Serious Games in a Graduate-Level Human Physiology Course

Arne Kristian Christensen
Westfield State University, Westfield, Massachusetts, USA

achristensen@westfield.ma.edu

Abstract: Healthcare, like most industry sectors, is facing the transformative implications of Artificial Intelligence (AI). We do not yet fully understand the benefits and pitfalls of AI, so it is critical that training in healthcare-related fields includes opportunities to learn how to leverage the former while being alert to the latter. We aim to evaluate the integration of serious games, as scenario-based teaching strategies, into a graduate-level human physiology course. Students use a publicly available large language model to diagnose a series of tailored medical case studies that have been carefully crafted to highlight the utility of AI while revealing its limitations. The course modality is online, alternating between asynchronous and synchronous instruction. Asynchronous course material includes video presentations and group discussions of peer-reviewed medical case studies, which serve to prepare the students for a tailored case study with a similar or related condition. Synchronous course meetings include an individual readiness assurance test, which covers the preparation materials, to incentivise students to develop a robust understanding of the relevant pathophysiology prior to meeting. Individuals are then grouped into randomised teams, using AI to interpret a case within a serious game framework. Teams are prompted with questions and respond using a scratch ticket-style scoring sheet that allows partial credit in a point-based game. Teams navigate hallucinated responses, apply prior knowledge, and prompt AI toward correct diagnoses. Evaluation of course participants' attitudes and opinions of the course methods will be conducted by survey. In addition to the course, the process for developing effective case studies that fit into this model is described, and the potential for extending this strategy to an undergraduate human anatomy and physiology curriculum is also explored. Moreover, opportunities for raising ethical considerations with the use of AI in healthcare are considered. Together, this study is an effort to respond to the rapid changes AI is bringing to healthcare education and may serve as a model for highlighting the utility and pitfalls of AI in other fields.

Keywords: Physiology, Healthcare, Games-Based learning, Serious games, Artificial intelligence, Generative AI

1. Introduction

Serious games are scenario-based instructional activities that have been shown to improve student motivation and engagement (Al-Awadai, 2024). These activities, using elements of game design such as challenges, goals, cooperation, and feedback, represent an emerging area of educational research focused on identifying which features are most successful in improving student learning (Silva, 2020). Serious games have been successfully implemented in courses across a range of disciplines, including those related to anatomy, physiology, and medicine, (Gorbanev et al, 2018; Moro, Phelps and Stromberga, 2020). The aim of the present study is to serve as a proof of principle for integrating serious games into a graduate-level human physiology course using tailored medical case studies, and student assistance from generative AI, as the situational context.

2. Game Context and Design

2.1 Course Context

Serious games will be integrated into Human Physiology (BIOL 0570; Westfield State University), a graduate-level course for students in healthcare-related programmes or high school science teachers meeting licensure requirements. The online course blends asynchronous and synchronous modality. Asynchronous course material provides students with the knowledge they apply during the synchronous meetings, when student teams engage with a tailored medical case that focuses on a medical condition. The first three cases are designed by the course instructor, who also serves as a guide through the exercise. The rationale for the instructor presenting the first three cases is to model serious games elements. The roles are reversed later in the course, students develop a case and guide their peers through it.

Case studies must align with the course learning objectives, incorporate key serious game attributes, including scaffolding, rules and gameplay, rewards, feedback, and reflection (Silva, 2020; Al-Awadai, 2024), and include this progression of activities:

- Individual Readiness Assurance Test
- Team Readiness Assurance Test
- Team-Based Case Study

2.2 Individual Readiness Assurance Test

Students are required to read and comment on assigned peer-reviewed case studies related to a specific condition to prepare for each synchronous Zoom meeting. At this meeting students answer five multiple-choice questions related to the preparation materials. This individual Readiness Assurance Test (iRAT) is mediated via Google Forms. The iRAT, and the next activity, the team Readiness Assurance Test (tRAT), are modelled after the Team Based Learning (TBL) technique introduced by Michaelsen and colleagues (2002).

