BMC Medical Education

RESEARCH Open Access

Applying ChatGPT to plan and create a realistic collection of virtual patients for clinical reasoning training

Joanna Fąferek[1], Andrzej A. Kononowicz[2], Nataliia Bogutska[3], Vital Da Silva Domingues[4,5], Nataliia Davydova[3], Ada Frankowska[2], Isabel Iguacel[6,7,8,9], Anja Mayer[10], Luc Morin[11], Nataliia Pavlyukovich[3], Iryna Popova[3], Tetiana Shchudrova[3], Małgorzata Sudacka[1], Renata Szydlak[2] and Inga Hege[12*]

Abstract

Background Virtual patients (VPs) are useful tools in training of medical students' clinical reasoning abilities. However, creating high-quality and peer-reviewed VPs is time-consuming and resource-intensive. Therefore, the aim of this study was to investigate whether generative artificial intelligence (AI) could facilitate the planning and creation of a diverse collection of VPs suitable for training medical students in clinical reasoning.

Methods We used ChatGPT to generate a blueprint for 200 diverse VPs that adequately represent the population in Europe. We selected five VPs from the blueprint to be created by humans and ChatGPT. We assessed the generated blueprint for representativeness and internal consistency, and we reviewed the VPs in a multi-step, partly blinded process for didactical quality and content accuracy. Finally, we received 44 VP evaluations from medical students.

Results The generated blueprint did not meet our expectations in terms of quality or representativeness and showed repetitive patterns and an unusually high number of atypical VP outlines.
The ChatGPT- and human-generated VPs were comparable in terms of didactic quality and medical accuracy. Neither contained any medically incorrect information and reviewers and students could not discern significant differences. However, the five human-created VPs demonstrated a greater variety in storytelling, differential diagnosis, and patient-doctor interaction. The ChatGPT-generated VPs also included AI-generated patient images; however, we could not generate realistic clinical images.

Conclusions While we do not consider ChatGPT in its current version capable of generating a realistic blueprint for a VP collection, we believe that the process of prompting, combined with iterative discussions and refinements after each step, is promising and warrants further exploration. Similarly, although ChatGPT-generated VPs can serve as a good starting point, the variety of VP scenarios in a large collection may be limited without interactions between authors and reviewers to further refine it.

Keywords Virtual patients, Generative artificial intelligence, ChatGPT, Large language models, Clinical reasoning, Medical education

*Correspondence:
Inga Hege
email@ingahege.de

Full list of author information is available at the end of the article

Background

Virtual patients (VPs) are interactive computer simulations that provide a safe environment for medical students to train their clinical reasoning (CR) abilities, such as gathering and interpreting information, generating differential diagnoses, and developing treatment plans [1–3]. The CR process is influenced by contextual factors related to the healthcare professional or student, the patient, and the setting of the encounter [4, 5]. Consequently, when designing VPs a deliberate selection of contextual factors, such as key symptoms, diagnoses, or age and gender of VP is crucial to allow learners to train CR by comparing and contrasting common conditions and experiencing many VPs with varying levels of complexity [6].

Despite the high costs required for the development [7], healthcare educators and collaborative initiatives have been engaged in developing VP collections for more than 30 years to facilitate the training of CR for students. Examples include the European "Electronic virtual patient project" (eViP), which shared a pool of 300 VPs [8], the MedU initiative which created VP collections in pediatrics and internal medicine [9], or the New England Journal of Medicine project Healer [10].

Similarly, based on the experiences of such previous projects, the aim of the Erasmus + funded iCoViP (international collection of virtual patients) project was to create a collection of 200 open-access VPs representing a realistic proportion of the European patient population in terms of sociodemographics, disease incidence, and reasons for consultation (= key symptoms) [11, 12]. The creation of a blueprint outlining the VP collection was completed within three months and required several rounds of consensus within the project consortium and careful refinement based on statistical population data to ensure its diversity and representativeness [13, 14].

The international project consortium started the VP creation process with a training workshop for the clinical educators, who created the VPs in the CASUS learning system. To facilitate CR, the VPs include quiz questions to help interpret findings, modeled patient-physician dialogs, that require students to identify and prioritize information provided by the VP, an interactive concept map to visualize the CR process [15], and the formulation of a summary statement [16] (see Fig. 1). The VPs also include an image of the fictitious patient from a stock photography collection and all relevant clinical images, such as X-rays and ECGs acquired with permission from the authors' clinical setting. Such images are essential for the training of CR because they allow learners to interpret them and refine their differential diagnoses accordingly. Similarly, patient images provide clinically relevant information, such as signs of being overweight or jaundiced,

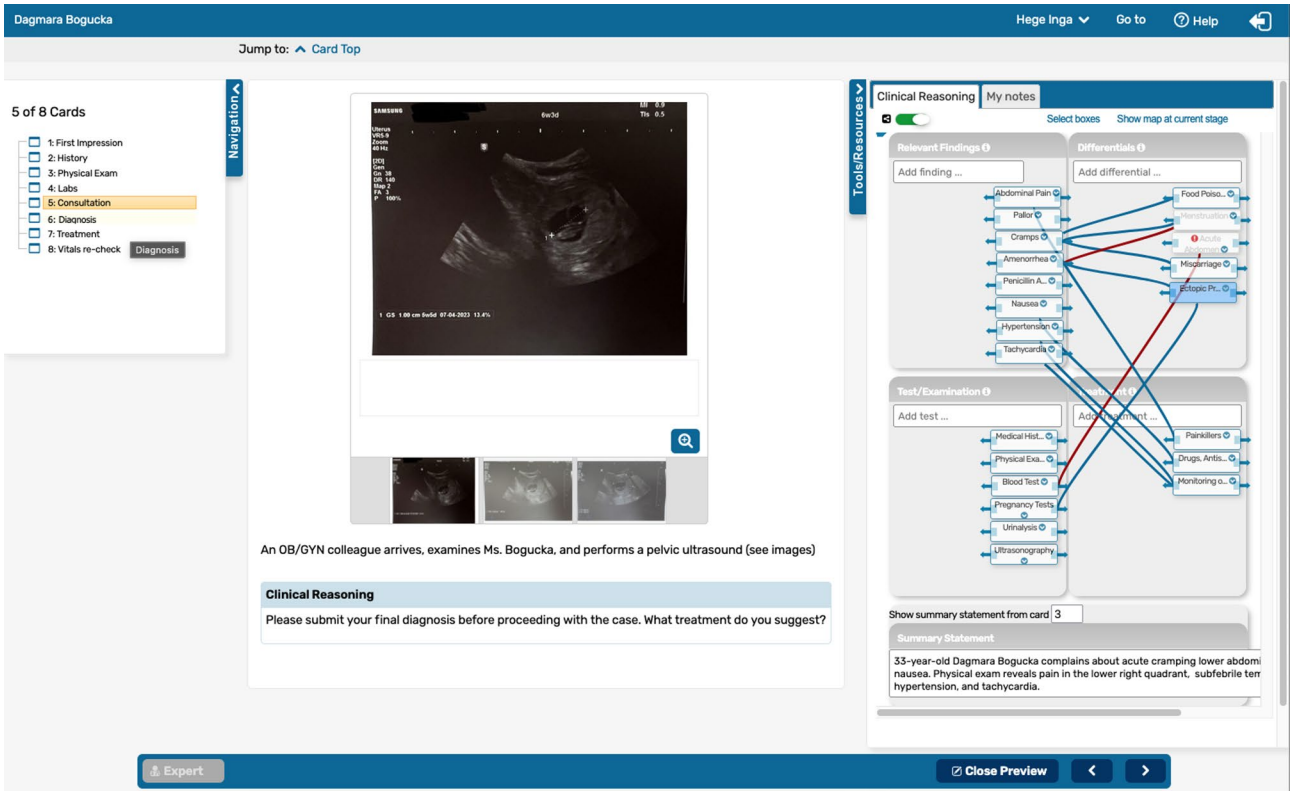

Fig. 1 Screenshot of an iCoViP VP with the CR concept map

provide contextual information, and lead to high learner satisfaction and perceived authenticity [17, 18].

