the Likert-scale midpoint. The total number of dialogues with Poor or Very poor scores ranges from 16% (naturalness) to 0% (user-understanding) for seen cases, and from 6% (naturalness) to 0% (speed) for unseen cases.

Regarding the system correctness, we analysed 8,078 turn-reply pairs from 131 dialogues (Tables 4 and 5). We removed 149 turn-reply pairs with out-of-task questions or statements. The two researchers who double-checked the subset of turn-reply pairs had a kappa agreement of 0.827. In the full set of dialogue logs (seen and unseen cases), when analysed per medical specialty, an average of 74.3% (±9.5) system replies were correct (min = 53.6%, max = 93.8%), i.e. answers were coherent with regard to inputs and provided accurate information from the record. An average of 14.9% (±6.3) of system replies were incorrect; however, unseen words only caused 2 errors. Incorrect replies affected the *system's faithfulness* (26.5%), *the dialogue flow* (56.2%)

UNCORRECTED PROOF

AUTHOR'S PROOF

JrnlID 10916_ArtID 1737_Proof#1 - 19/04/2021

Table 1 Evaluation rounds and medical specialities

	Development	Test				
	2016 through May 2017	July 2017	Oct 2017	Dec 2017	Jan 2018	Feb 2018
Evaluators	20	6	4	10	4	10
# cases	6	5	4	6 +3 (dev)	8	6 + 7 (from Jan 2018)
Medical specialities (# cases)	AN(1), CD(1), GP(1), PN(1), P(1), U(1)	N(2), CD(1), RH(1), ON(1)	OG(1), PN(1), GH(1), RH(1)	AN(1), CD(3), D(1), GE(1), GH(1), NE(1), PN(1), UC(1)	GH(3), ID(1), N(1), OG(1), PN(2)	GH(3), E(1), ID(2), N(3), PN(2), OG(1), OT(1)
Medical specialities in development+test (Total # of cases) [# of dialogues]						
AN: Anesthesiology (1) [11]		GP: General Practice (1) [6]		OT: Otolaryngology (1) [2]		
CD: Cardiology (1 + 3) [9 + 8]		ID: Infectious Diseases (2) [5]		PN: Pneumology (1 + 4) [13 + 10]		
D: Dermatology (1) [5]		NE: Nephrology (1) [2]		P: Psychiatry (1) [5]		
E: Endocrinology (1) [3]		N: Neurology (4) [15]		RH: Rheumatology (2) [7]		
GE: Geriatrics (1) [1]		OG: Obstetrics/Gynecaelogy (3) [4]		UC: Urgent Care (1) [1]		
GH: Gastroenterology/Hepatology (5) [13]		ON: Oncology (1) [5]		U: Urology (1) [6]		

and the *exhaustiveness of the information* provided by the virtual patient (17.3%) (Table 8, Appendix). The system determined that the rest of the questions were beyond the dialogue task and answered *I do not understand* (an average of 7.8% ±5.3) or asked for more precision (an average of 2.9% ±2.7). This defers giving an incorrect reply and is an additional average 10.7% of correct system behaviour, despite having a negative impact on the *dialogue flow*. When analysing the data per dialogue, results obtained were very similar (Table 5).

Table 2 Medical evaluators' profiles

Profile	Evaluators	S	U	Description
Students	♂	0	3	Students were in their 3rd year of medical studies and had limited experience with real patients (1-3 terms of part-time hospital internship).
	♀	3	4	
	Unique:	7 (3,	7)	
Residents	♂	2	5	Residents had at least 6 years of medical studies and passed the National Classifying Exam; they had broader experience than students (one or more full-time terms as practising physicians).
	♀	4	2	
	Unique:	10 (6,	7)	
Practitioners /Instructors	♂	8	4	Practitioners were private doctors or practising doctors in hospital or general practise.
	♀	0	1	
	Unique:	11 (8,	5)	
Researchers /Other	♂	5	1	Researchers included non-practising doctors, such as PhD students and postdoctoral researchers. Other profiles include doctors working for a drug database publisher or those whose profile was undeclared (anonymous evaluators).
	♀	0	0	
	NA	4	3	
	Unique:	11 (9,	4)	
Total unique	♂	15 (57.7% of 26)	13 (56.5% of 23)	
	♀	7 (26.9% of 26)	7 (30.4% of 23)	
	NA	4 (15.4% of 26)	3 (13.1% of 23)	
	Unique:	39 (26, 23)		

We report the number of evaluators for seen (S) and unseen (U) conditions. The total of unique participants of each profile is not always the sum of subjects in seen and unseen conditions, since some evaluators tested only seen or unseen cases, but others tested in both conditions. *NA* stands for 'not available' information

UNCORRECTED PROOF

Table 3 Description of aspects addressed in the qualitative evaluation; scores ranged from 5 (*Very good*) to 1 (*Very poor*)

Performance	An overall assessment of the system's global functioning.
Coherence	Adequateness of system answers in relation to user input.
Informativeness	Satisfaction with the information provided by the system.
User-understanding	Degree of comprehension of system replies by the user.
Speed	System quickness in replying to the user.
Tediousness	Verbosity of information answered by the system.
Answer concision	Quality of replies with regard to their length.
System-understanding	System degree of comprehension of user input.
Naturalness of replies	Realism of the utterances produced by the system.

