



HAL
open science

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.



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

Q1 Leonardo Campillos-Llanos^{1,2} · Catherine Thomas³ · Éric Bilinski³ · Antoine Neuraz³ · Sophie Rosset³ · Pierre Zweigenbaum³

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

0 Introduction

1 Developing diagnosis and clinical reasoning skills is a
 2 key element of medical education. In addition to clinical

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

Q2 ✉ 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

3 practice, medical students and practitioners can enhance
 4 these abilities by means of mannequins, role games and
 5 simulation systems. These have shown beneficial results
 6 [1–7] and are currently integrated in virtual patients [8–
 7 15]. Virtual patients (VPs)¹ are software through which
 8 students can train themselves by emulating the roles of
 9 health providers [16].

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

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

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

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

42 Objectives

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

54 Accordingly, a first requirement of the system is to
 55 cope with new contents across medical specialities. The
 56 second requirement is to provide unconstrained input,
 57 because the system aims at improving medical students'
 58 history-taking skills through the interaction with the VP.
 59 Figure 1 is a sample dialogue and illustrates natural dialogue
 60 phenomena. The system is integrated in a serious game
 61 developed with partner companies and a medical team
 62 [45]. The software features an animated avatar with text-to-
 63 speech, lip-synch and minor gestures.

64 To make the system able to handle plenty of cases, we
 65 gave it extensive conceptual and terminological coverage
 66 of the domain [27, 46]. The system can also adapt to
 67 new records dynamically. We provided it with components
 68 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 process-
 ing 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>

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



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 stomach ulcers
 D: Do you follow any treatment for that?
 P: Omeprazole for the chronic stomach ulcers. Maybe there is more information in my file

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

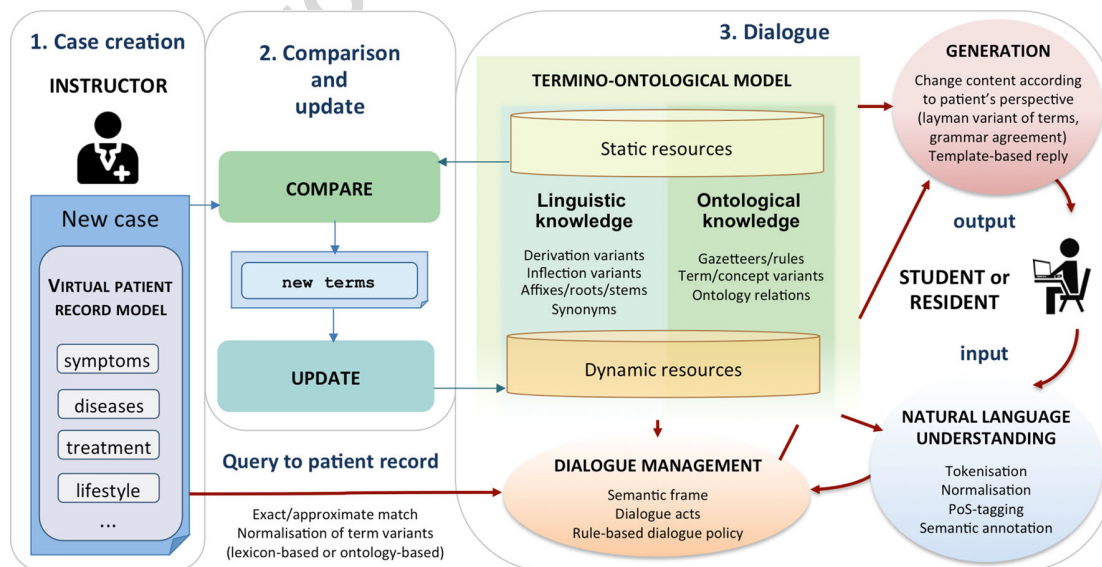


Fig. 2 Schema of the virtual patient dialogue system and update components

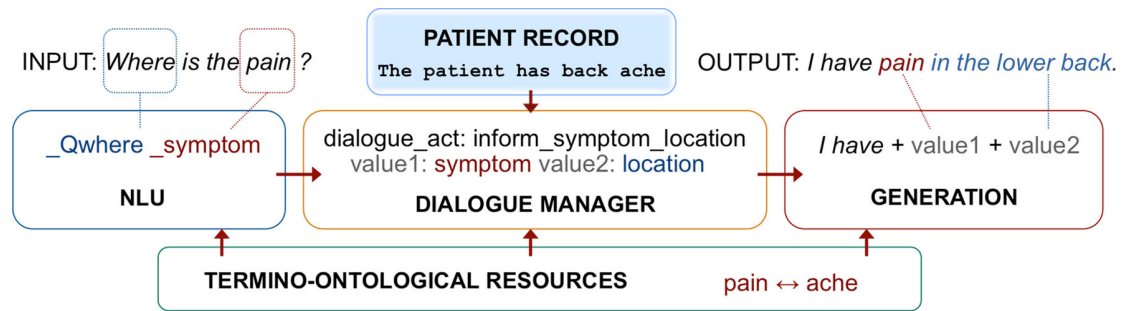


Fig. 3 Example of functioning of the dialogue system from input to output. The patient record is simplified; Table 9 shows a full example

157 6 unseen cases using the wordings of the clinical cases of
 158 French national classifying exams for medical students.³
 159 Tables 10 and 11 (Appendix) provide a brief description of
 160 each case.

161 We first conducted a user evaluation by means of 5-
 162 point Likert-scale questionnaires ranging from 1 (Very
 163 poor) to 5 (Very good). After each interaction, evaluators
 164 assessed the system on nine aspects (Table 3), which come
 165 from the evaluation framework of dialogue systems [49,
 166 50]. Evaluators were given instructions on the types of
 167 utterances the system can process, and an online link to the
 168 questionnaire.