Each created VP underwent a language and grammar check, followed by a thorough formal, didactic, and content review using review checklists and guidelines developed based on the aforementioned projects and frameworks [19, 20]. After the changes had been implemented, the VPs were copied and translated from English into six languages using DeepL [21]. In total, the project involved one public health expert, more than 20 clinicians, two medical education experts, and two project managers, and it took two years to complete the project.

Since the implementation of iCoViP (2021–2023), studies have shown that large language models (LLMs), such as ChatGPT, can be used to create VPs or case vignettes [22–25] or at least assist educators in the creation process [26]. Also, Moser et al. recently published 12 tips on how generative AI can create and optimize VPs [27]. However, ethical issues, such as biases, stereotyping, and false associations among patient characteristics need to be considered [23, 28–30]. For example, a study by Zack et al. found that GPT-4 exhibited deficiencies in accurately modelling demographic diversity and generated clinical vignettes with stereotypical presentations. Consequently, they emphasized the necessity for bias assessments of LLMs [31].

Using ChatGPT, Cook developed a conversational VP supporting management reasoning. While the overall results were promising, they encountered challenges, such as patients using atypical language that seemed overly detailed or polite [22].

Wong et al. described an approach in which faculty members created and reviewed nine cases for CR based on contextual information such as age, learning objectives, and discussion questions for three chief complaints. They found that the cases showed a high degree of simplification and uniformity across the chief complaints using standard formulations. Moreover, they suggested developing a framework that educators can apply for efficient and high-quality prompting [32]. Bakkum et al. created 30 case vignettes including patient images with LLMs which accelerated the vignette development and enhanced diversity. They concluded that their eight prompts are easily re-usable; however, the process itself requires computer skills not all educators may possess [33].

Thus, studies have demonstrated the potential of generative AI tools to generate one or more VPs. However, to our knowledge no studies have investigated the potential of generative AI to support the entire process of planning, creating, and reviewing a high-quality, diverse, and representative VP collection.

Therefore, the aim of this study was to investigate the potential of generative AI tools to facilitate each step of such an endeavor. The iCoViP project with the well-established process of planning, creating, and reviewing a VP collection and its publicly available guidelines and checklists served as a conceptual framework for this study.

Our research questions for this aim are:

1. Can ChatGPT develop a blueprint for a collection of VPs that represent the European (patient) population?
2. Can ChatGPT generate these VPs including the CR activities based on criteria established in the iCoViP project?
3. Can AI tools generate realistic patient and clinical images, that are aligned with the VP scenarios?
4. Which steps in the process of planning and developing a VP collection can be supported using generative AI and still be comparable in quality to VPs created by humans?

Methods

We deliberately chose ChatGPT (OpenAI, San Francisco, CA, United States) for answering the first two research questions as it is the most widely used generative AI tool in medical education [34, 35], which increases the likelihood that our results will be applicable to the medical education community.

Blueprint generation

Based on the metadata scheme of the iCoViP blueprint [13], we developed a series of prompts to generate a blueprint representing the European patient population. The blueprint includes sociodemographic data of a VP (age, name, gender, sexual orientation, migration background, disability), clinical data (diagnosis, key symptom, and acuity), and contextual data such as the encounter setting.

For developing the prompts, we adhered to best practice guides on prompt engineering and recommended prompt patterns [36, 37] and structured the task into several steps with two manual preparatory steps. First, for the creation of the initial list of diagnoses, we had to support ChatGPT in obtaining the incidences for diseases by providing the MeSH categories. Secondly, to align the diagnoses with key symptoms, we provided ChatGPT with a list of 40 key symptoms created by the iCoViP consortium [11]. We refined the prompts in an iterative process based on the quality criteria provided in Appendix 1. Apart from these two manual steps, no further manual interventions or corrections of the outputs were made.

Blueprint evaluation

Following the approach developed during the iCoViP project [14], we assessed the final blueprint concerning the representativeness of diagnoses, key symptoms, and

sociodemographic data of the population, in Europe and internal logic. To assess the representativeness of the final diagnoses in the ChatGPT blueprint we compared them to the most common reasons for encounter published by Finley et al. [38]. Concerning the key symptoms, we assessed how well ChatGPT included the 40 key symptoms it was given during the prompting. Finally, we compared the sociodemographic data (e.g. patient age, gender, or disability), with statistical sources (where available) covering the different aspects.

Two physicians (LM, VD) assessed the blueprint for inconsistencies in internal logic and documented their findings. Such inconsistencies were for example rare combinations of age or gender and diagnosis or unusual combinations of age and occupation. Two researchers (AM, JF) summarized these findings and clustered them into categories.

VP creation

From the blueprint we deliberately selected five VPs outlines representing a range of diagnoses, key symptoms, and VP sociodemographics (see Appendix 3 for details). Then, we developed a workflow to ensure the quality of the process for creating these five VPs manually and with ChatGPT (see Fig. 2). VPs were created manually in English in the CASUS VP system by five clinicians (NB, ND, NP, II, VD) during January-March 2025. They have experience in medical education and received a training session by healthcare educators (JF, IH). During the creation process we organized two online meetings to discuss the authors' progress and answer any questions. The completed VPs included all text elements, quiz questions, suitable patient and clinical images, and a CR concept map. We retrieved anonymized clinical images from hospitals or open access image databases.

The ChatGPT-supported creation of a VP included three steps: (1) the generation of the VP story and questions, (2) the creation or selection of the patient and clinical images, and (3) the development of the CR concept map. First, we developed a prompt for the generation of a VP story that received the VP specification from the blueprint as input. We tested the prompt with a randomly selected VP from the blueprint refining it until our criteria were met. Then, we prompted ChatGPT to generate each selected VP as an HTML file. The file was downloaded and imported into CASUS using an adapted pre-existing interface for file import.