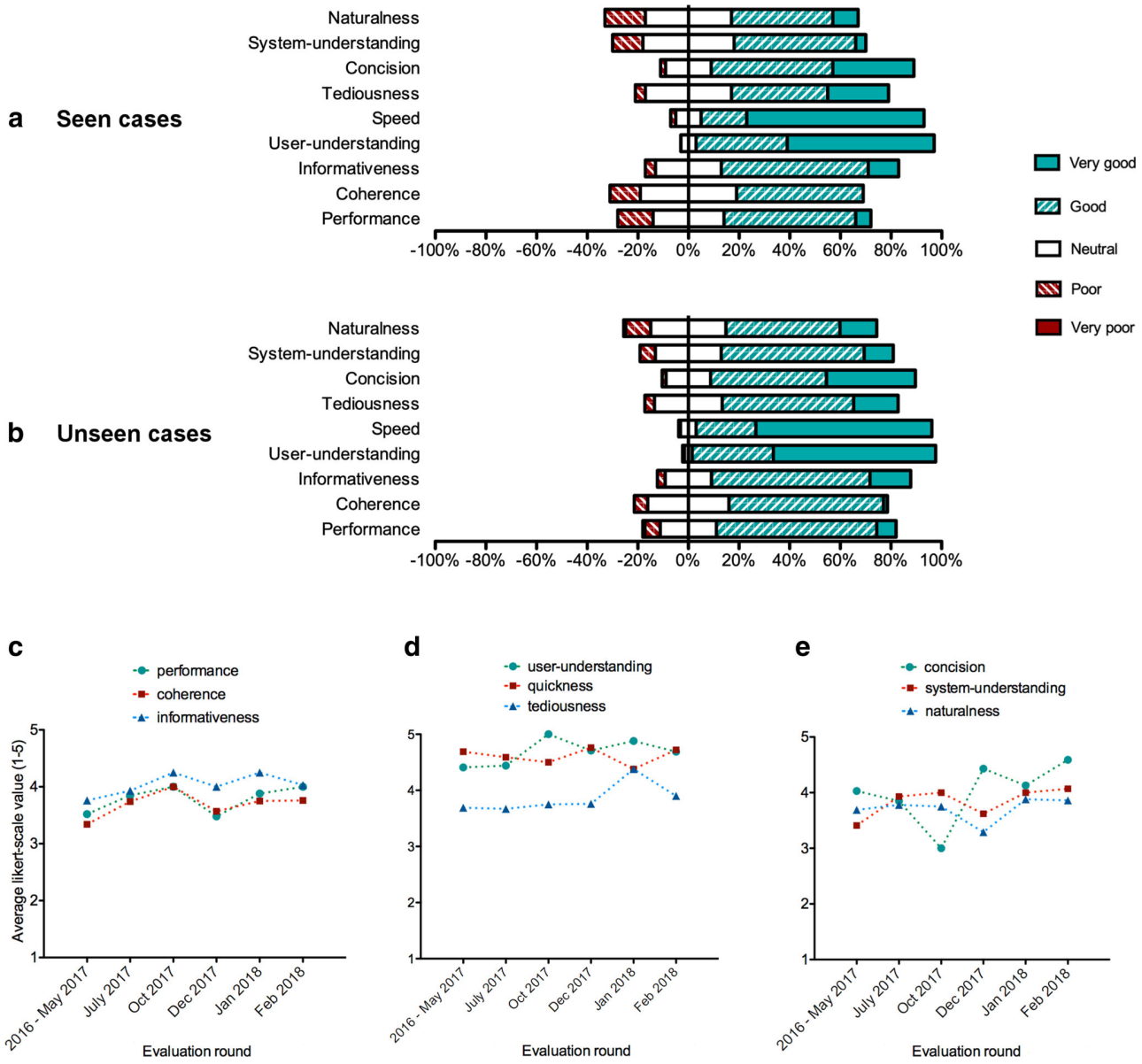


Fig. 4 Results of the qualitative evaluation and comparison between *seen* cases (used in development) and *unseen* cases

AUTHOR'S PROOF

Table 4 Evaluation data for all collected dialogues (*#d* = 131): *#T*: count of turns; *#W*: count of words; *stdev*: standard deviation; *#U/d*: average turns per dialogue; *#W/d*: average words per dialogue

	Turn reply-pairs		Words	
	#T	#T/d (stdev)	#W	#W/d (stdev)
User's input	4,044	30.9 (±11.7)	21,986	167.8 (±78.3)
System's reply	4,034	30.8 (±11.7)	21,921	167.3 (±78.5)
Total	8,078	61.7 (±11.7)	43,907	335.2 (±78.4)

Performance with unseen cases across specialities

Panels A and B of Fig. 4 display, respectively, the proportion of scores given to each aspect for the 6 seen and 29 unseen cases. Evaluators rated every aspect better in the unseen cases. The differences in evaluation scores were statistically significant for the following aspects: system performance (a mean of 3.50 (95% CI[3.27-3.73]) for seen cases versus 3.81 (95% CI[3.64-3.97]) for unseen cases, p-value = 0.029, Mann-Whitney test), coherence in replies (a mean of 3.38 (95% CI[3.18-3.58]) for seen cases versus 3.73 (95% CI[3.61-3.86]) for unseen cases, p = 0.004, Mann-Whitney test), informativeness (a mean of 3.78 (95% CI[3.58-3.98]) for seen cases versus 4.03 (95% CI[3.86-4.20]) for unseen cases, p = 0.047, Mann-Whitney test) and system-understanding (a mean of 3.44 (95% CI[3.22-3.66]) for seen cases versus 3.90 (95% CI[3.72-4.07]), p = 0.001, t-test).

We also examined the variation of scores along evaluation rounds; panels C-E in Fig. 4 show the average scores for each aspect. When we compared the scores given in the first evaluation round (using seen cases) with those in the last round (using unseen cases), the following aspects showed statistically significant differences: performance (a mean of 3.48 (95% CI[3.21-3.74]) in the first round versus 4.00 (95% CI[3.86-4.14]) in the last round, p = 0.003, Mann-Whitney test), coherence (a mean of 3.31 (95% CI[3.09-3.53]) in the first round versus 3.76 (95% CI[3.56-3.95]) in the last round, p = 0.005, t-test), informativeness (a mean of 3.69 (95% CI[3.48-3.90]) in the first round versus 4.03 (95% CI[3.87-4.19]) in the last round, p = 0.018, Mann-Whitney test), concision (a mean of 4.00 (95% CI[3.76-4.24]) in the first round versus 4.59 (95% CI[4.40-4.78]) in the last round, p = 0.001, Mann-Whitney test), and system-understanding (a mean of 3.36 (95% CI[3.11-3.60]) in the first round versus 4.07 (95% CI[3.89-4.24]) in the last round, p<0.0001, t-test).

Figure 5 plots the evaluation scores of the unseen cases grouped by speciality. From a qualitative point of view, we could not find any speciality that would consistently obtain scores below the others; outlier values correspond to cases where few dialogues were conducted.

Concerning the diversity of the vocabulary, unseen cases contained 1,488 types (unique word forms). 1,017 types (68.4%) appeared in isolated records; that means that only one third of the types (31.6%) occurred in more than one case. The average proportion of unique types per record is 34.6% (±7.4). Those numbers show to which extent the lexical content of each case differs across records in the unseen cases.

We also analysed the quantity of out-of-vocabulary words (OOVs) in unseen cases. Out of the total 1,488 types in the unseen cases, only 33 words (2.5%) were missing in system resources (avg = 1.2 OOVs per case, ±1.66). That is, our resources covered 97.5% of the vocabulary in the 29 new cases. Our analysis showed that most OOVs were spelling mistakes made when inputting data to create a new record. Our methods predicted the PoS category of these OOVs with a precision of 69.8%, a recall of 76.9%, and an F-measure of 73.2% (micro-average). Regarding the OOV words for which the system predicted the correct category, our methods to predict morphology data showed a precision of 59.4%, a recall of 61.3%, and an F-measure of 60.3% (micro-average). Table 7 (Appendix) shows further details about our results per category.

Lastly, Fig. 5 (bottom right) depicts differences in assessment according to the evaluators' profiles. The average scores of the majority or totality of evaluators agreed on user-understanding, quickness, tediousness and concision. Students and residents gave higher average scores to system performance, coherence of replies, informativeness, system- and user-understanding. Senior doctors generally gave lower scores.

Discussion

The quality of the natural language dialogue in seen and unseen cases received very positive, positive, or neutral

Table 5 Evaluation of system correctness expressed as average percentage (±standard deviation) [minimum - maximum]

	Per medical specialty	Per dialogue
Correct	74.3 (±9.5) [53.6–93.8]	74.9 (±12.6) [40.0–100.0]
Incorrect	14.9 (±6.3) [0.0–31.6]	14.7 (±9.4) [0.0–38.9]
Not understood	7.8 (±5.3) [0.0–25.0]	7.5 (±7.7) [0.0–40.0]
Request for repair	2.9 (±2.7) [0.0–11.5]	2.9 (±3.9) [0.0–20.0]

