









# Democratizing Virtual Patient Case Creation: A Proof-of-concept Technical Framework for Clinicians

Nikolaos Tsaftaridis, MD  \*<sup>1,2</sup>, Ioannis Koulas, MD, MSc <sup>2</sup>, Stefanos Zafeiropoulos,  
MD, PhD <sup>2,3</sup>, Veauthyela Saint-Joy, MD <sup>2</sup>, Marwa Ilali, MD <sup>2, 4</sup>, Michel Ibrahim,  
MD <sup>2, 5</sup>, Taina Brice, MD <sup>2</sup>, and Norrisa Haynes, MD <sup>2, 6, 7</sup>

<sup>1</sup>Feinstein Institutes for Medical Research, Northwell Health

<sup>2</sup>Global MedEd Network

<sup>3</sup>Department of Cardiology, University Hospital of Zurich

<sup>4</sup>Department of Family Medicine, McGill University

<sup>5</sup>Cardiovascular Medicine, ChenMed

<sup>6</sup>Yale School of Medicine

<sup>7</sup>Yale Institute for Global Health

## Abstract

**Objective:** Virtual patient cases are a scalable and engaging tool for training medical professionals. Strategies and frameworks for their implementation in teaching and training settings are few, technically complicated and/or expensive. We developed and evaluated open source and free virtual patient cases to test knowledge acquisition during an echocardiography training program for internal medicine trainees in Haiti. The objective of this paper is to describe the technical aspects of the GMENEcho virtual patient cases implementation and motivate similar work by resource-constrained teams.

**Methods:** We used an open source engine for text-based games (Twine) since it provides the necessary interaction mechanics and is usable out-of-the-box. The case code was written in SugarCube 2.30.0 notation and the tweego-generated .html file was hosted on Github Pages for continuous integration and deployment, making iterations by the clinical team seamless. Data from completed tests were reported back via email through a third party integration.

**Results:** The technical work was completed in two weeks by a team member with a clinical background and minimal computer programming experience. The virtual patient cases were deployed for a pretest (November 2023) and a second time unaltered for a posttest (June

**NOTE:** This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.  
\*ntsafaridis@northwell.edu

29 2024) after the interim hands-on and theoretical training had been completed. Qualitative  
30 feedback was positive or neutral. The overall score in the posttest was significantly higher  
31 with a large effect size (mean absolute improvement 15.26%,  $p < 0.001$ ; Cohen's  $d$ : 1.398),  
32 similarly to the diagnostic score (mean absolute difference 16.09%,  $p < 0.001$ ; Cohen's  $d$ :  
33 1.402). Management performance missed statistical significance by a small margin. The  
34 System Usability Scale (SUS) score was 74.6 ("Excellent"). There was reduced inter-trainee  
35 variability across metrics in the posttest, including the SUS score.

36 **Discussion:** This proof-of-concept methodology can be applied to create clinical patient  
37 cases for use within a class or a clinical training setting, through a friendly graphical user in-  
38 terface. A more complex software stack can allow for remote or larger scale implementations  
39 with additional features.

40 **Conclusion:** The rapid development time and positive qualitative and quantitative feedback  
41 highlight the potential of this approach for clinical education in resource-constrained set-  
42 tings. It can serve as a template for more streamlined adaptations of case-based learning  
43 in diverse healthcare settings.

44 **Keywords:** Twine, Virtual Patients, Case-based Learning, Echocardiography, Global Health

## 45 1 Introduction

46 Case-based learning is an adjunct teaching modality that aims to connect theory and prac-  
47 tice by engaging the trainee in a conversational and active process of learning through the  
48 provision of simulated (in text or otherwise) patient cases, along with rapid and individualized  
49 feedback.<sup>1,2</sup>

50 In the context of medical education several gamified educational apps have been created that  
51 aim to provide time-efficient learning to trainees (McCoy, Lewis and Dalton, 2016).<sup>3</sup> Virtual  
52 patient scenarios seem to be an engaging method to deliver clinical vignettes relevant to a  
53 physician's duties.<sup>4</sup> However, there are limited software options that can be deployed indepen-  
54 dently for the creation and delivery of such scenarios. When open source options are available,  
55 such as OpenLabyrinth, they require significant digital infrastructure and/or programming abil-  
56 ity.<sup>5</sup>