For creating the patient and clinical images we selected and tested 11 freely available generative AI tools (see Appendix 4 for details) prompting them to create a

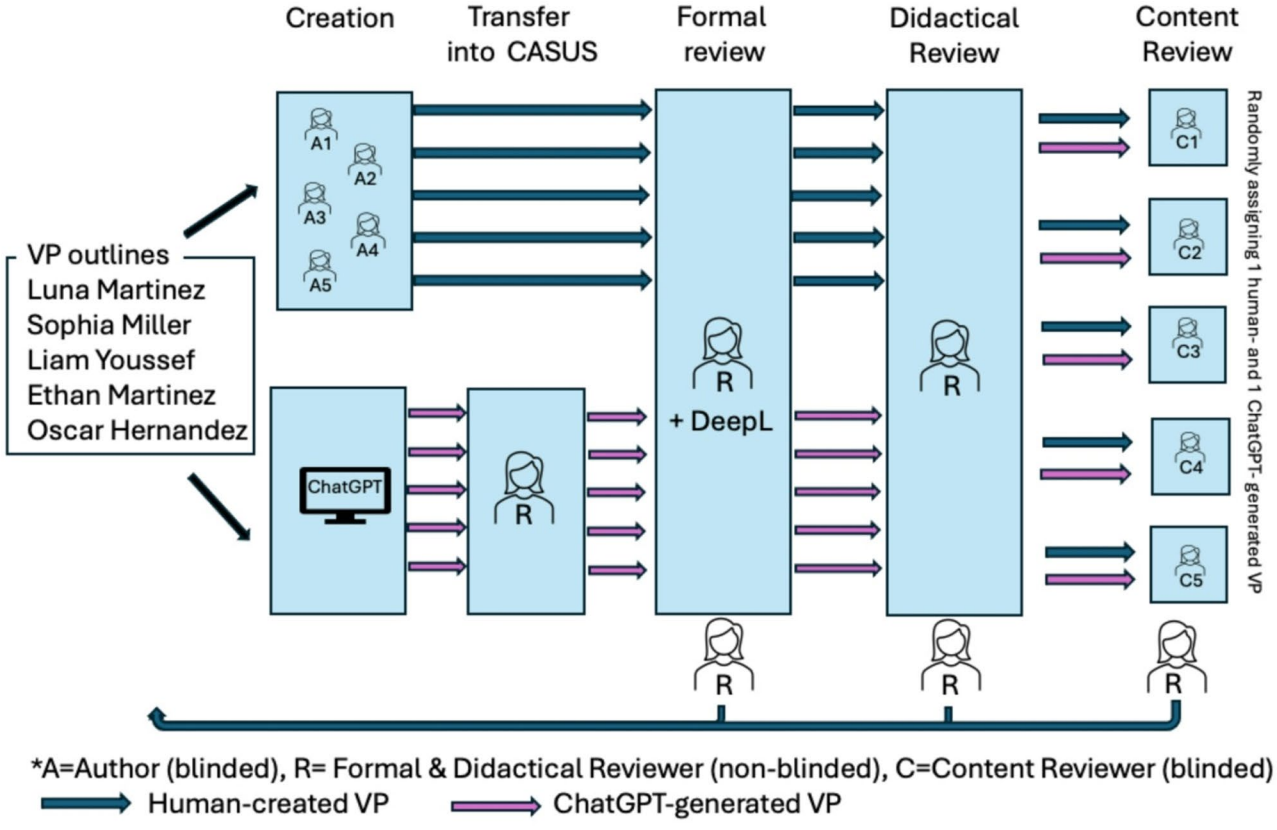

Fig. 2 Workflow for the creation and multi-step review process of the human- and ChatGPT-generated VPs

normal chest x-ray and a picture of a male patient. Based on these results, we selected Adobe Firefly for the creation of the patient images. Thus, we prompted Adobe Firefly to generate patient images for the five VPs during February 2025. Three tools (DALL-E, Craivon, and Adobe Firefly) provided promising results for the chest x-ray, so we further tested them with a range of clinical images including pathological findings. However, after evaluating the results, the clinicians on our team regarded them as insufficient for educational VPs. Therefore, we retrieved anonymized clinical images from hospitals or open access image databases.

Finally, we developed a prompt for the generation of the CR concept map that is included in each VP. We used the original prompt developed by Szydlak et al. [39] and refined it to be run based on an uploaded VP scenario and to provide an output that we could import into CASUS. After the didactical review had been completed, we exported the five VPs from CASUS, uploaded them to ChatGPT and executed the prompts. The generated maps were imported into CASUS via a programmed map importer and were checked and refined in case some map elements were not imported automatically.

All prompts for the three phases, applied prompt patterns, and quality criteria are included in Appendix 1. All prompts were executed between January and March 2025 using ChatGPT-4o.

VP Review

All ten VPs underwent (1) a language & grammar check using DeepL (DeepL SE, Cologne, Germany) [21] followed by a (2) formal and (3) didactical review by a medical educator (IH) using the templates and checklists from the iCoViP project [11]. It was not possible to blind this process as the reviewer also had to implement the necessary changes into the VPs generated by ChatGPT. When information was missing in the ChatGPT-generated VPs, this information was obtained by specific additional prompts (e.g. to provide spirometry results for a child with asthma).

For the subsequent content review (4), five clinicians (AF, IP, MS, NB, TS) each reviewed two VPs (one human-created and one ChatGPT-generated) based on the content review checklist from the iCoViP project [11]. They did not have any knowledge about the generation mode of the VPs they evaluated. After having completed their review in CASUS, they answered a 6-item questionnaire (see Appendix 2) about their overall impression of the VP, how much time they spent on the review, and whether they thought the VP was generated by a human or ChatGPT. Figure 2 summarizes this process.

To evaluate differences and similarities in the quality of human- and ChatGPT-generated VPs and concept maps we qualitatively analyzed and compared the VPs and the review documents.

VP evaluation by students

Based on previously implemented questionnaires [40, 41] to evaluate the quality of VPs for CR training by students we developed a 15-item questionnaire in LimeSurvey [42]. It included general demographic questions, questions about their overall impression of the VP concerning quality and difficulty, and whether they thought they VP was generated by a human or ChatGPT (see Appendix 2).

During May and June 2025, we distributed 84 access codes to medical students at Jagiellonian University, Poland, Bukovinian State Medical University, Ukraine, and University of Porto, Portugal. Each access code was randomly assigned to two different VPs (one ChatGPT-generated, and one human created) in CASUS and students consented to participate in the study upon logging into CASUS. After each completed VP session, students were asked to fill-out the questionnaire.

The study was approved by the ethics board of the Medical University Brandenburg Theodor Fontane (Waiver no 266012025-ANF).

Results

Blueprint

The blueprint covers 53 diagnoses, with the most common diagnoses being low back pain (n = 23), obesity (n = 16), and osteoarthritis (n = 14) (Table 1). Despite a high estimated initial incidence provided by ChatGPT, other common diagnoses, such as bronchitis, sinusitis, and osteoporosis-related fractures are not included. Similarly, diagnoses such as upper respiratory tract infections and hypertension are under-represented in the blueprint compared with the most common primary care encounters [38]. Notably, some diagnoses are less diseases and more signs and symptoms, such as low back pain (n = 23) or neck pain (n = 9) or a disease category, such as chronic liver disease (n = 4).

ChatGPT introduced 13 new key symptoms, such as stiffness, itching, or sneezing, despite being prompted with a list. Thus, 40 VP outlines are based on these newly introduced key symptoms. In addition, 13 key symptoms from this list, such as chest pain, constipation, or syncope are not covered by the blueprint (see Appendix 3 for details).

Although the blueprint is representative of gender and sexual orientation distribution, it includes only two children (1%) and too many adults between the ages of 15 and 64. Furthermore, VPs with a migration background from non-neighboring countries are slightly overrepresented (13.0% vs. 6–9%), while VPs with disabilities are slightly underrepresented (19% vs. 26.8%) (see Table 2 for details).