2.3 Team Readiness Assurance Test

Following the iRAT, students are organised into teams of three or four students that are randomised, which reduces social barriers and increases the mobility of knowledge between students (Liljedahl, Zager and Wheeler, 2021). Teams are assigned to Zoom breakout rooms to complete the tRAT, which is the same quiz as the iRAT, but each team responds on an Immediate Feedback Assessment Technique (IF-AT) form (Epstein Educational Enterprises, Cincinnati, OH). IF-AT forms are multiple choice (A-E) answer sheets with each choice covered by a metallic film that can be scratched off (similar to some lottery tickets). Teams can continue to discuss and scratch response options until they identify the correct response; 1 point for the first scratch, 0.5 points for the second scratch, or 0.25 points for a third (and last chance) scratch.

2.4 Team-Based Case Study

For the final activity, teams work through a tailored case related to the preparation material and tests. The case unfolds as a presentation (e.g. Google Slides), and can include a simple narrative, images of the patient, medical images, and lab values. Five questions are interspersed throughout the case, which may focus on telling clues, additional tests or information that may be required, or to determine a diagnosis. Teams respond on the IF-AT form, providing immediate feedback and determining the winning team of the serious game.

3. Generative AI

Generative AI has been shown to be prone to hallucinations and errors in accurately citing research (Alkaissi and McFarlane, 2022) and has been trained on data that reflect societal and cultural biases, compromising efficacy as a medical diagnostic tool (Cirillo et al, 2020). Recent studies have shown that when ChatGPT is challenged with medical case studies, it has a misdiagnosis rate of 51% with ChatGPT-3.5 (Hadi et al, 2024), 72% using paediatric cases with ChatGPT-3.5 (Barile et al, 2024), and 61% in complex cases with ChatGPT-4 (Kanjee, Crowe and Rodman, 2024).

The instructor-designed cases for the Human Physiology course have been tailored to exploit mismatches in how well the materials used to train large language models reflect a given clinical context. When cases omit terms that are typically paired with a condition, emphasize unrelated systems, or include signs that masquerade as another condition, ChatGPT is more likely to misdiagnose. ChatGPT was employed to create these “distractors”, or exploit known biases, to lead to a misdiagnosis. Students prepare for a case study by reading relevant peer-reviewed cases, then work in teams to navigate a case that is likely to lead ChatGPT to a misdiagnosis (Table 1). The case development offers an opportunity for participants designing cases to think deeply about what guides AI to a correct or incorrect diagnosis, and moreover, participants in the serious game must consider how they might shepherd ChatGPT in the right direction.

The following prompt was used to create the signs and symptoms that would be included in a case study that would likely lead ChatGPT to misdiagnose a condition:

> *Prompt 1: I am creating case studies as a teaching tool in a human physiology course. One of the course learning objectives is to demonstrate that ChatGPT can be a useful tool for interpreting patient signs and symptoms but requires human oversight because ChatGPT has been shown to misdiagnose cases and introduce bias and misinformation. Create a set of signs and symptoms for a patient with _____ that would lead ChatGPT to misdiagnose the condition.*

In all cases, ChatGPT came to a correct diagnosis with the first iteration despite distractor signs. ChatGPT was prompted to revise the signs and symptoms until they led to a misdiagnosis with the following prompt:

> *Revision Prompt: The signs and symptoms are leading ChatGPT to a correct diagnosis, modify the case signs and symptoms to increase the likelihood of a misdiagnosis.*

Table 1: Conditions used for tailored cases with the distractors that mislead ChatGPT

Condition (Sample student reading)	ChatGPT distractors	ChatGPT misdiagnoses (Required number of revisions)
Rhabdomyolysis **(Al Badi, Rasbi and Alalawi 2020).**	De-emphasize muscle involvement Remove trigger signs (e.g., brown urine) and excessive exercise	Viral myocarditis Early heart failure (Third revision)
Tetrodotoxin **(Almeida et al, 2019)**	Do not include seafood ingestion Highlight neurological presentations	Ischemic stroke (Second revision)
Endometriosis **(Hudson, 2021)**	Do not include menstrual cycle timing Add psychological symptoms	Irritable bowel syndrome Celiac disease (Second revision)

4. Evaluation

The evaluation of the project will be performed via a participant survey to measure respondents' attitudes and opinions of TBL, serious game elements, and inclusion of generative AI. The elements of serious games, and AI integration, that are successful will be trialled in our Human Anatomy and Physiology undergraduate-level courses. These courses have large enrolments that can be leveraged for quantitative evaluation of the methods.