UNCORRECTED PROOF

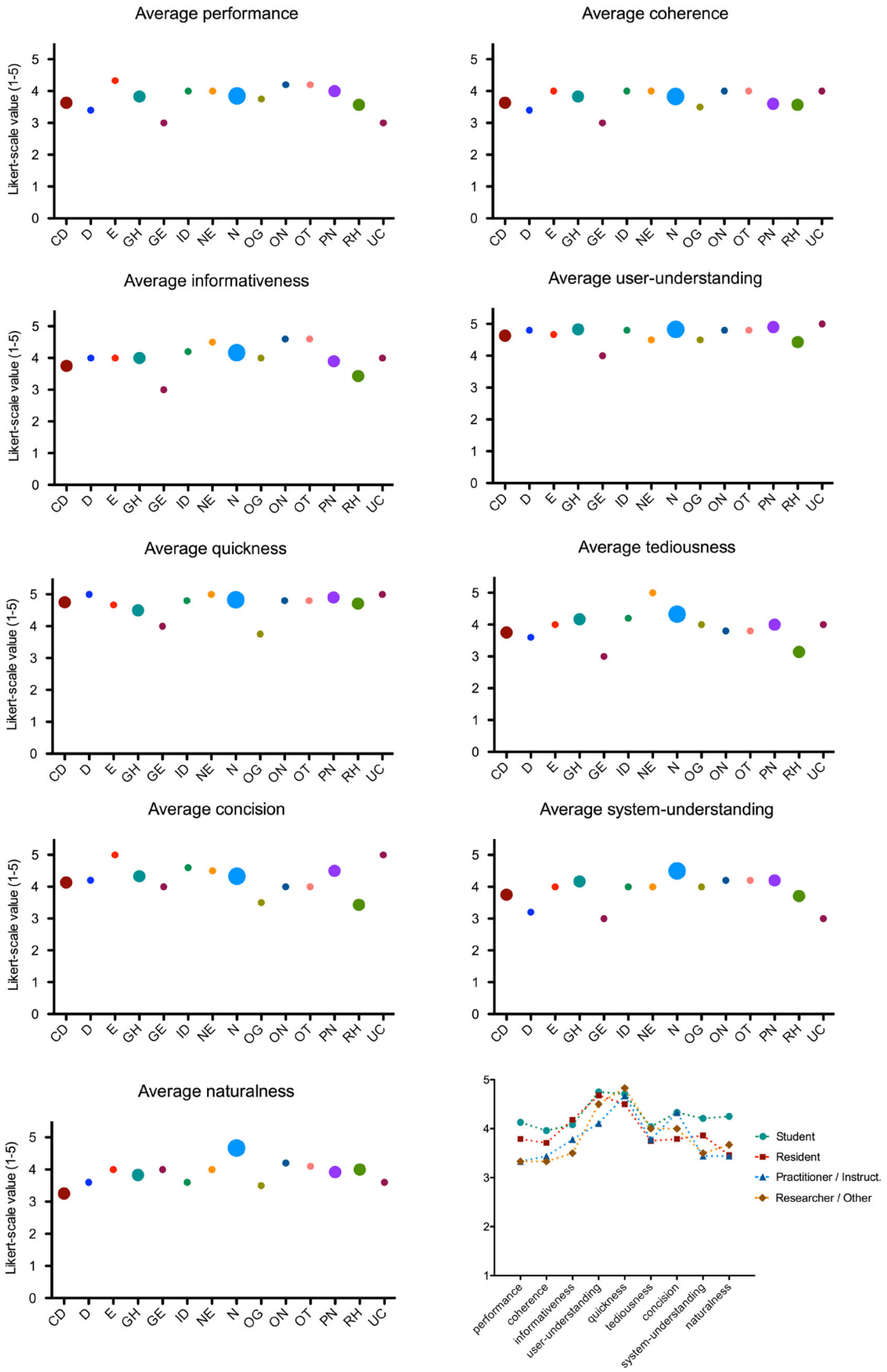


Fig. 5 Qualitative evaluation across medical specialities and evaluator profiles. The size of each point expresses the number of dialogues conducted: 1–5 (small size), 6–10 (medium size) and >10 (large size). The abbreviations of specialities are given in Table 1

AUTHOR'S PROOF

judgements from between 93% and 100% of the evaluators, allowing us to answer Q1 positively. System performance and coherence of replies received Good and Very good scores and overall satisfaction was high with an average of 3.84 (seen cases) and 4.06 (unseen cases) across all aspects. We cannot compare the error rate with other works (e.g. [34]) without bias, since we tested more patient cases.

Regarding Q2, in the test on unseen cases, every aspect received a higher user evaluation score than on seen cases. The improvement of some features proved statistically significant. The system was robust enough to cope with new cases without quality loss. The system's vocabulary coverage of unseen cases was very high (97.5%). Overall, we tested 35 different cases covering 18 medical specialities. To the best of our knowledge, this is much larger than what was reported so far in the literature.

The unseen cases covered varied medical specialities among which we could not highlight consistently less well-handled specialities from a qualitative point of view. To analyse this aspect from a quantitative perspective, a larger number of dialogues in each speciality would be needed. The comparison of scores across evaluators' profiles showed that medical students and residents evaluated the system better. This is a good point since they are the first targeted users of the system.

The correction rate of system replies varied across cases largely due to each record content: e.g. the performance was lower in a postpartum case, where some questions referred to the patient's newborn, but the system could not distinguish them from those related to the VP. Our analysis of logs across cases unveiled that most errors were due to the lack of variants of question formulations, missing question types, or processing errors (Table 6, Appendix). These weaknesses require fallback strategies, which we explored using machine learning [51].

At a technical level, we want to improve the performance of the dialogue manager and the comparison and update procedures. Given the lack of dialogue corpora for the task, we did not apply machine/deep learning approaches. Terminological components can mitigate the needs of the domain—rich in variant terms and acronyms, but without open training data available. This is the asset of our system. Once enough dialogue logs are collected via a rule- and terminology-based system, the data can be trained to complement the dialogue policy manager, or to generate word-embeddings for OOV terms. This is left for future work. The naturalness of system replies needs also refinement, especially the way it simplifies long sentences or outputs negative symptoms and layman terms. We are interested in evaluating the system in the overall framework of a simulated consultation, where medical students should diagnose the patient. This would allow us to know whether the system helps students to obtain all key elements of the history-taking step, and to ascertain whether students make a correct diagnosis. Finally, we need to gather dialogue data to evaluate the English and Spanish versions.

Lessons learned

Regarding development, several aspects demanded a heavy investment in resource creation: terminology components for concept mapping, update procedures to compare the existing knowledge base and OOVs, and linguistically-motivated modules to transform the data created by medical trainers according to the patient's perspective. Moreover, misspellings in trainers' input needed spelling correction tools. To fix the OOV errors related to spelling mistakes, the most reliable approach would be to include a correction module on the back-office interface that trainer doctors use to create the patient record. The system vocabulary could be mapped to misspellings, flag them, and the trainers could correct them before the interaction. Nevertheless, the developed modules were capable of adapting the system to new cases without causing problematic interactions, according to the end-user evaluation.

Regarding the system design and evaluation, we strongly advise that medical professionals be involved from the beginning. The closer to reality the patient data we received, the better the system was tested and improved. The more iterations were conducted for inspecting logs and fixing errors, the better the system was rated. Our evaluation revealed that experienced practitioners assessed the system as less satisfactory, given their greater diagnosis experience and different perception of these tools. This highlights the careful choice of the end-user and its impact on the framework design. This multi-case, adaptable VP system seems to fit medical students and interns, since they can bear infelicities in system replies and need to engage in the interaction to gain experience. A tool with canned answers would be rigid and necessitate more engineering to adapt to new cases. If no dialogue data are available for the task, collecting dialogue logs with potential end-users seems feasible before data-intensive methods (machine or deep learning) can be applied. Finally, this system is not yet suited for simulating VPs with chronic conditions needing follow-up consultations. Evolving symptoms would require a more advanced model of the VP's disease timeline.