Table 1 Frequencies of diagnoses estimated by ChatGPT

Diagnosis	Incidence during prompting	Frequency in blueprint	Finley et al. - rank [38]
Low Back Pain	5.0%	11.5% ($n=23$)	9
Obesity (BMI ≥ 30)	1.5%	8.0% ($n=16$)	-
Osteoarthritis	3.5%	7.0% ($n=14$)	4
Neck Pain	2.0%	4.5% ($n=9$)	-
Anxiety Disorders	1.0%	4.0% ($n=8$)	6
Type 2 Diabetes Mellitus	0.5%	3.5% ($n=7$)	5
Chronic Obstructive Pulmonary Disease (COPD)	0.13%	3.5% ($n=7$)	-
Depression	1.0%	3.5% ($n=7$)	6
Peptic Ulcer Disease	0.15%	3.0% ($n=6$)	-
Gout	0.5%	3.0% ($n=6$)	-
Heart Failure	0.12%	3.0% ($n=6$)	-
Hypertension	1.0%	3.0% ($n=6$)	2
Psoriasis	< 0.1%	2.5% ($n=5$)	-
Stroke	0.18%	2.0% ($n=4$)	-
Acute Myeloid Leukemia (AML)	< 0.1%	2.0% ($n=4$)	-
Chronic Kidney Disease	0.15%	2.0% ($n=4$)	-
Community-Acquired Pneumonia	1.5%	2.0% ($n=4$)	7
Chronic Liver Disease	< 0.1%	2.0% ($n=4$)	-
Parkinson's Disease	< 0.1%	2.0% ($n=4$)	-
Rheumatoid Arthritis	0.35%	2.0% ($n=4$)	-

The 20 most frequent diagnoses estimated by ChatGPT during step 2 of prompting and in the final blueprint in comparison to the rank of the diagnosis according to Finely et al. [38]

Most of the VP encounters (47%, $n=94$) take place in an urgent care facility/emergency department, less frequent in a primary care facility (27%, $n=54$) or a specialist clinic (21.5%, $n=43$). Additionally, the blueprint includes home visits (3.5%, $n=7$) and telemedicine consultations (1%, $n=2$). In terms of acuity of the complaints, 47% of VPs are classified as acute ($n=94$), 45.5% as chronic ($n=91$), and 7.5% as subacute ($n=15$).

In terms of internal logic, we found two main aspects that make the blueprint less representative of the European population. First, it contains a high number ($n=82$) of unusual combinations of final diagnosis, key symptom, and sociodemographic data. For example, diagnoses are combined with atypical key symptoms, such as mucosal ulceration in psoriasis ($n=4$) without providing sufficient typical cases ($n=1$) or the diagnosis is rare for the given age (e.g. acute myeloid leukemia in a 19-year-old VP ($n=2$), or the onset is rare, such as stroke and a chronic onset. The blueprint also includes some impossible combinations of diagnosis and weight, such as obesity in VPs with a normal weight ($n=11$ out of 16). Additionally, the combination of occupation and age is unusual, e.g. of the six university/college students, four are 30 years or older, and of the 49 retired VPs, eight are younger than 60 years. Of the 26 VPs with a non-European migration background, only five are of a "non-European" ethnicity.

Second, we noticed repetitive patterns and a lack of variety. For example, the blueprint provides only 21 different occupations, including being retired, a student, or unemployed, and the variety of the fictitious names is also limited. Diagnoses are repeatedly combined with the same key symptom, such as obesity with sleep disorders ($n=11$ out of 16) or acute myocardial infarction with dyspnea ($n=3$ out of 3). The full list of atypical presentations is included in the blueprint in Appendix 3.

VPs

All ten VPs are available in CASUS as a demo course at https://icovip.casus.net. They are similarly structured and of comparable length and underwent formal, didactical, and content review with adaptations implemented after each review round.

During the formal review, DeepL suggested more adaptations for the VPs created by human authors (non-native speakers) and found more misspellings but also suggested a few improvements for the ChatGPT-generated VPs.

During the didactical review we found that the human-created VPs more often did not match the specifications from the blueprint, the dialogs between physician and patient were not continued till the end of the scenario, and the CR concept map was not well aligned with the case progression. For the ChatGPT-generated VPs some clinical data were missing, the scenario offered options for the diagnostic and therapeutic process, and in general the scope of the scenario was often too narrowly focused on the final diagnosis. In general, the manually created VPs showed a greater variety.

Neither the human- nor the ChatGPT-generated VPs contained any medically inaccurate information. However, the content reviewers suggested changes to the diagnostic or management plan for both versions, recommending either additional or fewer tests or treatments (Table 3).

Our findings of the didactical and content review were consistent across the five VPs, and the review of the fifth VP did not provide any additional results. Therefore, we agreed on having reached theoretical saturation for this step.

Image creation

For the five generated patient images several rounds of prompting were required, as some images included obvious inconsistencies such as six fingers or a blend of patient and physician. With the current version of Adobe Firefly, we found an increased likelihood of inconsistencies the more persons were depicted. Therefore, most images are relatively simple, showing only a patient.

Table 2 Overview of sociodemographic distribution of VPs

Criterion	Values	% and *n*	References
Age	0–14 years	1.0% (*n* = 2)	14.7% [43]
	15–64 years	81.5% (*n* = 163)	63.8% [43]
	≥ 65 years	17.5% (*n* = 35)	21.5% [43]
Gender[4]	Female	49.5% (*n* = 99)	51.7% [44]3
	Male	49.0% (*n* = 98)	48.3% [44]3
	Transgender (male & female)	1.5% (*n* = 3)	3.7% (survey data) [45]
Sexual orientation[4]	Heterosexual	91.5% (*n* = 183)	N/A
	Homosexual	4.5% (*n* = 9)	Approx. 9% LGBT+ [46]
	Bisexual	2.5% (*n* = 5)	
	Not stated (children, other)	1.5% (*n* = 3)	N/A
Ethnicity[4]	European	86.0% (*n* = 172)	N/A
	African	3.0% (*n* = 6)	
	Asian	2.0% (*n* = 4)	
	Middle Eastern	7.5% (*n* = 15)	
	Other	1.5% (*n* = 3)	
Occupation	Most frequent occupations	retail worker (*n* = 11), factory worker (*n* = 10), driver (*n* = 10), self-employed (*n* = 10), construction worker (*n* = 9), farmer (*n* = 9), teacher (*n* = 9)	sales workers, office associate professionals, teaching professionals, office professionals, personal service workers [47]
	Unemployed[1]	4.3% (*n* = 6)	5.9% [48]
	Retired	24.5% (*n* = 49)	23.5% (2017) [49]
	Student (University/College)	3.0% (*n* = 6)	Approx. 2.5% [50]
Disability[2,4]	Physical	11.5% (*n* = 23)	26.8% [51]
	Mental	6.0% (*n* = 12)	
	Sensory	1.5% (*n* = 3)	
Migration background[4]	East Asian	2.5% (*n* = 5)	Approx. 6% non-EU citizens, 9% born outside the EU [52]
	South Asian	3.5% (*n* = 7)	
	Latin American	2.0% (*n* = 4)	
	North African	4.5% (*n* = 9)	
	West African	0.5% (*n* = 1)	
	Neighboring country	9.0% (*n* = 18)	
Body Mass Index (BMI) [53]	Underweight	0	3% [54]
	Normal weight	56.0% (*n* = 112)	45% [54]
	Overweight	26.5% (*n* = 53)	36% [54]
	Obese	16.5% (*n* = 33)	17% [54]

Overview of sociodemographic distribution of VPs in the blueprint and comparison to EU data where available

N/A Not applicable

[1]Referring to age group 15-74 years (n=134)

[2]Referring to age group ≥16 years (n=198)

[3]does not include non-binary gender

[4]categorizations suggested by ChatGPT

Evaluation

The evaluation of the content reviewer of the VPs concerning quality, timeliness, and accuracy was comparable for the human- and ChatGPT-generated VPs. The time they spent for the review was on average four hours for both versions and depended more on thoroughness and expertise as a reviewer than the type of VP they reviewed.