169 We also evaluated the dialogue system's correctness. We
 170 gathered data from the dialogues with all the 35 VP cases.
 171 We analysed dialogue logs and quantified the number of
 172 correct replies. We considered correct those replies giving a
 173 coherent answer (consistent according to the user input and
 174 correct regarding the data in the record). Table 6 (Appendix)
 175 describes some examples of correct, incorrect and deferred
 176 replies. An author of this paper (LC) annotated all data;
 177 another author (SR) checked the annotations of a subset of
 178 84 (2%) turn-reply pairs that were unclear about how to
 179 classify; finally, a consensus was reached. We computed the
 180 kappa agreement between both annotators.

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

190 To measure the diversity of the unseen cases, we counted
 191 the word types (i.e. different word forms) appearing in
 192 only one record, and the types shared across different
 193 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' 194
 profiles. 195

Results 196

Quality of natural language dialogue 197

Each case was tested by an average of 3.74 evaluators (±2.8; 198
 minimum number of evaluators per case = 1; maximum = 199
 13). Panels A and B of Fig. 4 display the average evaluator 200
 scores for the seen and unseen cases respectively. Lower 201
 scores are placed to the left of each Y axis; neutral scores, in 202
 the middle; and higher scores, to the right. The bars show the 203
 cumulated percentages of evaluator scores that were Very 204
 good, Good, Neutral, Poor and Very poor. For example, 205
 in the seen cases, performance was assessed as Very good 206
 by 6% of the evaluators, as Good by an additional 52% 207
 of evaluators, as Neutral by 28% of them, and as Poor by 208
 the remaining 14%. The overall average score, obtained by 209
 averaging the mean scores given to the 9 evaluated aspects, 210
 was of 3.84 out of 5 for seen cases, and of 4.05 for unseen 211
 cases. This is above the Likert-scale midpoint. The total 212
 number of dialogues with Poor or Very poor scores ranges 213
 from 16% (naturalness) to 0% (user-understanding) for seen 214
 cases, and from 6% (naturalness) to 0% (speed) for unseen 215
 cases. 216

Regarding the system correctness, we analysed 8,078 217
 turn-reply pairs from 131 dialogues (Tables 4 and 5). We 218
 removed 149 turn-reply pairs with out-of-task questions or 219
 statements. The two researchers who double-checked the 220
 subset of turn-reply pairs had a kappa agreement of 0.827. 221
 In the full set of dialogue logs (seen and unseen cases), 222
 when analysed per medical specialty, an average of 74.3% 223
 (±9.5) system replies were correct (min = 53.6%, max = 224
 93.8%), i.e. answers were coherent with regard to inputs and 225
 provided accurate information from the record. An average 226
 of 14.9% (±6.3) of system replies were incorrect; however, 227
 unseen words only caused 2 errors. Incorrect replies affected 228
 the system's faithfulness (26.5%), the dialogue flow (56.2%) 229

Table 1 Evaluation rounds and medical specialities

	Development 2016 through May 2017	Test				
		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/Gynaecology (3) [4]	UC: Urgent Care (1) [1]			
	GH: Gastroenterology/Hepatology (5) [13]	ON: Oncology (1) [5]	U: Urology (1) [6]			

230 and the *exhaustiveness of the information* provided by the 235
 231 virtual patient (17.3%) (Table 8, Appendix). The system 236
 232 determined that the rest of the questions were beyond the 237
 233 dialogue task and answered *I do not understand* (an average 238
 234 of 7.8% ±5.3) or asked for more precision (an average of 239
 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

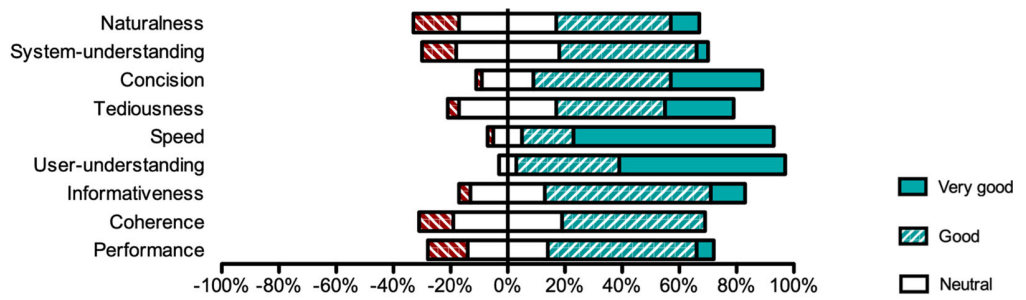
Profile	Evaluators		Description
	S	U	
Students	♂	0	Students were in their 3rd year of medical studies and had limited experience with real patients (1-3 terms of part-time hospital internship).
	♀	3	
	Unique:	7 (3, 7)	
Residents	♂	2	Residents had at least 6 years of medical studies and passed the National Classifying Exam; they had broader experience than students (one or more full-time terms as practising physicians).
	♀	4	
	Unique:	10 (6, 7)	
Practitioners /Instructors	♂	8	Practitioners were private doctors or practising doctors in hospital or general practise.
	♀	0	
	Unique:	11 (8, 5)	
Researchers /Other	♂	5	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	
	NA	4	
	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

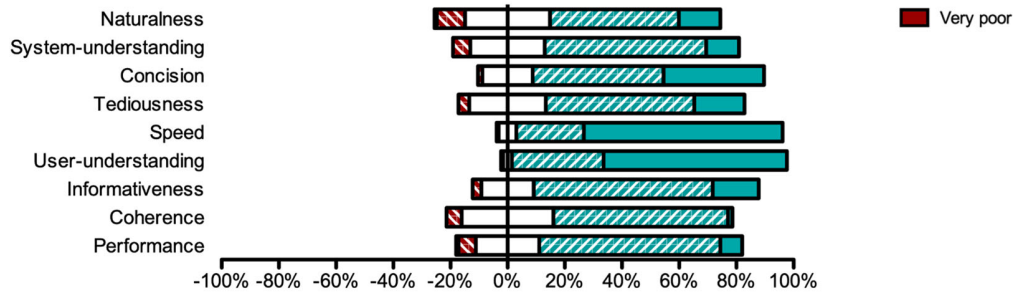
Table 3 Description of aspects addressed in the qualitative evaluation; scores ranged from 5 (*Very good*) to 1 (*Very poor*)

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.