Overall, the tradeoff between adaptability and naturalness has design implications related to immediate vs long-term needs, or sophisticated case-specific vs generic applications. Table 12 (Appendix) outlines our observations.

UNCORRECTED PROOF

AUTHOR'S PROOF

JrnlID 10916_ArtID 1737_Proof#1 - 19/04/2021

Conclusion

Medical doctors need to master medical history taking and these abilities may be enhanced through practice by using software simulations. To complement the direct contact with patients, we proposed a dialogue system for simulating the interview with multiple virtual patient cases. Because this system features interaction through natural language, it provides favourable conditions to improve medical students' anamnesis skills. We reported here the usability evaluation of the French system. We assessed to which extent it is mature enough in a real use context.

The agent was tested with 35 different cases from 18 different specialities. Medical evaluators considered that this system provides quality dialogue through natural language, that it does so across heterogeneous cases and medical specialities, and that it processes new records without quality loss compared to already known cases. Our usability evaluation showed that this multi-case system can support student training in history taking and provided us with lessons we thought useful to share regarding its strengths and limits.

Appendix

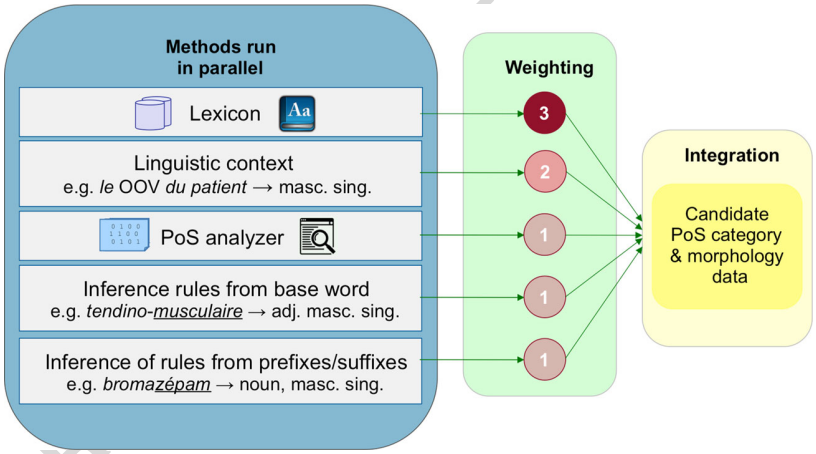


Fig. 6 Interface to input data to create a new virtual patient record

Table 6 Examples of correct, incorrect and deferred replies (*I*: 'input'; *R*: 'system reply'); we show the English translation of dialogue interactions using the French system

Correct	The information is accurate with regard to the patient record (*correctness*)	I: Since when have you had these symptoms? R: I have been nauseous for 2 weeks
	No information is invented (*faithfulness*)	I: Do your parents suffer from any disease? R: I can not answer this question (This information is not available in the patient record)
	No data are omitted (*exhaustiveness*)	I: Do you smoke? R: I smoke one packet a day since I am 18 years old. I want to stop.
Deferred or	Context information	I: How did it start?

UNCORR

AUTHOR'S PROOF

Table 6 (continued)

Request for repair	unprocessed	R: What do you refer to? Could you give more details?
	Declarative statement	R: I have a dry coughing since yesterday evening at 23 h I: That is very precise! R: Excuse me?
	Joint questions or dialogue acts	I: Yes, I see you have difficulty breathing but since when are you feeling like that? R: What do you refer to? Could you give more details?
	Processing error	I: Do you have coagulation problems? R: I do not understand what you are asking for
	Question type does not exist	I: Have you lifted something too heavy? R: Excuse me?
	Spelling	I: Poid ? (poids) '(your) weight?' R: Pardon ? 'Excuse me?'
	Term missing in vocabulary list	I: Do you limp? R: I do not understand what you are asking for
	Variation of query phrasing	I: Avez-vous un concubin? 'Do you have a partner?' R: Pardon ? 'Excuse me ?' (The system replies correctly to a synonym question: 'Avez-vous un couple?')
Incorrect	Context information unprocessed	I: Do you have a history of diseases? R: Bronchitis I: Frequently? R: Yes, bronchitis
	Question type does not exist	I: How much sport do you do each week? R: Biking, and I go swimming occasionally
	Processing error	I: Never had a surgery procedure? R: I have never had that type of surgery
	Spelling	I: Avez-vous des maux de tete ? (tête, 'head') 'Do you have headaches?' R: J'ai des coliques, j'ai extrêmement mal au ventre 'I have colics, I have a strong stomach ache'
	Term missing in vocabulary list	I: Do you often eat fast food? R: I do not follow any special diet
	Variation of query phrasing	I: Rhythm of the pain? R: I felt a big pain in my right side (The system replies correctly to the question: 'How often do you have pain?')

UNCORRECTED PROOF

AUTHOR'S PROOF

Table 7 Results of prediction methods of part-of-speech (PoS) category and morphology data for out-of-vocabulary (OOV) words (in percentage); the number of instances per class is shown in brackets; results of morphology data were only computed on OOVs for which the PoS category was predicted correctly

PoS category	P	R	F1	Morphology data	P	R	F1
Adjective (9)	100.0	66.7	80.0	Conjugation data (3)	25.0	33.3	28.6
Adverb (1)	100.0	100.0	100.0	Feminine singular (6)	80.0	66.7	72.7
Noun (20)	75.0	90.0	81.8	Feminine plural (2)	0.0	0.0	0.0
Preposition (2)	100.0	100.0	100.0	Masculine singular (15)	77.8	93.3	84.8
Proper name (3)	0.0	0.0	0.0	Masculine plural (2)	0.0	0.0	0.0
Verb (4)	30.0	75.0	42.9	No inflection (3)	0.0	0.0	0.0
Micro-average	69.8	76.9	73.2	Micro-average	59.4	61.3	60.3

Table 8 Analysis of incorrect replies with examples (*I*: 'user input'; *R*: 'system reply'); we show the English translation of dialogue interactions using the French system

Severe errors (26.5%)	Answers that give misleading information (including those in coherent replies). These errors have an impact on systems' *faithfulness*. It would be preferable to request for repair or reply *I have not understood*
Example:	I: Have you lost consciousness? R: Yes [Error due to an incorrect match between 'lose consciousness' (input) and 'sight loss' (record)]
Moderate errors (56.2%)	Replies that are incoherent regarding the user question, but provide correct information from the patient record; or when the system should reply that no information is available in the record. Because of the incoherence, the user might not pay attention to the information and would reformulate the question or change the topic. These have an impact on the *dialogue flow*. Especially, when no data are available, the user should be informed to stop reformulating questions on the same topic. These errors could be more harmful if the user changes the topic without having found key information from the record.
Example:	I: And does your wife have any symptom? R: I am married [The system should reply that details are not available in the record]
Errors that impact the exhaustiveness of information	Answers that do not provide all the information in the patient record. The severity would depend on each dialogue context and on the design strategy of the agent; i.e., whether the system should always provide all the information (as in our system), or let the user be trained in asking for further details.
Example:	I: Where do you have pain? R: Yes [The virtual patient should reply: 'I have thorax pain']

UNCORRECTED PROOF

AUTHOR'S PROOF

Table 9 Sample clinical record (top) and sample of the output for OOV words in a new VP record (bottom); *adj* stands for 'adjective'; *fp*, for 'feminine plural'; the format is YAML