Five content reviewers responded that they do not know whether they had reviewed a human- or a ChatGPT-generated VP, three reviewers guessed correctly,

Table 3 Overview of the findings from the formal and didactical review

Aspect	Review type	Human-created	ChatGPT-generated
Alignment with blueprint	DiR	Mismatches ($n=2$)	
Structure	DiR	Well structured	Well structured
Storytelling & Patient	DiR	Storytelling and use of dialog form not continued until the end	Consistent use of dialog form Open variations of scenarios, e.g. stating "If the patient also had any abdominal ultrasound, you might see …" or "medication, such as…"
	CoR	More patient involvement Unnatural patient language	Minor inconsistency of patient image and weight Unnatural patient language
Questions	DiR	"Giving away" the final diagnosis Feedback missing	"Giving away" the final diagnosis Focused on background knowledge not supporting CR
Clinical data	DiR		Narrow lab values Missing data
	CoR	Additional history questions, physical exam findings, tests, details of treatment, additional/changed lab values	Additional history questions, physical exam findings, tests, details of treatment, additional/changed lab values
			Unnecessary examinations (over-diagnostics) Minor inconsistency in timing of treatment and diagnostics
Concept Map	DiR	Additional differentials Connections missing Not aligned with case progression	Additional Tests Differentials too narrow MeSH identifiers incorrect
	CoR	Additional connections Additional differentials Minor inconsistencies Better alignment with case progression	Additional findings Additional treatments Less connections Minor timing inconsistencies
Summary Statement	DiR	Minor inconsistencies	
References	DiR		Links missing
	CoR	Additional/other references suggested	Additional/other references suggested

Findings from the formal and didactical review (DiR) by a medical educator and the content review (CoR) by clinicians

and two were wrong. When looking at the comments the reviewers provided for their assessment similar arguments for a VP being either human or ChatGPT generated. "It is a well-written case, which suggests that it may have been a human", when indeed it was a ChatGPT-generated VP or "Story inconsistency [and] missing information" led to the false assumption of having reviewed a ChatGPT-generated VP. The content reviewers' assessment concerning quality (3.4 for both versions), timeliness (4.2 ChatGPT vs. 4.4 human), and accuracy (3.4 ChatGPT vs. 3.8 human) was comparable for ChatGPT- and human-generated VPs.

In total, 31 students from Jagiellonian University (2nd year), University of Porto (3rd year), and Bukovinian State Medical University (5th and 6th year) participated in the evaluation. 23 students completed two VP sessions, and eight students completed one. Thus, we recorded a total of 54 VP sessions (26 ChatGPT-generated and 28 human-created). There were no significant differences in time spent on ChatGPT- and the human-generated VPs, which averaged 31 min (range: 2–114). Diagnostic accuracy and richness of the concept maps created by the participants for the ChatGPT- and human-generated VPs were comparable.

Table 4 Overview of how students assessed the creation mode of the VP they worked on

Assessed by students as	VP version	
	Human-created	ChatGPT-generated
Human-created VP	11	10
ChatGPT-generated VP	4	8
Do not know	7	4

We received a total of 44 survey responses regarding 22 ChatGPT- and 22 human-generated VPs. 20 responses were from female, 22 from male, and two were from a non-binary participant.

Of the responses eleven (25.0%) indicated that they did not know whether the VP was created by ChatGPT or a human, 19 responses (43.2%) were correct, and 14 (31.8%) were incorrect, mainly suggesting that a ChatGPT-generated VP was created by a human ($n=10$) (see Table 4). There were differences between the VPs, with two VPs being mostly rated as human-created for both versions.

Students assessed the ChatGPT-generated VPs, or those they thought were created by ChatGPT often as "straight forward" and "typical" scenarios that are easy to follow and do not include irrelevant information or

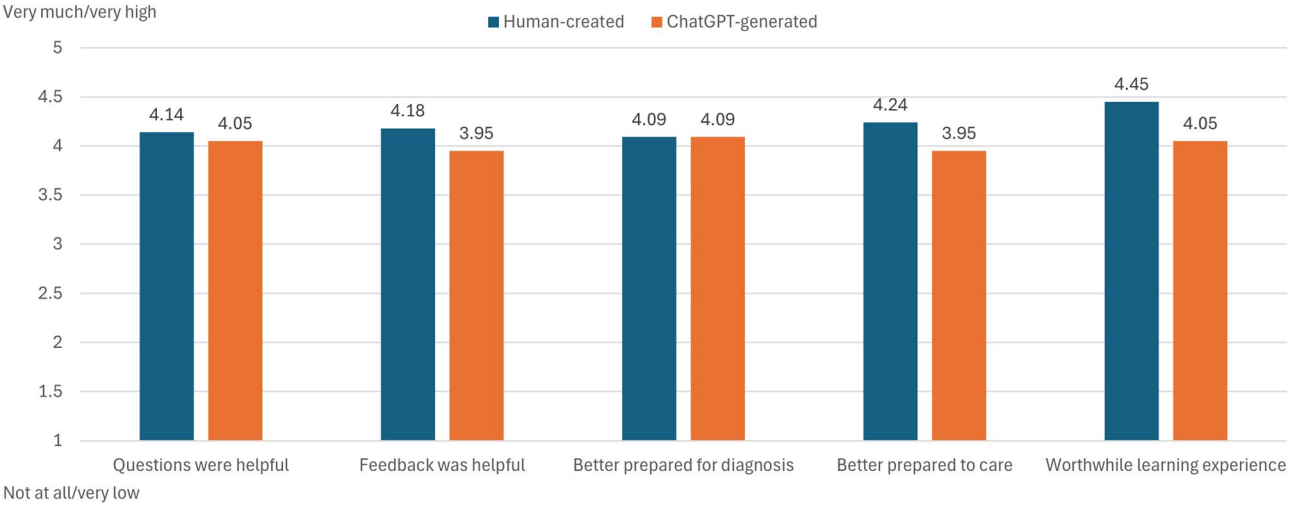

Fig. 3 Students responses ($n = 44$) concerning helpfulness, preparedness, and learning experience of the VPs. 1 = strongly disagree, 5 = strongly agree

details. For example, a student wrote "The language used was straight forward and it explained everything in an easy-to-understand way. The case presented typical a presentation of disease and was easy to follow".

On the other hand, students associated human-created VPs or VPs supposedly created by humans with being realistic, providing details and emotional responses. For example, one student wrote that the VP "had some specific, field-related information, which was not obvious, and it stands out to me". Another student, who believed a ChatGPT-generated VP was human-created, justified this by saying, "because it was realistic with emotions and real-life scenarios".

Regarding the level of difficulty, eleven responses indicated that the VP was too easy ($n = 3$ for human-created and $n = 8$ for ChatGPT-generated VPs). However, it appears that particularly one of the ChatGPT-generated VPs was rated as being too easy.

Ratings of the human-created VPs were slightly better in terms of helpfulness of questions and feedback, feeling better prepared for patient care, and considering the VP a worthwhile learning experience (see Fig. 3).

Discussion

Former research has shown a promising potential of generative AI, often ChatGPT, in creating educational content, such as multiple-choice questions [55, 56] assessment items [57], feedback to learners [58], and VPs and case vignettes [32, 33]. In our study we investigated the potential of ChatGPT for the entire process of planning and creating a VP collection for CR training.