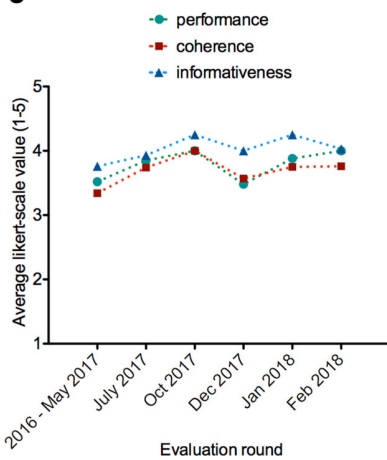
a Seen cases



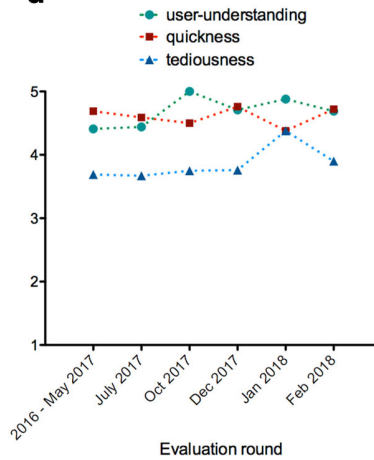
b Unseen cases



c



d



e

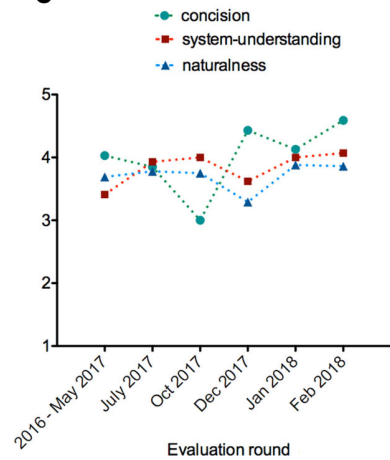


Fig. 4 Results of the qualitative evaluation and comparison between *seen* cases (used in development) and *unseen* cases

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)

240 Performance with unseen cases across specialities

241 Panels A and B of Fig. 4 display, respectively, the proportion
 242 of scores given to each aspect for the 6 seen and 29
 243 unseen cases. Evaluators rated every aspect better in the
 244 unseen cases. The differences in evaluation scores were
 245 statistically significant for the following aspects: system
 246 performance (a mean of 3.50 (95% CI[3.27-3.73]) for seen
 247 cases versus 3.81 (95% CI[3.64-3.97]) for unseen cases, p-
 248 value = 0.029, Mann-Whitney test), coherence in replies
 249 (a mean of 3.38 (95% CI[3.18-3.58]) for seen cases versus
 250 3.73 (95% CI[3.61-3.86]) for unseen cases, p = 0.004,
 251 Mann-Whitney test), informativeness (a mean of 3.78 (95%
 252 CI[3.58-3.98]) for seen cases versus 4.03 (95% CI[3.86-
 253 4.20]) for unseen cases, p = 0.047, Mann-Whitney test) and
 254 system-understanding (a mean of 3.44 (95% CI[3.22-3.66])
 255 for seen cases versus 3.90 (95% CI[3.72-4.07]), p = 0.001,
 256 t-test).

257 We also examined the variation of scores along
 258 evaluation rounds; panels C-E in Fig. 4 show the average
 259 scores for each aspect. When we compared the scores given
 260 in the first evaluation round (using seen cases) with those
 261 in the last round (using unseen cases), the following aspects
 262 showed statistically significant differences: performance (a
 263 mean of 3.48 (95% CI[3.21-3.74]) in the first round versus
 264 4.00 (95% CI[3.86-4.14]) in the last round, p = 0.003,
 265 Mann-Whitney test), coherence (a mean of 3.31 (95%
 266 CI[3.09-3.53]) in the first round versus 3.76 (95% CI[3.56-
 267 3.95]) in the last round, p = 0.005, t-test), informativeness
 268 (a mean of 3.69 (95% CI[3.48-3.90]) in the first round
 269 versus 4.03 (95% CI[3.87-4.19]) in the last round, p =
 270 0.018, Mann-Whitney test), concision (a mean of 4.00 (95%
 271 CI[3.76-4.24]) in the first round versus 4.59 (95% CI[4.40-
 272 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]) 273
 in the first round versus 4.07 (95% CI[3.89-4.24]) in the last 274
 round, p<0.0001, t-test). 275

Figure 5 plots the evaluation scores of the unseen cases 276
 grouped by speciality. From a qualitative point of view, we 277
 could not find any speciality that would consistently obtain 278
 scores below the others; outlier values correspond to cases 279
 where few dialogues were conducted. 280

Concerning the diversity of the vocabulary, unseen cases 281
 contained 1,488 types (unique word forms). 1,017 types 282
 (68.4%) appeared in isolated records; that means that only 283
 one third of the types (31.6%) occurred in more than one 284
 case. The average proportion of unique types per record is 285
 34.6% (± 7.4). Those numbers show to which extent the 286
 lexical content of each case differs across records in the 287
 unseen cases. 288

We also analysed the quantity of out-of-vocabulary 289
 words (OOVs) in unseen cases. Out of the total 1,488 types 290
 in the unseen cases, only 33 words (2.5%) were missing in 291
 system resources (avg = 1.2 OOVs per case, ± 1.66). That 292
 is, our resources covered 97.5% of the vocabulary in the 293
 29 new cases. Our analysis showed that most OOVs were 294
 spelling mistakes made when inputting data to create a new 295
 record. Our methods predicted the PoS category of these 296
 OOVs with a precision of 69.8%, a recall of 76.9%, and an 297
 F-measure of 73.2% (micro-average). Regarding the OOV 298
 words for which the system predicted the correct category, 299
 our methods to predict morphology data showed a precision 300
 of 59.4%, a recall of 61.3%, and an F-measure of 60.3% 301
 (micro-average). Table 7 (Appendix) shows further details 302
 about our results per category. 303