Sample clinical record

```
aimOfConsultation:
     aim: the patient is consulting you about persistent backache.
informations:
     patientFirstName: Patricia
     patientLastName: Hurst
     patientAge: 65
     maritalStatus: single
     profession: accountant
     children: none
     weight: 72 kilograms
     height: 162 centimetres
lifestyle:
     food:
       items:
         - the patient often eats fish and chips; the patient hates vegetables
     physicalActivity:
       items:
         - the patient goes to country and western dance club twice a week
     addictions:
       items:
          - the patient drinks about two pints of dark beer every day.
      socialBehaviour:
       items:
         - the patient lives alone but often spends time with her family
medicalRecord:
     allergies:
       nonmedicationAllergy:
         - allergy: tree pollen
         observationsValue: the patient is allergic to many types of tree pollen
     medicalHistory:
       - disease: stomach ulcers
         durationValue: for 8 years
         treatment:
           - therapeuticClassValue: proton pump inhibitor (omeprazole)
     surgery:
       - operation: the patient had a broken leg and a dislocated knee
         age: at the age of three
         observationsValue: the patient has a slight limp
complaints:
      - symptom: pain in the lower back
       observationsValue: the pain is in the lower back and sometimes down the legs
       durationValue: for months
      - symptom: the patient has a pain that disrupts sleep
       frequencyValue: often
```

UNCORRECTED PROOF

AUTHOR'S PROOF

Table 9 (continued)

Sample clinical record

```
      observationsValue: the pain often makes it difficult to fall asleep
currentTreatment:
      - therapeuticClassValue: proton pump inhibitor
        methodOfAdministrationValue: oral
        frequencyValue: three times a day
        observationsValue: the patient used to be on esomeprazole magnesium
      - therapeuticClassValue: pain-killer
        methodOfAdministrationValue: oral
        doseValue: 1 gram
        frequencyValue: 3 a day
        observationsValue: the patient's pain is not relieved
```

Linguistic data output for OOV words in a new VP record

```
symptoms:
   token: insomniantes
   lemma: insomniant
   data:
     cat: adj
     mor: fp
     string:
       douleurs parfois insomniantes ('pain often causing insomnia')
```

CORRECTED PROOF

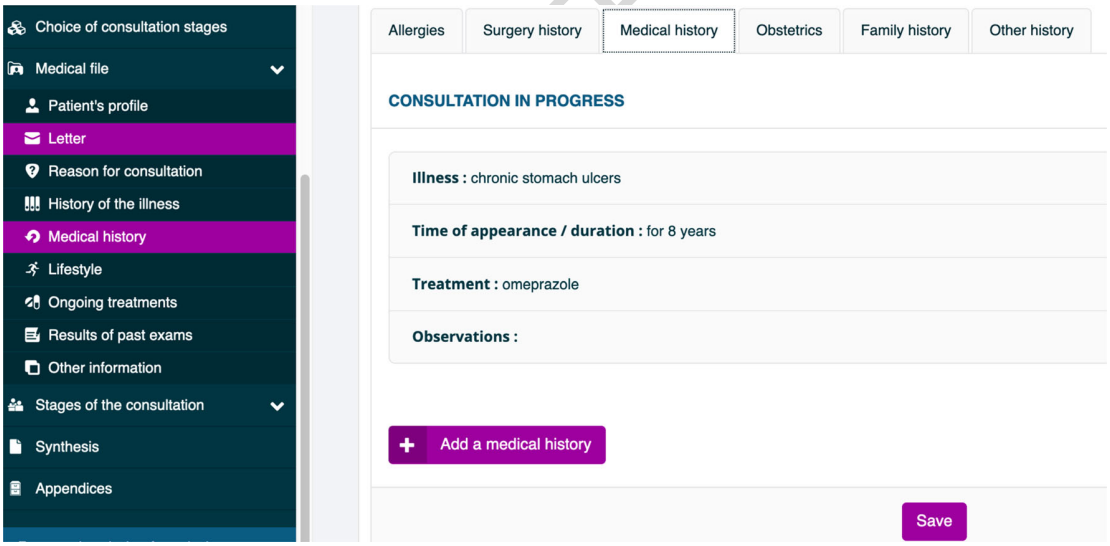


Fig. 7 Procedures and weighting scheme to predict linguistic information for OOV items

AUTHOR'S PROOF

Table 10 Description of the seen cases used in the usabi/lity study

Description	Diagnosis	Spec.
A 41-year-old woman comes for a pre-anesthesia checkup before a gallbladder surgery.	NA	AN
A 41-year-old man comes for a medical certificate for a sport competition.	CD Essential hypertension	
A 49-year-old man consults about a violent thoracic pain since last night.	Pneumopathy	PN
A 35-year-old man complains of a considerable fatigue and weight loss.	Depressive episode	P
A 40-year-old woman complains of a sore throat.	Throat infection	GP
A 49-year-old man consults about urinary problems.	Prostatic hyperplasia	U

Abbreviations of medical specialities (Spec.) are given in Table 1; *NA* stands for *not available* (no diagnosis): not all consultations lead to a diagnosis (e.g., pre-anesthesia checkup), and some cases only contained the case description for the dialogue system, without further training feedback

Table 11 Description of the unseen cases used in the usability study

Description	Diagnosis	Spec.
A 57-year-old man comes for a medical check-up after an episode of cardiac insufficiency.	Cardiac insufficiency	CD
A 64-year-old man consults because he had a myocardial infarction.	Extended anterior myocardial infarction	CD
A 65-year-old man consults for a thigh wound that developed progressively	Psoriasis	D
A 27-year-old woman complains of diarrhoea, hot flushes and palpitations. for one year.	Thyroid disorders	E
A 70-year-old woman consults for knee pain.	Knee osteoarthritis	GE
A 29-year-old man consults for a disabling diarrhoea and increasing tiredness.	NA	GH
A 60-year-old man consults for epigastric pain.	Chronic gastroesophageal reflux	GH
A 56-year-old man complains of weight loss and abdominal pain.	NA	GH
A 31-year-old woman has been having abdominal pain within the last 24 h.	Mesenteric adenitis	GH
A 78-year-old man consults for bloody stools and loss of appetite.	NA	GH
A 24-year-old woman consults for pains in her lower abdomen and foul-smelling vaginal discharge.	Sexually transmitted disease	IT
A 24-year-old man consults for hair loss and a rash on his feet.	Syphilis	IT
A 24-year-old woman has been having gait problems and tingling recently.	Multiple sclerosis	N
A 32-year-old woman has been suffering from regular headaches over the last year.	Migraine	N
A 70-year-old man has suffered a sudden vision loss.	Cerebrovascular accident	N
A 28-year-old woman has suffered a progressive vision loss.	Possible multiple sclerosis	N
A 67-year-old man comes with alteration of the general state, left lumbar pain and vomiting.	Renal Insufficiency	NE
A 66-year-old woman complains of vaginal bleeding.	NA	OG
A 32-year-old woman gave birth two months ago and feels very tired.	Postpartum depression	OG
A 25-year-old woman complains of right leg pain and a fever.	Phlebitis	OG
A 59-year-old man comes to a follow-up consultation for a multiple myeloma.	Multiple myeloma	ON
A 71-year-old man complains of difficulty swallowing over the past months.	Possible oesophageal cancer	OT
A 66-year-old man complains of shortness of breath on any exertion.	NA	PN