Acknowledgements

Simon Holdaway was involved in many discussions about the use of AI as a learning tool and these discussions helped guide how this project would integrate AI. Westfield State University provided support for this project.

Ethics Declaration: Ethics clearance was not needed for this paper.

AI Declaration: ChatGPT-4o was used to proof for spelling and grammar errors and generate signs and symptoms leading to AI misdiagnosis.

References

Al-Awadai, Z. (2024) A Conceptual Framework for the Development of Effective Serious Games for Learning. *18th European Conference on Games Based Learning (ECGBL)*. DOI: https://doi.org/10.34190/ecgbl.18.1.2974

Al Badi, A., Al Rasbi, S. and Alalawi, A. M. (2020) Exercise-Induced Rhabdomyolysis: A Case Report and Literature Review. *Cureus*, *12*(8), e10037. https://doi.org/10.7759/cureus.10037

Alkaissi, H. and McFarlane, S. I. (2023) Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*, *15*(2), e35179. https://doi.org/10.7759/cureus.35179

Almeida, P., Diaz, R., Hernandez, F. and Ferrer, G. (2019) Blow: A Case of Pufferfish Intoxication in South Florida. *BMJ Case Reports*, *12*(6), e229272. https://doi.org/10.1136/bcr-2019-229272

Aolayan, K., Almohammadi, T. and Alotaibi, A. (2023) A Diaphragmatic Hernia in a Traumatic Patient Simulating a Hemorrhage: A case report. *Trauma Case Reports*, *43*, 100754. https://doi.org/10.1016/j.tcr.2023.100754

Barile, J., Margolis, A., Cason, G., Kim, R., Kalash, S., Tchaconas, A. and Milanaik, R. (2024) Diagnostic Accuracy of a Large Language Model in Pediatric Case Studies. *JAMA Pediatrics*, *178*(3), 313–315. https://doi.org/10.1001/jamapediatrics.2023.5750

Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M. J., Chadha, A. S. and Mavridis, N. (2020) Sex and Gender Differences and Biases in Artificial Intelligence for Biomedicine and Healthcare. *NPJ Digital Medicine*, *3*, 81. https://doi.org/10.1038/s41746-020-0288-5

Gorbanev, I., Agudelo-Londoño, S., González, R. A., Cortes, A., Pomares, A., Delgadillo, V., … Muñoz, Ó. (2018) A Systematic Review of Serious Games in Medical Education: Quality of Evidence and Pedagogical Strategy. *Medical Education Online*, *23*(1). https://doi.org/10.1080/10872981.2018.1438718

Hadi, A., Tran, E., Nagarajan, B. and Kirpalani, A. (2024) Evaluation of ChatGPT as a Diagnostic Tool for Medical Learners and Clinicians. *PloS One*, 19(7), e0307383. https://doi.org/10.1371/journal.pone.0307383

Hudson N. (2021) The Missed Disease? Endometriosis as an Example of 'Undone Science'. *Reproductive Biomedicine & Society Online*, *14*, 20–27. https://doi.org/10.1016/j.rbms.2021.07.003

Kanjee, Z., Crowe, B. and Rodman, A. (2023) Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. *JAMA, 330*(1), 78–80. https://doi.org/10.1001/jama.2023.8288

Liljedahl, P., Zager, T. and Wheeler, L. (2021) *Building Thinking Classrooms in Mathematics, Grades K-12: 14 Teaching Practices for Enhancing Learning.* Corwin Mathematics.

Michaelsen, L. K., Knight, A. B. and Fink, L. D. (Eds.). (2023) *Team-based Learning: A Transformative Use of Small Groups in College Teaching*. Taylor & Francis.

Moro, C., Phelps, C. and Stromberga, Z. (2020) Utilizing Serious Games for Physiology and Anatomy Learning and Revision. *Advances in Physiology Education*, *44*(3), 505–507. https://doi.org/10.1152/advan.00074.2020

Silva, F. G. M. (2020) Practical Methodology for the Design of Educational Serious Games. *Information*, *11*(1), 14. https://doi.org/10.3390/info11010014