Lastly, Fig. 5 (bottom right) depicts differences in 304
 assessment according to the evaluators' profiles. The 305
 average scores of the majority or totality of evaluators 306
 agreed on user-understanding, quickness, tediousness and 307
 concision. Students and residents gave higher average 308
 scores to system performance, coherence of replies, 309
 informativeness, system- and user-understanding. Senior 310
 doctors generally gave lower scores. 311

Discussion 312

The quality of the natural language dialogue in seen and 313
 unseen cases received very positive, positive, or neutral 314

Table 5 Evaluation of system correctness expressed as average percentage (\pm 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]

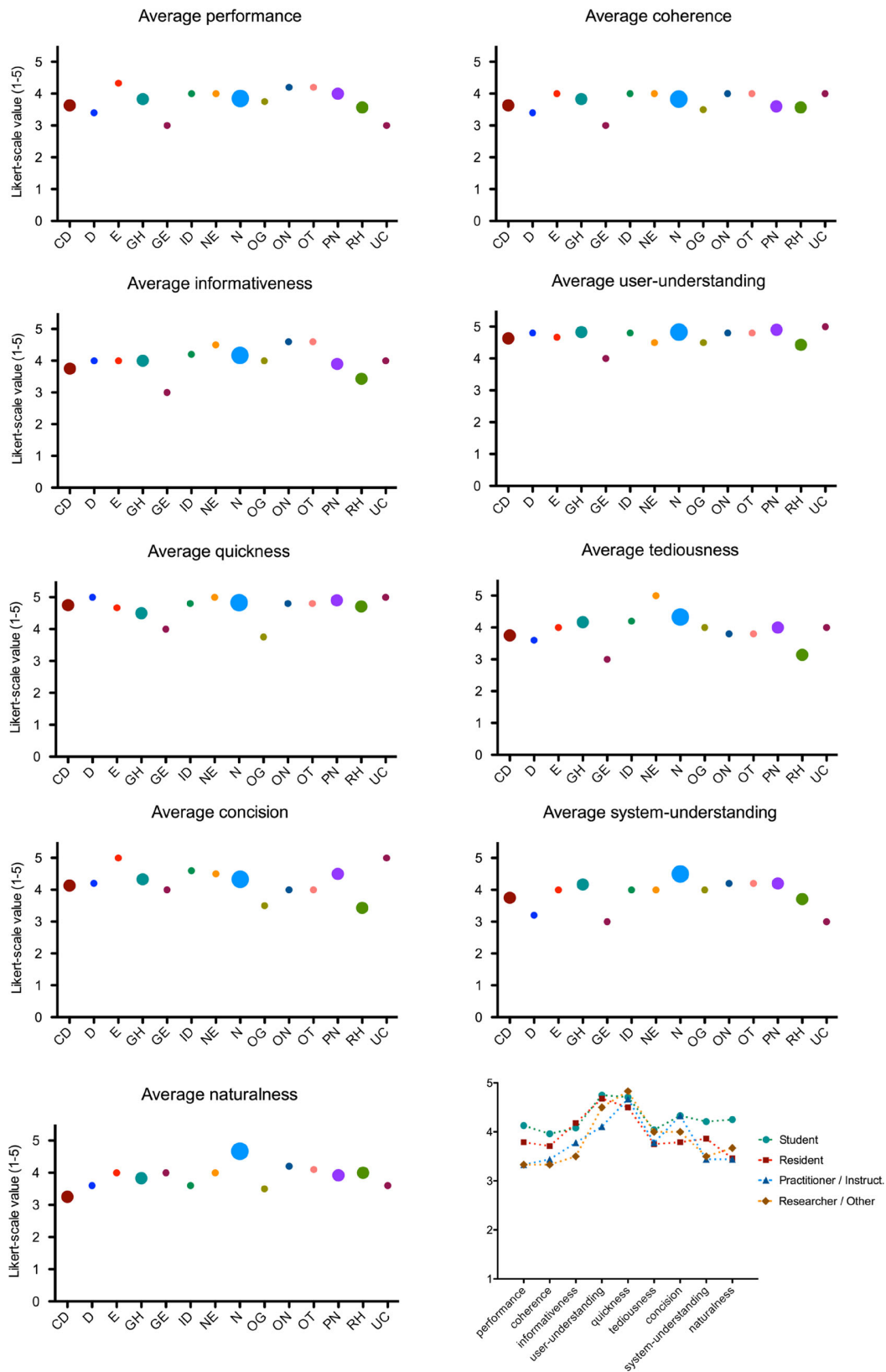


Fig. 5 Qualitative evaluation across medical specialties and evaluator profiles. The size of each point expresses the number of dialogues conducted: 1–5 (small size), 6–10 (medium size) and >10 (large size). The abbreviations of specialties are given in Table 1

315 judgements from between 93% and 100% of the evaluators,
 316 allowing us to answer Q1 positively. System performance
 317 and coherence of replies received Good and Very good
 318 scores and overall satisfaction was high with an average of
 319 3.84 (seen cases) and 4.06 (unseen cases) across all aspects.
 320 We cannot compare the error rate with other works (e.g.
 321 [34]) without bias, since we tested more patient cases.

322 Regarding Q2, in the test on unseen cases, every
 323 aspect received a higher user evaluation score than on
 324 seen cases. The improvement of some features proved
 325 statistically significant. The system was robust enough to
 326 cope with new cases without quality loss. The system's
 327 vocabulary coverage of unseen cases was very high
 328 (97.5%). Overall, we tested 35 different cases covering
 329 18 medical specialities. To the best of our knowledge,
 330 this is much larger than what was reported so far in the
 331 literature.

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

341 The correction rate of system replies varied across cases
 342 largely due to each record content: e.g. the performance
 343 was lower in a postpartum case, where some questions
 344 referred to the patient's newborn, but the system could
 345 not distinguish them from those related to the VP.
 346 Our analysis of logs across cases unveiled that most
 347 errors were due to the lack of variants of question
 348 formulations, missing question types, or processing errors
 349 (Table 6, Appendix). These weaknesses require fallback
 350 strategies, which we explored using machine learning
 351 [51].