UNCORRECTED PROOF

AUTHOR'S PROOF

Table 11 (continued)

Description	Diagnosis	Spec.
A 21-year-old woman has suffered an episode of respiratory distress on effort.	NA	PN
A 55-year-old man consults for coughing, often with blood-tainted sputum.	NA	PN
A 37-year-old man complains of coughing with sputum and shortness of breath.	Bronchitis	PN
A 60-year-old man complains of a back pain that does not go away.	Persistent sciatica	RH
A 57-year-old man presents with a back pain started suddenly 5 days ago.	Acute lumbar sciatica	RH
A 55-year-old woman comes into urgent care with a fever and abdominal pain.	Cholecystitis	UC

Abbreviations of medical specialities (Spec.) are given in Table 1. *NA* stands for *not available* (no diagnosis)

Table 12 Summary of lessons learned from the development and usability evaluation and implications on design and development

Design	• Create a patient record model for the medical trainers to input the virtual patient's health state in a semistructured template • Devise a knowledge model for the task: range of question types, dialogue acts and entity types concerning history taking • Conceive the appropriate dialogue strategies: – Careful fallback replies when user's question is not in the patient record or it is out-of-scope or out-of-domain – Accurate information regarding the patient record (*correctness*), without inventing information (*faithfulness*) nor omitting data (*exhaustiveness*) – And all the above, in a *dynamic dialogue flow*: maximising user engagement in interaction and minimising tiredness or boredom • Outline the end-users' profile (students, interns or experienced practicing doctors) • Analyse the users' needs in order to balance the trade-off between generalisability (adaptable system) and specialisation (a tailored, engineered application for a specific case or a medical specialty)
Development	• Invest in creating termino-ontologic resources: – Terminology modules for concept mapping and term variation – Components to compare the existing knowledge base, detect out-of-vocabulary words in new cases and update system resources – Linguistically-motivated modules to change the patient record from the input description to patient's perspective (3rd to 1st person) – Term simplification modules to map technical to laymen words – Spelling correction tools • Minimise human intervention or engineering needs to adapt the system to unseen cases on-the-fly • Have medical professionals involved from the start of the project • If no training dialogue data are available, collect dialogue logs simulating the task with real end-users via a rule-based and terminology-based system, crowdsourcing, or a wizard-of-oz protocol
Evaluation	• Get close-to-reality patient cases to simulate a wide range of virtual patient profiles (e.g. medical transcripts or cases prepared by medical trainers and aimed at medical students) • Conduct tests by real end-users as soon as possible • Iteratively inspect patient logs to detect and fix dialogue errors before each evaluation round • Warn the users about the system limitations (what it can do and it cannot do)

UNCORRECTED PROOF

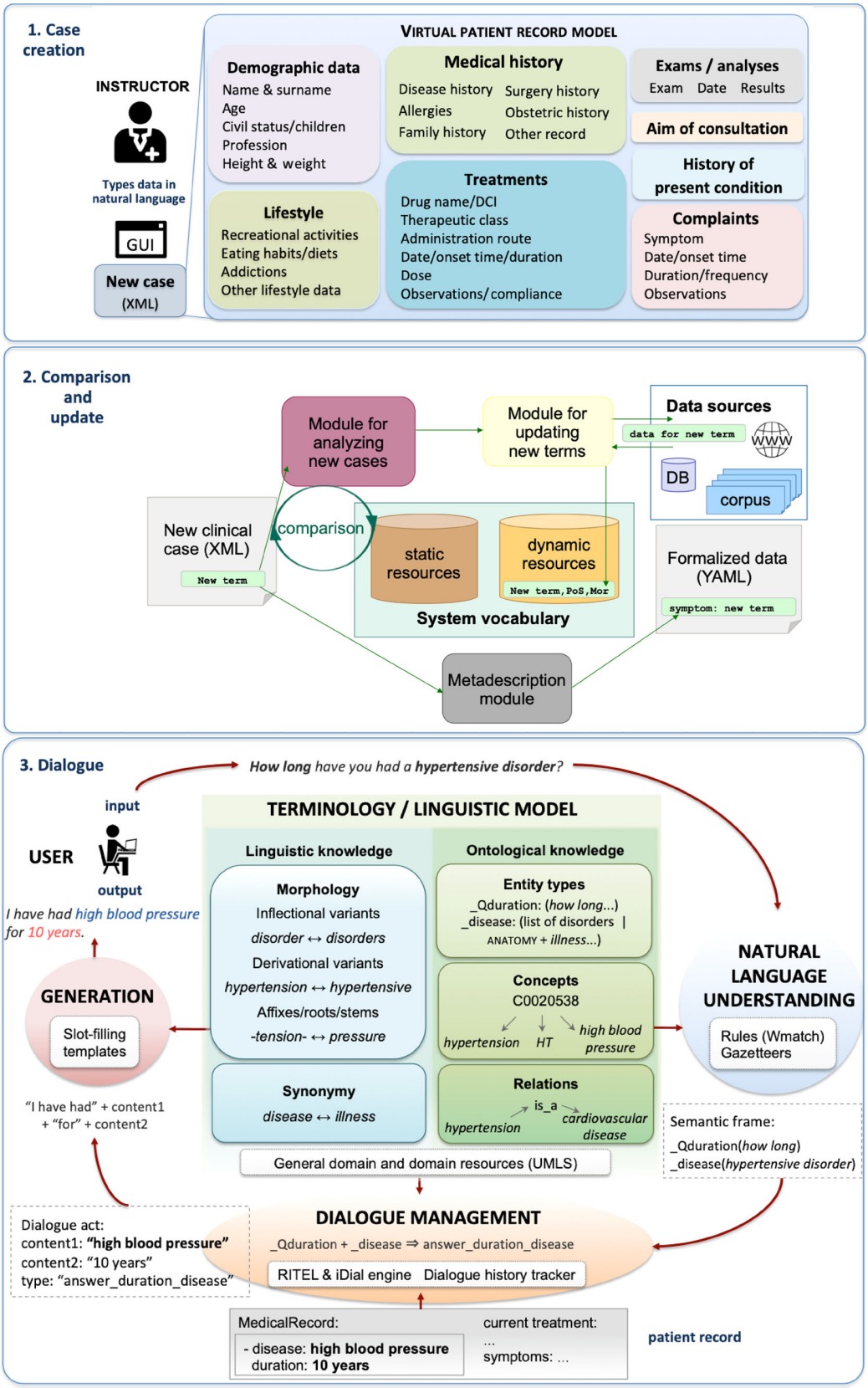


Fig. 8 Overall functioning of the dialogue system and update components; further technical details are provided in [27, 46, 47]

AUTHOR'S PROOF

Acknowledgements We greatly thank all doctors who evaluated the system and gave valuable remarks, and also Dr. Aurélie Névéol for her helpful comments on the manuscript. We developed the dialogue system in a collaborative project led by Interaction Healthcare and having as partners VIDAL, Angers University Hospital, Voxygen and LIMSI.[4]

Author contributions Sophie Rosset (SR), Leonardo Campillos-Llanos (LC) and Catherine Thomas (CT) developed the VP dialogue system, and Pierre Zweigenbaum (PZ) contributed to the medical terminology components and patient record model. Éric Bilinski (EB) implemented the web evaluation tool and the online demonstration of the dialogue system. Antoine Neuraz (AN) helped to engage the evaluation participants and made valuable remarks about the system and article. SR and PZ designed the evaluation protocol, and LC collected and analysed the evaluation data. LC and SR double-checked a subset of the data. LC, SR and PZ wrote the manuscript, and all authors read and approved the final article.