Blueprint generation

Overall, the generated blueprint did not meet our expectations in terms of quality or representativeness. A main issue arose in the first step of the prompt, when ChatGPT could not provide a reasonable list of common diseases, which led to inconsistencies in the subsequent steps. These findings are consistent with previous studies showing that ChatGPT and generative AI in general has limitations in differentiating between reliable and unreliable sources, recalling details, understanding conceptual relationships, and adhering to standardized guidelines [31, 59]. Another issue is the repetitive patterns and relatively high number of unusual VP outlines. To evaluate ChatGPT's capabilities, we deliberately did not manually adapt the prompt outputs and did not discuss necessary adaptations. However, such discussions about the blueprint fostered a deeper understanding and awareness of the importance of a diverse VP collection among the research team in the iCoViP project. These discussions also led to fruitful conversations about differences among partner countries and enabled us to deliberately amend the blueprint for educational purposes. Furthermore, when discussing the blueprint, the VP authors developed some initial ideas for creating "their" VPs. This sense of ownership and responsibility for a VP did not emerge with the ChatGPT-generated VPs - a phenomenon that has also been described for essay writing in a recent study by Kosmyna et al. [60]. In summary, while we do not consider the blueprint created by ChatGPT to be usable in its current form, we believe that the process of prompting combined with discussions and refinements after each step could improve quality and representativeness. Such an approach combines ChatGPT's efficiency in quickly providing a basis for discussions with the involvement of educators and clinicians in the creation process. Given the rapid advancements in LLMs, the outputs could significantly improve within the next few months or years, providing a more reliable foundation for discussions.

Creation of VPs

In contrast to a study by Wong et al., in which ChatGPT-generated cases included inaccurate information and hallucinations, neither the human-created nor the ChatGPT-generated VP scenarios and concept maps in our study contained any medically wrong information [32]. Furthermore, our analysis revealed that the results from the didactical and content reviews were similar and content reviewers and students could not reliably discern significant differences when presented with one human-created and one ChatGPT-generated VP. These results are consistent with previous studies showing that LLMs can generate high-quality VPs or case vignettes [26, 33]. Unlike Takahasi et al., the ChatGPT-generated VPs in our study provided consistent and realistic patient-physician dialogs [25] as suggested by Moser et al. in their tip 4 [27].

These improvements in the VP quality compared to previous studies are most likely due to technical advances and differences in the prompt design.

However, considering all VPs, the five human-created VPs demonstrated a greater variety in storytelling, differential diagnosis, and patient and doctor interaction. Since five authors from three different countries created the VPs, there was naturally greater variety in how they wrote such scenarios than with ChatGPT. Therefore, when creating a large pool of ChatGPT-generated VPs, the collection may lack the variety of styles and storytelling that each human author brings to the creation process.

Student evaluations confirm this, showing a tendency to rate the ChatGPT-generated VPs lower than the human-created VPs concerning aspects in terms of helpfulness of questions and feedback, preparation for patient care, and overall rating. Additionally, students associated the VPs presumably generated by ChatGPT with typical presentations and an inability to generate complex scenarios.

We were able to generate realistic, albeit simple, patient images using Adobe Firefly, as was done by Bakkum et al. [33] and suggested by Moser et al. [27]. However, we were unable to generate medically accurate clinical images with the currently available AI tools, confirming previous findings by Moser et al. [27] and Benítez et al. [61]. Given the rapid development of AI tools, we anticipate that this will improve further and become feasible in the near future. Both the iCoViP project and this study involved significant efforts to retrieve real-world clinical images, including legal discussions with healthcare institutions and obtaining patient consent. AI-supported generation could reduce this effort but would still require careful review by subject matter experts to assess the accuracy and appropriateness of the generated images.

Comparison of human-created and ChatGPT-generated VPs

After designing and piloting the prompts for the VP generation, the creation process initiated by executing the prompts was mostly automated. The resulting VPs and the concept maps could be imported into the VP system CASUS for further editing. Depending to some extent on their experience, VP creation by human authors is time-consuming and resource intensive [62]. Clinicians need at least a few hours to create one VP, often spreading the work over a few weeks to fit into their busy schedules. In addition to the limited availability of clinical educators, they often feel more comfortable when basing a VP on patients they have encountered in their practice. This can make it difficult to recruit experienced authors for certain topics and increases the likelihood of an imbalanced VP collection. ChatGPT generates VPs independently of such preferences or experience. Therefore, we conclude that the initial VP creation process can be supported using ChatGPT.

However, all VPs (human- or ChatGPT-generated) need to be reviewed and refined regarding formal, didactical, and content-related aspects. For these steps the ChatGPT-generated VPs did not save time. Review processes benefit from interactions between reviewers and authors, as new ideas arise that enrich the case scenario.

Thus, even when the initial VPs are generated by ChatGPT we suggest assigning dedicated authors to the VPs for refinement rounds based on the review results. Assigning different authors increases the variety lacking in the initial VP versions and enables fruitful collaboration and exchange between reviewers and authors.

One aspect to keep in mind is the possibility of deskilling [63] and automation bias among VP authors and reviewers. To review and refine ChatGPT-generated VPs, they still need training and experience in creating and reviewing VPs, which could be lost by relying on ChatGPT [64].

Limitations

Despite careful planning, our study has some limitations. First, due to constraints in our team composition we could not blind the formal and didactical review, so we cannot exclude that the comments and feedback of these review steps might have been biased. However, the didactical reviewer assessed the VPs based on a previously standardized guideline and a checklist of quality criteria.

Second, unlike in some previous studies [31], we deliberately did not run several rounds of the same prompts to choose the best output from. Instead, we piloted and iteratively refined the prompts based on quality criteria and then prompted ChatGPT once.

Although this might have improved the output, it would have been time-consuming to compare the outputs and would be an unrealistic endeavor for educators to do. Third, in some instances we were missing accurate references for some EU data items, which made it difficult to reliably judge representativeness. Additionally, some categories assigned by ChatGPT were unrealistic for comparison with EU references, e.g., ethnicity or disability.

Future research should investigate whether ChatGPT can support the didactical and content review and regular update process in addition to planning and creating VPs. Furthermore, it would be interesting to investigate how to avoid de-skilling and automation bias among authors and reviewers when AI tools are routinely used to create educational material, such as VPs. Finally, future studies could explore the potential of applying custom AI tools for specific steps in the VP creation process, such as generating clinical images, in order to improve the results.

Conclusions

The results of our study confirm the potential of generative AI in creating educational content on the specific example of a VP collection for CR training. In addition, the study provides relevant details on how to organize such a multi-step process and emphasizes the importance of combining ChatGPT's ability to quickly generate blueprints and VPs with humans' strengths, such as carefully assessing and refining ChatGPT's outputs and providing cultural and contextual variety relevant for CR training.

Each step of the blueprinting and VP creation process including the CR concept map can be initiated by ChatGPT using and adapting our published prompts, but all subsequent steps of refining the outputs require the inputs of human experts. We recommend holding structured discussion rounds with stakeholders for each step and making adaptations accordingly before starting the next step. This way ChatGPT can support the process, while ensuring that the necessary and fruitful discussions take place to guarantee high-quality outputs. Although human factors make the VP creation process time-consuming, the variety of VP scenarios may be negatively affected without interactions between authors, didactic reviewers, and content reviewers, and biases may be more likely to go undetected.

Abbreviations

AI	Artificial Intelligence
BMI	Body Mass Index
CR	Clinical reasoning
iCoViP	International collection of virtual patients project
LLM	Large Language Models
VPs	Virtual patients

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12909-025-08006-9.

Appendix 1: Prompts & quality criteria.