352 At a technical level, we want to improve the performance
 353 of the dialogue manager and the comparison and update
 354 procedures. Given the lack of dialogue corpora for the
 355 task, we did not apply machine/deep learning approaches.
 356 Terminological components can mitigate the needs of the
 357 domain—rich in variant terms and acronyms, but without
 358 open training data available. This is the asset of our
 359 system. Once enough dialogue logs are collected via
 360 a rule- and terminology-based system, the data can be
 361 trained to complement the dialogue policy manager, or to
 362 generate word-embeddings for OOV terms. This is left for
 363 future work. The naturalness of system replies needs also
 364 refinement, especially the way it simplifies long sentences
 365 or outputs negative symptoms and layman terms. We are
 366 interested in evaluating the system in the overall framework
 367 of a simulated consultation, where medical students should

diagnose the patient. This would allow us to know whether 368
 the system helps students to obtain all key elements of the 369
 history-taking step, and to ascertain whether students make 370
 a correct diagnosis. Finally, we need to gather dialogue data 371
 to evaluate the English and Spanish versions. 372

Lessons learned 373

Regarding development, several aspects demanded a heavy 374
 investment in resource creation: terminology components 375
 for concept mapping, update procedures to compare the 376
 existing knowledge base and OOVs, and linguistically- 377
 motivated modules to transform the data created by 378
 medical trainers according to the patient's perspective. 379
 Moreover, misspellings in trainers' input needed spelling 380
 correction tools. To fix the OOV errors related to spelling 381
 mistakes, the most reliable approach would be to include 382
 a correction module on the back-office interface that 383
 trainer doctors use to create the patient record. The 384
 system vocabulary could be mapped to misspellings, flag 385
 them, and the trainers could correct them before the 386
 interaction. Nevertheless, the developed modules were 387
 capable of adapting the system to new cases without 388
 causing problematic interactions, according to the end-user 389
 evaluation. 390

Regarding the system design and evaluation, we strongly 391
 advise that medical professionals be involved from the 392
 beginning. The closer to reality the patient data we received, 393
 the better the system was tested and improved. The more 394
 iterations were conducted for inspecting logs and fixing 395
 errors, the better the system was rated. Our evaluation 396
 revealed that experienced practitioners assessed the system 397
 as less satisfactory, given their greater diagnosis experi- 398
 ence and different perception of these tools. This high- 399
 lights the careful choice of the end-user and its impact 400
 on the framework design. This multi-case, adaptable VP 401
 system seems to fit medical students and interns, since 402
 they can bear infelicities in system replies and need to 403
 engage in the interaction to gain experience. A tool with 404
 canned answers would be rigid and necessitate more engi- 405
 neering to adapt to new cases. If no dialogue data are 406
 available for the task, collecting dialogue logs with poten- 407
 tial end-users seems feasible before data-intensive methods 408
 (machine or deep learning) can be applied. Finally, this sys- 409
 tem is not yet suited for simulating VPs with chronic condi- 410
 tions needing follow-up consultations. Evolving symptoms 411
 would require a more advanced model of the VP's disease 412
 timeline. 413

Overall, the tradeoff between adaptability and nat- 414
 uralness has design implications related to immediate 415
 vs long-term needs, or sophisticated case-specific vs 416
 generic applications. Table 12 (Appendix) outlines our 417
 observations. 418

419 Conclusion

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

The agent was tested with 35 different cases from 431
 18 different specialities. Medical evaluators considered 432
 that this system provides quality dialogue through natural 433
 language, that it does so across heterogeneous cases and 434
 medical specialities, and that it processes new records 435
 without quality loss compared to already known cases. 436
 Our usability evaluation showed that this multi-case system 437
 can support student training in history taking and provided 438
 us with lessons we thought useful to share regarding its 439
 strengths and limits. 440

Appendix

441

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

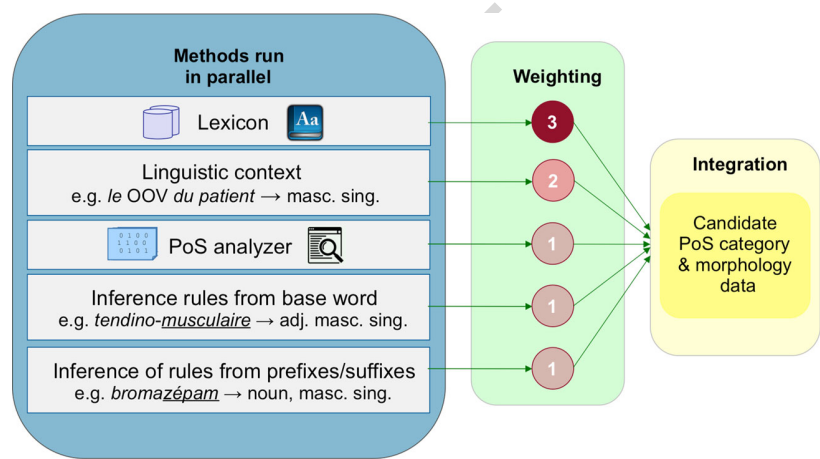


Table 6 Examples of correct, incorrect and deferred replies (I: 'input'; R: 'system reply'); we show the English translation of dialogue interactions using the French system

Correct	The information is accurate with regard to the patient record (<i>correctness</i>)	I: Since when have you had these symptoms? R: I have been nauseous for 2 weeks
	No information is invented (<i>faithfulness</i>)	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 (<i>exhaustiveness</i>)	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 <u>it</u> start?

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: <u>Poid</u> ? (poids) '(your) weight?' R: Pardon ? 'Excuse me?'
		Term missing in vocabulary list	I: Do you <u>limp</u> ?
		Variation of query phrasing	R: I do not understand what you are asking for I: Avez-vous un <u>concubin</u> ? '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 <u>tete</u> ? (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 <u>fast food</u> ?	
Variation of query phrasing		R: I do not follow any special diet 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?')	