Funding This work was funded by BPI (FUI Project PatientGenesys, F1310002-P) and by the Société d'Accélération de Transfert Technologique (SATT) Paris Saclay (PVDial project). The funding bodies did not take part in the design of the study, analysis and interpretation of data and writing the manuscript.

Data availability The dialogue data collected during development and evaluation is available at: https://pvdial.limsi.fr/data/PG-logs-eval.zip A demonstration of the dialogue system can be tested at: http://vps-9069f76a.vps.ovh.net

Code availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Washburn, M., Bordnick, P., and Rizzo, A. S., A pilot feasibility study of virtual patient simulation to enhance social work students' brief mental health assessment skills. *Soc. Work Health Care* 55(9):675–693, 2016.
2. Barnett, S. G., Gallimore, C. E., Pitterle, M., and Morrill, J., Impact of a paper vs virtual simulated patient case on student-perceived confidence and engagement. *Am. J. Pharm. Educ.* 80(1):16, 2016.
3. McCoy, L., Pettit, R. K., Lewis, J. H., Allgood, J. A., Bay, C., and Schwartz, F. N., Evaluating medical student engagement during virtual patient simulations: A sequential, mixed methods study. *BMC Med. Educ.* 16:20, 2016.
4. Tait, L., Lee, K., Rasiah, R., Cooper, J. M., Ling, T., Geelan, B., and Bindoff, I., Simulation and feedback in health education: A mixed methods study comparing three simulation modalities. Pharmacy (Basel) 6(2), 2018.
5. Courteille, O., Fahlstedt, M., Ho, J., Hedman, L., Fors, U., von Holst, H., Fellander-Tsai, L., and Moller, H., Learning through a virtual patient vs. recorded lecture: A comparison of knowledge retention in a trauma case. *Int. J. Med. Educ.* 9:86–92, 2018.
6. Gupta, A., Singh, S., Khaliq, F., Dhaliwal, U., and Madhu, S. V., Development and validation of simulated virtual patients to impart early clinical exposure in endocrine physiology. *Adv. Physiol. Educ.* 42(1):15–20, 2018.
7. de Cock, C., Milne-Ives, M., van Velthoven, M. H., Alturkistani, A., Lam, C., and Meinert, E., Effectiveness of conversational agents (virtual assistants) in health care: Protocol for a systematic review. *JMIR Res. Protoc.* 9(3):e16934, 2020.
8. Ellaway, R., Candler, C., Greene, P., and Smothers, V., An architectural model for MedBiquitous virtual patients. http://groups.medbiq.org/medbiq/display/VPWG/MedBiquitous+Virtual+Patient+Architecture. Accessed: 8 Dec 2018, 2006.
9. Sijstermans, R., Jaspers, M. W., Bloemendaal, P., and Schoonderwaldt, E., Training inter-physician communication using the dynamic patient simulator®. *Int. J. Med. Inf.* 76(5–6):336–343, 2007.
10. Danforth, D. R., Procter, M., Chen, R., Johnson, M., and Heller, R., Development of virtual patient simulations for medical education. *J. Virtual Worlds Res.* 2(2):4–11, 2009.
11. Rombauts, N., Patients virtuels: pédagogie, état de l'art et développement du simulateur Alphadiag. PhD dissertation, Faculty of Medicine, Claude Bernard University, Lyon France, 2014.
12. Menendez, E., Balisa-Rocha, B., Jabbur-Lopes, M., Costa, W., Nascimento, J. R., Dósea, M., Silva, L., and Junior, D. L., Using a virtual patient system for the teaching of pharmaceutical care. *Int. J. Med. Inf.* 84(9):640–646, 2015.
13. Lin, C. J., Pao, C. W., Chen, Y. H., Liu, C. T., and Hsu, H. H., Ellipsis and coreference resolution in a computerized virtual patient dialogue system. *J. Med. Syst.* 40(9):206–221, 2016.
14. Laleye, F. A., Blanié, A., Brouquet, A., Behnamou, D., and de Chalendar, G., Semantic similarity to improve question understanding in a virtual patient. In: *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pp. 859–866, 2020.
15. Chen, F., Lee, Y., and Hubal, R., Work-in-progress—testing of a virtual patient: Linguistic and display engagement findings. In: *2020 6th International Conference of the Immersive Learning Research Network (iLRN)*, pp. 348–350: IEEE, 2020.
16. Candler, C., Effective use of educational technology in medical education. In: *Colloquium on Educational Technology: Recommendations and Guidelines for Medical Educators*, pp. 1–19. Washington, DC: AAMC Institute for Improving Medical Education, 2007.
17. Schmidlen, T., Schwartz, M., DiLoreto, K., Kirchner, H. L., and Sturm, A. C., Patient assessment of chatbots for the scalable delivery of genetic counseling. *J. Genet. Couns.* 28(6):1166–1177, 2019.
18. Chetlen, A., Artrip, R., Drury, B., Arbaiza, A., and Moore, M., Novel use of chatbot technology to educate patients before breast biopsy. *J. Am. Coll. Radiol.* 16(9 Pt B):1305–1308, 2019.
19. Kokciyan, N., Chapman, M., Balatsoukas, P., Sassoon, I., Essers, K., Ashworth, M., Curcin, V., Modgil, S., Parsons, S., and Sklar, E. I., A collaborative decision support tool for managing chronic conditions. *Stud. Health Technol. Inform.* 264:644–648, 2019.
20. Cook, D. A., Erwin, P. J., and Triola, M. M., Computerized virtual patients in health professions education: A systematic review and meta-analysis. *Acad. Med.* 85(10):1589–1602, 2010. https://doi.org/10.1097/ACM.0b013e3181edfe13.
21. Wattanasoontorn, V., Hernández, R. J. G., and Sbert, M., Embodied conversational virtual patients. In: Diana, P. M., and Nieto, I. P. (Eds.) *Conversational Agents and Natural Language*

[4] https://pvdial.limsi.fr

UNCORRECTED PROOF

AUTHOR'S PROOF

Interaction: Techniques and Effective Practices, pp. 254–281. Hershey: Information Science Reference, IGI Global, 2011. https://doi.org/10.4018/978-1-60960-617-6.ch011.