Appendix 2: Content reviewer and student questionnaires.

Appendix 3: Blueprint and internal consistency check results.

Appendix 4: Image generation.

Acknowledgements

We would like to thank our colleagues Begoña Martinez Jarreta, Begoña Abecia, Daniela Antunes, Léa Linglart, and Tetiana Antofiichuk for supporting this study with their advice and encouragement. We also thank Martin Adler who supported the import of ChatGPT generated VPs into CASUS and the implementation of the student questionnaire in LimeSurvey. Finally, we want to thank all students who participated in the study and provided valuable feedback.

Authors' contributions

JF made substantial contributions to the design of the study, data collection, analysis, and interpretation of results and the drafting and revision of the manuscript. AK made substantial contributions to the design of the study, analysis, and interpretation of results, and the drafting and revision of the manuscript. NB, VD, ND, AF, II, LM, NP, IP, TS, MS, RS made substantial contributions to the data collection and critically revised the manuscript. AM made substantial contributions to the data collection, analysis, and interpretation or results and critically revised the manuscript. IH made substantial contributions to the design of the study, data collection, analysis, and interpretation and the drafting and revision of the manuscript. All authors approved the submitted version and have agreed both to be personally accountable for their own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

Funding

The study was partially funded by internal funds from Jagiellonian University Medical College.

Data availability

The prompts and results of the study are included as appendices. Additional material, such as the intermediate outputs of the ChatGPT-generated blueprint, VPs, and concept maps are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

The study complies with the principles outlined in the Declaration of Helsinki. The study was approved by the ethics board of the Medical University Brandenburg Theodor Fontane (Waiver no 266012025-ANF). All participants voluntarily participated, and we obtained their informed consent prior to their involvement in the study.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Centre for Innovative Medical Education, Jagiellonian University Medical College, Kraków, Poland
[2]Department of Bioinformatics and Telemedicine, Faculty of Medicine, Jagiellonian University Medical College, Kraków, Poland
[3]Bukovinian State Medical University, Chernivtsi, Ukraine
[4]Unidade Local de Saúde de Santo António, Porto, Portugal
[5]Instituto de Biociências Abel Salazar, University of Porto, Porto, Portugal

[6]Growth, Exercise, Nutrition and Development (EXER-GENUD) Research Group, Universidad de Zaragoza, Zaragoza, Spain
[7]Instituto de Investigación Sanitaria de Aragón (IIS Aragón), Zaragoza, Spain
[8]Instituto Agroalimentario de Aragón-IA2 (Universidad de Zaragoza-CITA), Zaragoza, Spain
[9]Centro de Investigación Biomédica en Red de Fisiopatología de la Obesidad y Nutrición (CIBERObn), Madrid, Spain
[10]University of Augsburg, Augsburg, Germany
[11]Pediatric Intensive Care Unit, Bicetre Hospital and Faculty of Medicine, APHP Paris Saclay University, Le Kremlin-Bicetre, France
[12]Institute of Research in Health Sciences Education, Faculty of Health Sciences Brandenburg, Brandenburg Medical School Theodor Fontane (MHB), Neuruppin, Germany

Received: 26 June 2025 / Accepted: 15 September 2025
Published online: 02 October 2025

References

1. Kononowicz AA, Woodham LA, Edelbring S, Stathakarou N, Davies D, Saxena N, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res. 2019. https://doi.org/10.2196/14676.
2. Plackett R, Kassianos A, Kambouri M, Kay N, Mylan S, Hopwood J, Schartau P, Gray S, Timmis J, Bennett S, et al. Online patient simulation training to improve clinical reasoning: A feasibility randomised controlled trial. BMC Med Educ. 2020;20:245.
3. Watari T, Tokuda Y, Owada M, Onigata K. The utility of virtual patient simulations for clinical reasoning education. Int J Environ Res Public Health. 2020;17:5325.
4. Penner JC, Schuwirth L, Durning SJ. From noise to music: reframing the role of context in clinical reasoning. J Gen Intern Med. 2024;39(5):851–7.
5. Schuler K, Jung IC, Zerlik M, Hahn W, Sedlmayr M, Sedlmayr B. Context factors in clinical decision-making: a scoping review. BMC Med Inform Decis Mak. 2025;25(1):133.
6. Hege I, Kononowicz AA, Berman NB, Lenzer B, Kiesewetter J. Advancing clinical reasoning in virtual patients - development and application of a conceptual framework. GMS J Med Educ. 2018;35(1):Doc12.
7. Huang G, Reynolds R, Candler C. Virtual patient simulation at US and Canadian medical schools. Acad Med. 2007;82(5):446–51.
8. Zary N, Hege I, Heid J, Woodham L, Donkers J, Kononowicz AA. Enabling interoperability, accessibility and reusability of virtual patients across Europe-design and implementation. Stud Health Technol Inf. 2009;150:826–30.
9. Berman NB, Fall LH, Chessman AW, Dell MR, Lang VJ, Leong SL, et al. A collaborative model for developing and maintaining virtual patients for medical education. Med Teach. 2011;33(4):319–24.
10. Abdulnour RE, Parsons AS, Muller D, Drazen J, Rubin EJ, Rencic J. Deliberate practice at the virtual bedside to improve clinical reasoning. N Engl J Med. 2022;386(20):1946–7.
11. iCoViP project and review documents. https://www.icovip.eu. Accessed 15 June 2025.
12. iCoViP collection of virtual patients in CASUS. https://crt.casus.net. Accessed 15 June 2025.
13. iCoViP blueprint. http://blueprint.icovip.eu. Accessed 15 June 2025.
14. Mayer A, Da Silva Domingues V, Hege I, Kononowicz AA, Larrosa M, Martínez-Jarreta B, et al. Planning a collection of virtual patients to train clinical reasoning: a blueprint representative of the European population. IJERPH. 2022;19(10):6175.
15. Hege I, Kononowicz AA, Adler M. A clinical reasoning tool for virtual patients: Design-Based research study. JMIR Med Educ. 2017;3(2):e21.
16. Chang RW, Bordage G, Connell KJ. The importance of early problem representation during case presentation. Acad Med. 1998;73(10):S109-11.
17. Huwendiek S, Reichert F, Bosse HM, de Leng BA, van der Vleuten CP, Haag M, et al. Design principles for virtual patients: a focus group study among students. Med Educ. 2009;43(6):580–8.
18. Botezatu M, Hult H, Fors UG. Virtual patient simulation: what do students make of it? A focus group study. BMC Med Educ. 2010;4:10:91.
19. Kim S, Phillips WR, Pinsky L, Brock D, Phillips K, Keary J. A conceptual framework for developing teaching cases: a review and synthesis of the literature across disciplines. Med Educ. 2006;40(9):867–76.
20. Fischer MR. CASEPORT: systemintegrierendes sortal für die fallbasierte lehre in der medizin. In: Jäckel A., editor Telemedizinführer Deutschland, Ober-Mörlen: Medizin Forum AG, 2003:146–50.
21. DeepL. https://www.deepl.com. Accessed 15 June 2025.
22. Cook DA. Creating virtual patients using large language models: scalable, global, and low cost. Med Teach. 2024;(1):3. https://doi.org/10.1080/0142159X.2024.2376879.
23. Benítez TM, Xu Y, Boudreau JD, Kow AWC, Bello F, Van Phuoc L, et al. Harnessing the potential of large Language models in medical education: promise and pitfalls. J Am Med Inform Assoc. 2024;31(3):776–83.
24. Silvestri-Elmore A, Burton C. How can nursing faculty create case studies using AI and educational technology? Nurse Educ. 2025;50(1):35–9.
25. Takahashi H, Shikino K, Kondo T, Komori A, Yamada Y, Saita M, Naito T. Educational utility of clinical vignettes generated in Japanese by ChatGPT-4: mixed methods study. JMIR Med Educ. 2024;10:e59133.
26. Berbenyuk A, Powell L, Zary N. Feasibility and educational value of clinical cases generated using large language models. In: Mantas J, Hasman A, Demiris G, Saranto K, Marschollek M, Arvanitis TN, et al. editors. Studies in health technology and informatics. IOS, Amsterdam, The Netherlands. 2024;316:1524–8. https://ebooks.iospress.nl/doi/10.3233/SHTI240705.
27. Moser M, Posel N, Ganescu O, Fleiszer D. Twelve tips: using generative AI to create and optimize content for virtual patient simulations. Med Teach. 2025:1–7. https://doi.org/10.1080/0142159X.2025.2501252.
28. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health. 2023;5(4):e179–81.
29. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ. 2023;9(1):e46885.
30. Combs CD, Combs PF. Emerging roles of virtual patients in the age of AI. AMA J Ethics. 2019;21(2):E153.
31. Zack T, et al. Assessing the potential of GPT-4 to perpetuate Racial and gender biases in health care: a model evaluation study. Lancet Digit Health. 2024;6(1):e12–22.
32. Wong K, Fayngersh A, Traba C, Cennimo D, Kothari N, Chen S. Using ChatGPT in the Development of Clinical Reasoning Cases: A Qualitative Study. Cureus. 2024;16(5):e61438.
33. Bakkum MJ, et al. Using artificial intelligence to create diverse and inclusive medical case vignettes for education. Br J Clin Pharmacol. 2024;90(3):640–8.
34. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ. 2024. https://doi.org/10.1111/medu.15402.
35. Hale J, Seth A, Towner Wright S, Gilliland K. Generative AI in undergraduate medical education: a rapid review. J Med Educ Curric Dev. 2024;11:1–15.
36. https://platform.openai.com/docs/guides/prompt-engineering. Accessed 15 June 2025.
37. White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv:2302.11382v1 [cs.SE]; 2023. https://doi.org/10.48550/arXiv.2302.11382.
38. Finley CR, Chan DS, Garrison S, Korownyk C, Kolber MR, Campbell S, et al. What are the most common conditions in primary care? Systematic review. Can Fam Physician. 2018;64(11):832–40.
39. Szydlak R, Kiyak YS, Hege I, Torre D, Kononowicz AA. Comparison of human and GPT-generated concept maps in a clinical reasoning collection of educational virtual patients. Stud Health Technol Inf. 2025;327:1024–28.
40. Huwendiek S, De Leng BA, Kononowicz AA, Kunzmann R, Muijtjens AMM, Van Der Vleuten CPM, Hoffmann GF, Tönshoff B, Dolmans D. Exploring the validity and reliability of a questionnaire for evaluating virtual patient design with a special emphasis on fostering clinical reasoning. Med Teach. 2015;37(8):775–82.
41. iCoViP. evaluation tools. https://icovip.eu/knowledge-base/evaluation-tools/. Accessed 15 June 2025.
42. LimeSurvey GmbH. https://www.limesurvey.org. Accessed 15 June 2025.
43. European Union. Age distribution of inhabitants from 2013 – 2023. https://www.statista.com/statistics/253408/age-distribution-in-the-european-union-eu/. Accessed 15 June 2025.
44. Estimated population of Europe. from 1950 to 2024, by gender. https://www.statista.com/statistics/755225/population-of-europe-by-gender/. Accessed 15 June 2025.
45. Russell CB, Sanders F, Watkins F. Intersections. Diving into the FRAU LGBTI II survey data – Trans and non-binary briefing. Available from https://www.ilga-europe.org/files/uploads/2023/07/FRA-Intersections-Report-Trans-Non-binary.pdf. Accessed 15 June 2025.