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’ <i>faithfulness</i> . It would be preferable to request for repair or reply <i>I have not understood</i>
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 <i>dialogue flow</i> . 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’]

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

```

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')

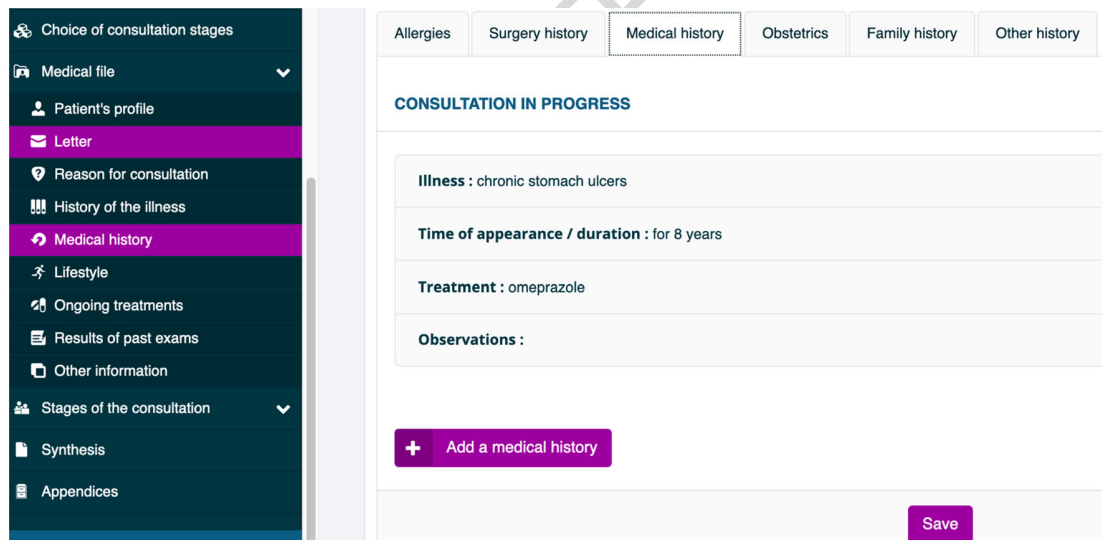


Fig. 7 Procedures and weighting scheme to predict linguistic information for OOV items

Table 10 Description of the seen cases used in the usability 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	
A 49-year-old man consults about a violent thoracic pain since last night.	Essential hypertension 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

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	<ul style="list-style-type: none"> • 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: <ul style="list-style-type: none"> – 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 (<i>correctness</i>), without inventing information (<i>faithfulness</i>) nor omitting data (<i>exhaustiveness</i>) – And all the above, in a <i>dynamic dialogue flow</i>: 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	<ul style="list-style-type: none"> • Invest in creating termino-ontologic resources: <ul style="list-style-type: none"> – 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	<ul style="list-style-type: none"> • 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)

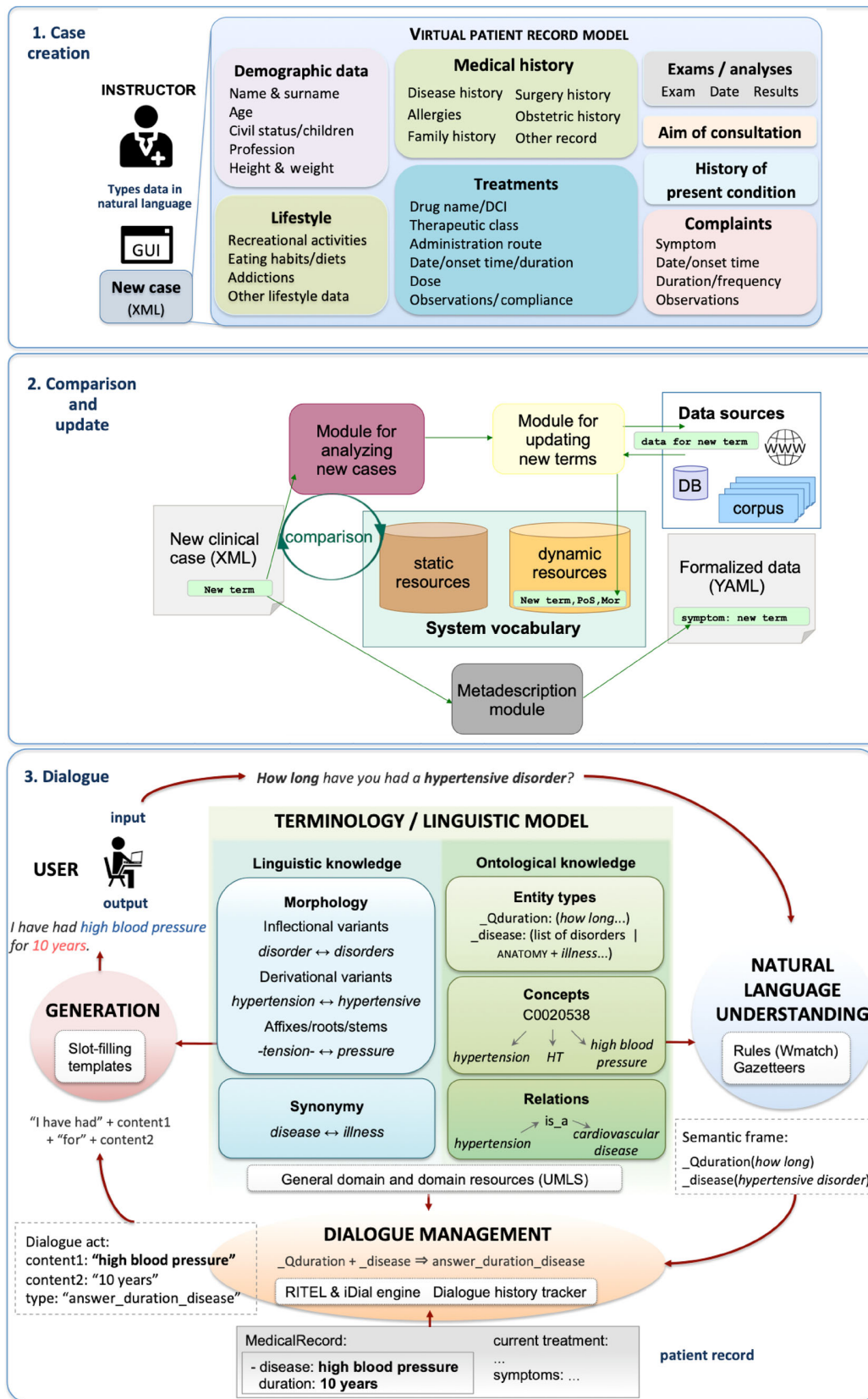


Fig. 8 Overall functioning of the dialogue system and update components; further technical details are provided in [27, 46, 47]

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

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

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

464 **Data availability** The dialogue data collected during development and
465 evaluation is available at: <https://pvdial.limsi.fr/data/PG-logs-eval.zip>
466 A demonstration of the dialogue system can be tested at: <http://vps-9069f76a.vps.ovh.net>
467

468 **Code availability** Not applicable.

469 Declarations

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

472 References

- 473 1. Washburn, M., Bordnick, P., and Rizzo, A. S., A pilot feasibility
474 study of virtual patient simulation to enhance social work
475 students' brief mental health assessment skills. *Soc. Work Health*
476 *Care* 55(9):675–693, 2016.
- 477 2. Barnett, S. G., Gallimore, C. E., Pitterle, M., and Morrill, J.,
478 Impact of a paper vs virtual simulated patient case on student-
479 perceived confidence and engagement. *Am. J. Pharm. Educ.*
480 80(1):16, 2016.
- 481 3. McCoy, L., Pettit, R. K., Lewis, J. H., Allgood, J. A., Bay, C., and
482 Schwartz, F. N., Evaluating medical student engagement during
483 virtual patient simulations: A sequential, mixed methods study.
484 *BMC Med. Educ.* 16:20, 2016.
- 485 4. Tait, L., Lee, K., Rasiah, R., Cooper, J. M., Ling, T., Geelan,
486 B., and Bindoff, I., Simulation and feedback in health education:
487 A mixed methods study comparing three simulation modalities.
488 *Pharmacy (Basel)* 6(2), 2018.
- 489 5. Courteille, O., Fahlstedt, M., Ho, J., Hedman, L., Fors, U., von
490 Holst, H., Fellander-Tsai, L., and Moller, H., Learning through a

- virtual patient vs. recorded lecture: A comparison of knowledge
491 retention in a trauma case. *Int. J. Med. Educ.* 9:86–92, 2018. 492
6. Gupta, A., Singh, S., Khaliq, F., Dhaliwal, U., and Madhu, S. V.,
493 Development and validation of simulated virtual patients to impart
494 early clinical exposure in endocrine physiology. *Adv. Physiol.*
495 *Educ.* 42(1):15–20, 2018. 496
7. de Cock, C., Milne-Ives, M., van Velthoven, M. H., Alturkistani,
497 A., Lam, C., and Meinert, E., Effectiveness of conversational