22. Rossen, B., and Lok, B., A crowdsourcing method to develop virtual human conversational agents. *Int. J. Hum. Comput. Stud.* 70(4):301–319, 2012.
23. Lelardeux, C., Panzoli, D., Alvarez, J., Galaup, M., and Lagarrigue, P., Serious game, simulateur, serious play: État de l'art pour la formation en santé. In: *Actes du colloque Serious Games en Médecine et Santé (SeGaMED) 2013*, pp. 27–38. Nice: e-virtuoses, 2013.
24. Wattanasoontorn, V., Hernández, R. J. G., and Sbert, M., Serious games for e-health care. In: Cai, Y., and Goei, S. (Eds.) *Simulations, Serious Games and Their Applications*, pp. 127–146. Singapore: Springer, 2014. https://doi.org/10.1007/978-981-4560-32-0_9.
25. Reiswich, A., and Haag, M., Evaluation of chatbot prototypes for taking the virtual patient's history. *Stud. Health Technol. Inform.* 260:73–80, 2019.
26. Nirenburg, S., Beale, S., McShane, M., Jarrell, B., and Fantry, G., Language understanding in Maryland virtual patient. In: *Proceedings of the International Conference on on Computational Linguistics*, pp. 36–39. Manchester: Citeseer, 2008.
27. Campillos-Llanos, L., Bouamor, D., Bilinski, É., Ligozat, A. L., Zweigenbaum, P., and Rosset, S., Description of the PatientGenesys dialogue system. In: *Proceedings of SIGDIAL*, pp. 438-440. Prague: Association for Computational Linguistics, 2015.
28. Leuski, A., and Traum, D., Practical language processing for virtual humans. In: *Proceedings on Innovative Applications of Artificial Intelligence Conference*, pp. 1740–1747. Atlanta, 2010.
29. Rizk, Y., Kshoury, K., Chehab, M., Chidiac, P., Awad, M., and Antoun, J., Virtual patient. In: *Proceedings of WINLP*, pp. 1–3. Vancouver, 2017.
30. Datta, D., Brashers, V., Owen, J., White, C., and Barnes, L. E., A deep learning methodology for semantic utterance classification in virtual human dialogue systems. In: Traum, D., Swartout, W., Khooshabeh, P., Kopp, S., Scherer, S., and Leuski, A. (Eds.) *Intelligent Virtual Agents, Los Angeles*, pp. 451–455. Berlin: Springer, 2016.
31. Jin, L., White, M., Jaffe, E., Zimmerman, L., and Danforth, D., Combining cnns and pattern matching for question interpretation in a virtual patient dialogue system. In: *Proceedings on Workshop Innovative Use NLP Building Educational Applications*, pp. 11–21: Copenhagen, 2017.
32. Dickerson, R., Johnsen, K., Raij, A., Lok, B., Hernandez, J., Stevens, A., and Lind, D. S., Evaluating a script-based approach for simulating patient-doctor interaction. In: *Proceedings of the International, Conference on Human-Computer Interface Advances Modeling and Simulation*, pp. 79–84. New Orleans, 2005.
33. Pence, T. B., Dukes, L. C., Hodges, L. F., Meehan, N. K., and Johnson, A., The effects of interaction and visual fidelity on learning outcomes for a virtual pediatric patient system. In: *IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 209–218. Philadelphia: IEEE, 2013. https://doi.org/10.1109/ICHI.2013.36.
34. Maicher, K., Danforth, D., Price, A., Zimmerman, L., Wilcox, B., Liston, B., Cronau, H., Belknap, L., Ledford, C., Way, D. et al., Developing a conversational virtual standardized patient to enable students to practice history-taking skills. *Simul. Healthc.* 12(2):124–131, 2017. https://doi.org/10.1097/SIH.0000000000000195.
35. Talbot, T. B., Sagae, K., John, B., and Rizzo, A. A., Sorting out the virtual patient: How to exploit artificial intelligence, game technology and sound educational practices to create engaging role-playing simulations. *Int. J. Gaming Comput. Mediat. Simul.* 4(3):1–19, 2012. https://doi.org/10.4018/jgcms.2012070101.
36. Scherly, D., and Nendaz, M., Simulation du raisonnement clinique sur ordinateur: Le patient virtuel. In: Boet, S., Granry, J., and Savoldelli, G. (Eds.) *La Simulation en Santé. De la Théorie à la Pratique*, pp. 43–50. Paris: Springer, 2013. https://doi.org/10.1007/978-2-8178-0469-9_5.
37. Hubal, R. C., Kizakevich, P. N., Guinn, C. I., Merino, K. D., and West, S. L., The virtual standardized patient. *Stud. Health Technol. Inform.* 70:133–138, 2000.
38. Stevens, A., Hernandez, J., Johnsen, K., Dickerson, R., Raij, A., Harrison, C., DiPietro, M., Allen, B., Ferdig, R., Foti, S. et al., The use of virtual patients to teach medical students history taking and communication skills. *Am. J. Surg.* 191(6):806–811, 2006.
39. Kenny, P., Rizzo, A. A., Parsons, T. D., Gratch, J., and Swartout, W., A virtual human agent for training novice therapists clinical interviewing skills. *Annu. Rev. CyberTherapy Telemed.* 5:77–83, 2007. https://doi.org/10.1145/159544.159587.
40. Kenny, P., Parsons, T. D., Gratch, J., and Rizzo, A. A., Evaluation of Justina: A virtual patient with PTSD. In: Prendinger, H., Lester, J., and Ishizuka, M. (Eds.) *Intelligent Virtual Agents*, pp. 394–408. Berlin: Springer, 2008.
41. Parsons, T. D., Virtual standardized patients for assessing the competencies of psychologists. In: *Encyclopedia of Information Science and Technology, 3rd edn*, pp. 6484–6492: IGI Global, 2015. https://doi.org/10.4018/978-1-4666-5888-2.ch637.
42. Persad, A., Stroulia, E., and Forgie, S., A novel approach to virtual patient simulation using natural language processing. *Med. Educ.* 50(11):1162–1163, 2016. https://doi.org/10.1111/medu.13197.
43. Gokcen, A., Jaffe, E., Erdmann, J., White, M., and Danforth, D., A corpus of word-aligned asked and anticipated questions in a virtual patient dialogue system. In: *LREC International Conference on Language Resources and Evaluation*, pp. 3174–3179. Portorož, 2016.
44. Talbot, T. B., Kalisch, N., Christoffersen, K., Lucas, G., and Forbell, E., Natural language understanding performance and use considerations in virtual medical encounters. *Stud Health Technol. Inform.* 220:407–413, 2016.
45. Leleu, J., Caillat-Grenier, R., Pierard, N., Rica, P., Granry, J. C., Lehousse, T., Pereira, S., Bretier, P., Rosec, O., Bilinski, É., Bouamor, D., Campillos-Llanos, L., Grau, B., Ligozat, A. L., Zweigenbaum, P., and Rosset, S., Patient Genesys: Outil de création de cas cliniques de simulation médicale proposant des cas patients virtuels en 3D. In: *Applications Pratiques de l'Intelligence Artificielle*, p. 2. Rennes, 2015.
46. Campillos-Llanos, L., Bouamor, D., Zweigenbaum, P., and Rosset, S., Managing linguistic and terminological variation in a medical dialogue system. In: *LREC International Conference on Language Resources and Evaluation*, pp. 3167–3173. Portorož, 2016.
47. Campillos-Llanos, L., Thomas, C., Bilinski, É., Zweigenbaum, P., and Rosset, S., Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation. *Nat. Lang. Eng.* 26(2):183–220, 2020.
48. Bodenreider, O., The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* 32(suppl 1):D267–D270, 2004.
49. Dybkjær, L., and Bernsen, N. O., Usability evaluation in spoken language dialogue systems. In: *Proceedings of Workshop on Evaluation for Language and Dialogue Systems*, pp. 9–18: Association for Computational Linguistics, 2001.
50. Duplessis, G. D., Letard, V., Ligozat, A. L., and Rosset, S., Purely corpus-based automatic conversation authoring. In: *LREC*

UNCORRECTED PROOF

AUTHOR'S PROOF

International Conference on Language Resources and Evaluation, pp. 2728–2735. Portorož, 2016.

51. Campillos-Llanos, L., Rosset, S., and Zweigenbaum, P., Automatic classification of doctor-patient questions for a virtual patient record query task. In: *Proceedings of BioNLP*, pp. 333–341. Vancouver: Association for Computational Linguistics, 2017.

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

UNCORRECTED PROOF