57 As a digital learning modality requiring relevant infrastructure and expertise, virtual patient sce-  
58 narios have been less readily available in resource-constrained settings.<sup>6</sup> Lack of infrastructure  
59 and difficulties with cultural and educational adaptation are examples of the typical barriers edu-

60 cators face in the implementation of simulation-based didactics in low income countries.<sup>7</sup>

61 Additionally, most available online clinical cases are targeted at specific exams (i.e. USMLE,  
62 COMLEX, board licensing exams, etc.), with high costs to build, deploy and access. These plat-  
63 forms are not cognizant of local conditions and cultural characteristics except for the state of the  
64 art in high-resource academic settings. They are also not modifiable, extendable or replicable,  
65 limiting the potential for dissemination in other contexts.

66 Given the socioeconomic and political challenges in Haiti, there is a great need to establish re-  
67 silient cardiology training modalities and programs.<sup>8</sup> As part of the "Focused Echo InTervention  
68 Package" (FETIP) project, we created virtual patient cases that would test the echocardiogra-  
69 phy knowledge of internal medicine trainees in Haiti in a manner responsive to local needs and  
70 realities. The purpose of this paper is to describe the technical implementation of our virtual  
71 patient cases in a manner approachable to interested parties with minimal technical expertise  
72 and encourage case-based learning in all settings.

## 73 **2 Methods**

74 The GMENEcho digital patient cases were intended to bookend the FETIP program as an  
75 assessment module. The FETIP project was a 6-month-long educational intervention imple-  
76 mented in the University Hospital of La Paix, one of the four University Hospitals in Haiti. The  
77 technological infrastructure includes computers and smartphones with internet access, allowing  
78 for digital educational interventions. There are, however, significant challenges, including elec-  
79 tricity interruptions and political instability, which can affect the consistency of access to these  
80 technologies. Additionally, there is a lack of hands-on programs for technical skill development  
81 with regards to echocardiography training.

82 This virtual patient case system was developed to serve as both an assessment tool for the  
83 GMENEcho training program and a potential standalone educational resource for cardiovascular  
84 medicine in resource-constrained settings.

85 To quantitatively measure the impact of our cases, we calculated absolute improvements and  
86 standardized differences between the pretest and posttest scores. Three scores were calcu-  
87 lated as main outcomes by evaluating performance across quizzes and scenario-based actions  
88 for all four cases: overall performance, diagnostics and management scores. Interpretation of

89 echocardiographic images was presented to the trainees as quiz questions and incorporated  
90 into diagnostic scores. Details on the contribution of the case questions and choices to the  
91 scores are available in Figure 4 of the Supplement. Statistical significance testing was per-  
92 formed using paired samples t-tests, given the small number of participants (<20). Effect size  
93 assessment was conducted using Cohen's d. We used the rule of thumb thresholds set for small  
94 (0.1 to <0.30), moderate (0.3 to <0.5), and large ( $\geq 0.5$ ) effect sizes. Statistical analysis was  
95 conducted and graphs were created using Jamovi Version 2.5.

## 96 **2.1 Rationale**

97 The development of the cases was a multidisciplinary effort involving clinical and research cardi-  
98 ologists along with medical researchers from Haiti, Greece and the United States. The clinical  
99 content of the cases was created with the direction of on-the-ground physicians to address  
100 pain-points in cardiology care in Haiti.

101 The cases tied together clinical presentation, workup, echocardiographic imaging acquisition,  
102 interpretation and clinical management in a choose-your-own-adventure format, where choices  
103 affected patient outcomes and the performance was graded qualitatively on scale from insuffi-  
104 cient to optimal.

105 The content was culturally adapted, with scenarios tailored to the Haitian healthcare context and  
106 incorporation of locally available resources and treatment options. This adaptation was crucial  
107 to ensure the relevance and applicability of the cases to the target audience. The clinical details  
108 of the scenarios are further discussed in a separate manuscript (in preparation).