498 agents (virtual assistants) in health care: Protocol for a systematic
499 review. *JMIR Res. Protoc.* 9(3):e16934, 2020. 500
8. Ellaway, R., Candler, C., Greene, P., and Smothers,
501 V., An architectural model for MedBiquitous virtual
502 patients. [http://groups.medbiq.org/medbiq/display/VPWG/
503 MedBiquitous+Virtual+Patient+Architecture](http://groups.medbiq.org/medbiq/display/VPWG/MedBiquitous+Virtual+Patient+Architecture). Accessed: 8 Dec
504 2018, 2006. 505
9. Sijstermans, R., Jaspers, M. W., Bloemendaal, P., and Schoonder-
506 waldt, E., Training inter-physician communication using the
507 dynamic patient simulator®. *Int. J. Med. Inf.* 76(5–6):336–343,
508 2007. 509
10. Danforth, D. R., Procter, M., Chen, R., Johnson, M., and
510 Heller, R., Development of virtual patient simulations for medical
511 education. *J. Virtual Worlds Res.* 2(2):4–11, 2009. 512
11. Rombauts, N., Patients virtuels: pédagogie, état de l'art et
513 développement du simulateur Alphadiag. PhD dissertation,
514 Faculty of Medicine, Claude Bernard University, Lyon France,
515 2014. 516
12. Menendez, E., Balisa-Rocha, B., Jabbur-Lopes, M., Costa, W.,
517 Nascimento, J. R., Dósea, M., Silva, L., and Junior, D. L., Using a
518 virtual patient system for the teaching of pharmaceutical care. *Int.*
519 *J. Med. Inf.* 84(9):640–646, 2015. 520
13. Lin, C. J., Pao, C. W., Chen, Y. H., Liu, C. T., and Hsu, H.
521 H., Ellipsis and coreference resolution in a computerized virtual
522 patient dialogue system. *J. Med. Syst.* 40(9):206–221, 2016. 523
14. Laleye, F. A., Blanié, A., Brouquet, A., Behnamou, D., and
524 de Chalendar, G., Semantic similarity to improve question
525 understanding in a virtual patient. In: *Proceedings of the 35th*
526 *Annual ACM Symposium on Applied Computing*, pp. 859–866,
527 2020. 528
15. Chen, F., Lee, Y., and Hubal, R., Work-in-progress—testing of a
529 virtual patient: Linguistic and display engagement findings. In:
530 *2020 6th International Conference of the Immersive Learning*
531 *Research Network (iLRN)*, pp. 348–350: IEEE, 2020. 532
16. Candler, C., Effective use of educational technology in medical
533 education. In: *Colloquium on Educational Technology: Recom-*
534 *mendations and Guidelines for Medical Educators*, pp. 1–19.
535 Washington, DC: AAMC Institute for Improving Medical Educa-
536 tion, 2007. 537
17. Schmidlen, T., Schwartz, M., DiLoreto, K., Kirchner, H. L., and
538 Sturm, A. C., Patient assessment of chatbots for the scalable
539 delivery of genetic counseling. *J. Genet. Couns.* 28(6):1166–1177,
540 2019. 541
18. Chetlen, A., Artrip, R., Drury, B., Arbaiza, A., and Moore, M.,
542 Novel use of chatbot technology to educate patients before breast
543 biopsy. *J. Am. Coll. Radiol.* 16(9 Pt B):1305–1308, 2019. 544
19. Kokciyan, N., Chapman, M., Balatsoukas, P., Sassoan, I., Essers,
545 K., Ashworth, M., Curcin, V., Modgil, S., Parsons, S., and Sklar,
546 E. I., A collaborative decision support tool for managing chronic
547 conditions. *Stud. Health Technol. Inform.* 264:644–648, 2019. 548
20. Cook, D. A., Erwin, P. J., and Triola, M. M., Computerized
549 virtual patients in health professions education: A systematic
550 review and meta-analysis. *Acad. Med.* 85(10):1589–1602, 2010.
551 <https://doi.org/10.1097/ACM.0b013e3181edfe13>. 552
21. Wattanasoontorn, V., Hernández, R. J. G., and Sbert, M.,
553 Embodied conversational virtual patients. In: Diana, P. M., and
554 Nieto, I. P. (Eds.) *Conversational Agents and Natural Language* 555

⁴<https://pvdial.limsi.fr>

- 556 *Interaction: Techniques and Effective Practices*, pp. 254–281.
 557 Hershey: Information Science Reference, IGI Global, 2011.
 558 <https://doi.org/10.4018/978-1-60960-617-6.ch011>.
- 559 22. Rossen, B., and Lok, B., A crowdsourcing method to develop
 560 virtual human conversational agents. *Int. J. Hum. Comput. Stud.*
 561 70(4):301–319, 2012.
- 562 23. Lelardeux, C., Panzoli, D., Alvarez, J., Galaup, M., and