46. Pride month 2024. 9% of adults identify as LGBT+. https://www.ipsos.com/sites/default/files/ct/news/documents/2024-06/Pride-Report-2024_2.pdf. Accessed 15 June 2025.
47. Occupation employment in sector. https://www.cedefop.europa.eu/en/tools/skills-intelligence/occupation-employment-sector?year=2023&country=EU. Accessed 15 June 2025.
48. Unemployment statistics and beyond. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Unemployment_statistics_and_beyond Accessed 15 June 2025.
49. Ageing Europe - statistics on pensions. income and expenditure https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Ageing_Europe_-_statistics_on_pensions,_income_and_expenditure#Source_data_for_tables_and_graphs Accessed 15 June 2025.
50. Students enrolled in. tertiary education by education level, programme orientation, sex and age. https://ec.europa.eu/eurostat/databrowser/view/EDUC_UOE_ENRT02__custom_2559245/bookmark/table?lang=en&bookmarkId=6f6941ab-9c03-4ba6-a498-0c62e4cad301 Accessed 15 June 2025.
51. Self-reported disability. (limitation in usual activities due to health problems. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_with_disability#Self-reported_disability_.28limitation_in_usual_activities_due_to_health_problems.29 Accessed 15 June 2025.
52. People living in the EU. in 2023. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/promoting-our-european-way-life/statistics-migration-europe_en Accessed 15 June 2025.
53. Glossary. Body mass index (BMI). https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Body_mass_index_(BMI) Accessed 15 June 2025.
54. Over half of adults in the EU are overweight. 2021. https://ec.europa.eu/eurostat/de/web/products-eurostat-news/-/ddn-20210721-2 Accessed 15 June 2025.
55. Kıyak YS, Emekli E. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. Postgrad Med J. 2024;100(1189):858–65.
56. Rezigalla AA. AI in medical education: uses of AI in construction type A MCQs. BMC Med Educ. 2024;24(1):247.
57. Lam G, Shammoon Y, Coulson A, Lalloo F, Maini A, Amin A, et al. Utility of large language models for creating clinical assessment items. Med Teach. 2025;47(5):878–82.
58. Çiçek FE, Ülker M, Özer M, Kıyak YS. ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: a randomized controlled trial. Postgrad Med J. 2025;101(1195):458–63.
59. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595.
60. Kosmyna N, Hauptmann E, Yuan YT, Situ J, Liao XH, Beresnitzky AV, Braunstein I, Pattie M. Your brain on ChatGPT: accumulation of cognitive debt when using an AI assistant for essay writing task. arXiv: 2506.08872; 2025. Available from: http://arxiv.org/abs/2506.08872.
61. Benítez TM, Xu Y, Boudreau JD, Kow AWC, Bello F, Van Phuoc L, Wang X, Sun X, Leung GK, Lan Y, Wang Y, Cheng D, Tham YC, Wong TY, Chung KC. Harnessing the potential of large language models in medical education: promise and pitfalls. J Am Med Inf Assoc. 2024;31(3):776–83.
62. Stretton B, Kovoor J, Arnold M, Bacchi S. ChatGPT-based learning: generative artificial intelligence in medical education. Med Sci Educ. 2024;34(1):215–7.
63. Choudhury A, Chaudhry Z. Large language models and user trust: consequence of self-referential learning loop and the deskilling of health care professionals. J Med Internet Res. 2024;26:e56764.
64. Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu NY, Bartlett R, et al. A scoping review of artificial intelligence in medical education: BEME guide 84. Med Teach. 2024;46(4):446–70.

Publisher's Note