 563 Lagarrigue, P., Serious game, simulateur, serious play: État de
 564 l'art pour la formation en santé. In: *Actes du colloque Serious*
 565 *Games en Médecine et Santé (SeGaMED) 2013*, pp. 27–38. Nice:
 566 e-virtuoses, 2013.
- 567 24. Wattanasontorn, V., Hernández, R. J. G., and Sbert,
 568 M., Serious games for e-health care. In: Cai, Y., and
 569 Goei, S. (Eds.) *Simulations, Serious Games and Their*
 570 *Applications*, pp. 127–146. Singapore: Springer, 2014.
 571 https://doi.org/10.1007/978-981-4560-32-0_9.
- 572 25. Reisch, A., and Haag, M., Evaluation of chatbot prototypes for
 573 taking the virtual patient's history. *Stud. Health Technol. Inform.*
 574 260:73–80, 2019.
- 575 26. Nirenburg, S., Beale, S., McShane, M., Jarrell, B., and Fantry,
 576 G., Language understanding in Maryland virtual patient. In:
 577 *Proceedings of the International Conference on Computational*
 578 *Linguistics*, pp. 36–39. Manchester: Citeseer, 2008.
- 579 27. Campillos-Llanos, L., Bouamor, D., Bilinski, É., Ligozat, A.
 580 L., Zweigenbaum, P., and Rosset, S., Description of the
 581 PatientGenesys dialogue system. In: *Proceedings of SIGDIAL*,
 582 pp. 438–440. Prague: Association for Computational Linguistics,
 583 2015.
- 584 28. Leuski, A., and Traum, D., Practical language processing for
 585 virtual humans. In: *Proceedings on Innovative Applications of*
 586 *Artificial Intelligence Conference*, pp. 1740–1747. Atlanta, 2010.
- 587 29. Rizk, Y., Kshoury, K., Chehab, M., Chidiac, P., Awad, M., and
 588 Antoun, J., Virtual patient. In: *Proceedings of WINLP*, pp. 1–3.
 589 Vancouver, 2017.
- 590 30. Datta, D., Brashers, V., Owen, J., White, C., and Barnes, L. E., A
 591 deep learning methodology for semantic utterance classification
 592 in virtual human dialogue systems. In: Traum, D., Swartout, W.,
 593 Khooshabeh, P., Kopp, S., Scherer, S., and Leuski, A. (Eds.)
 594 *Intelligent Virtual Agents, Los Angeles*, pp. 451–455. Berlin:
 595 Springer, 2016.
- 596 31. Jin, L., White, M., Jaffe, E., Zimmerman, L., and Danforth, D.,
 597 Combining cnns and pattern matching for question interpretation
 598 in a virtual patient dialogue system. In: *Proceedings on Workshop*
 599 *Innovative Use NLP Building Educational Applications*, pp. 11–
 600 21: Copenhagen, 2017.
- 601 32. Dickerson, R., Johnsen, K., Raij, A., Lok, B., Hernandez, J.,
 602 Stevens, A., and Lind, D. S., Evaluating a script-based approach
 603 for simulating patient-doctor interaction. In: *Proceedings of*
 604 *the International Conference on Human-Computer Interface*
 605 *Advances Modeling and Simulation*, pp. 79–84. New Orleans,
 606 2005.
- 607 33. Pence, T. B., Dukes, L. C., Hodges, L. F., Meehan, N.
 608 K., and Johnson, A., The effects of interaction and visual
 609 fidelity on learning outcomes for a virtual pediatric patient
 610 system. In: *IEEE International Conference on Healthcare*
 611 *Informatics (ICHI)*, pp. 209–218. Philadelphia: IEEE, 2013.
 612 <https://doi.org/10.1109/ICHI.2013.36>.
- 613 34. Maicher, K., Danforth, D., Price, A., Zimmerman, L.,
 614 Wilcox, B., Liston, B., Cronau, H., Belknap, L., Led-
 615 ford, C., Way, D. et al., Developing a conversational
 616 virtual standardized patient to enable students to practice
 617 history-taking skills. *Simul. Healthc.* 12(2):124–131, 2017.
 618 <https://doi.org/10.1097/SIH.000000000000195>.
- 619 35. Talbot, T. B., Sagae, K., John, B., and Rizzo, A. A., Sorting out
 620 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., Lehouste, 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*

686 *International Conference on Language Resources and Evaluation,*
687 pp. 2728–2735. Portorož, 2016. 691
688 51. Campillos-Llanos, L., Rosset, S., and Zweigenbaum, P., 692
689 Automatic classification of doctor-patient questions for a
690 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