## 109 **2.2 Design**

110 Given the limited budget, to enhance immersion and engagement, we used stable diffusion-  
111 generated images depicting human interactions between patients and medical staff and con-  
112 veying positive or negative sentiment depending on how the trainee was doing in the case.<sup>9</sup>

113 The prompts were adapted to generate culturally appropriate images and to minimize the bias  
114 inherent in the image generating models. Examples of the prompts and resulting images are  
115 shown in Figure 1.

116 Multimedia including ECGs, heart sound audio, chest X-rays, transthoracic and transesophageal  
117 echo recordings were selected from open educational repositories to fit each specific case.<sup>10-13</sup>



Figure 1: Examples of AI-generated images used as illustrations in the GMENEcho Project. Prompts used were "A realistic digital painting of an African nurse in a hospital ward looking at the camera doing the stop sign frightened, precise, high resolution" (left) and "A realistic digital painting of a doctor in Haiti smiling at a patient who is talking to him from his bed, detailed, vibrant colors, nurses in the background, high resolution" (right). Notice doctor on the right is presented as white, in an example of bias inherent to the image generating model.

## 118 2.3 Development

119 Development began in March 2023 and concluded in September 2023. The core technical work  
120 was completed in a period of two weeks, taking about 15 hours of work and the rest of the time  
121 was dedicated to the iterative improvement of the clinical case content.

122 After developing the academic content for each case, lessons learned were taken into consid-  
123 eration to improve the implementation of the next one. The initial design of the interaction  
124 model featured multiple-choice questions allowing for a single selection. This was later evolved  
125 to allow trainees to select multiple or no treatment options, mirroring real-life clinical decision-  
126 making processes. Many of the nodes in each scenario were interconnected so that users could  
127 select between different diagnostic modalities and then returning back to the main node where  
128 they would decide the treatment based on their assessment. In the case of an erroneous choice,  
129 the script offered advice to guide the user into the correct choice, in the example of a nurse  
130 reminding the user that their selection is wrong and prompting them to reconsider it. These in-  
131 termediate nodes were only available to the user if they had selected a wrong choice and aimed  
132 to provide immediate feedback and redirect the user to the right steps in order to complete the

133 scenario.

134 Each case was structured with an interactive progression, beginning with a patient presentation  
135 and vital signs. Users were required to navigate through various diagnostic choices, including  
136 physical exam, EKG, chest X-ray, labs, and echocardiography. The cases incorporated multi-  
137 media elements such as audio clips of heart sounds, EKG images, chest X-ray images, and  
138 echocardiogram videos and still images. Based on their diagnostic findings, users were then  
139 prompted to make treatment decisions. The cases also included follow-up management and  
140 patient outcomes to provide a comprehensive learning experience.

141 A scoring system was implemented to track user performance in diagnosis, treatment, and  
142 follow-up management. Initial implementations incorporated both choice tracking and real-  
143 time scoring. However, due to complexity concerns, later iterations focused solely on choice  
144 tracking, deferring scoring to post-hoc data analysis.

145 Given that the cases were originally created in English, translation into French and Haitian  
146 Creole was undertaken by team members fluent in French and local Haitian Creole idioms,  
147 aided via automatic translation of the English text through free online translation services. This  
148 significantly reduced the time necessary to deploy the French version of the cases.

149 In order to evaluate this effort we aimed to quantify how the end users felt when using the  
150 scenarios and how well they did in their medical decision making throughout the scenarios.  
151 The System Usability Scale (SUS) questionnaire (Brooke, 1996) was deployed after all cases  
152 were completed to collect information about the usability of the virtual patients platform, with  
153 responses graded on a 5-point Likert scale from 1 = completely disagree to 5 = completely  
154 agree. This is an industry-standard tool that can provide an overall score based on the end-  
155 user responses as follows: <25 = “Worst Imaginable”, 25.1–51.6 = “Poor”, 51.7–62.6 = “OK/Fair”,  
156 62.7–72.5 = “Good”, 72.6–84 = “Excellent”, >84.1 = “Best Imaginable”.

157 The details and results of the academic scoring of the scenarios are described in more detail  
158 in the clinically-directed manuscript of the main GMENEcho study (in preparation).

## 159 **2.4 Technical Implementation**

160 Each virtual patient case was scripted using Twine’s SugarCube notation (2.30.0) and saved as  
161 an individual source file. Additional files were created to handle user identification, usability

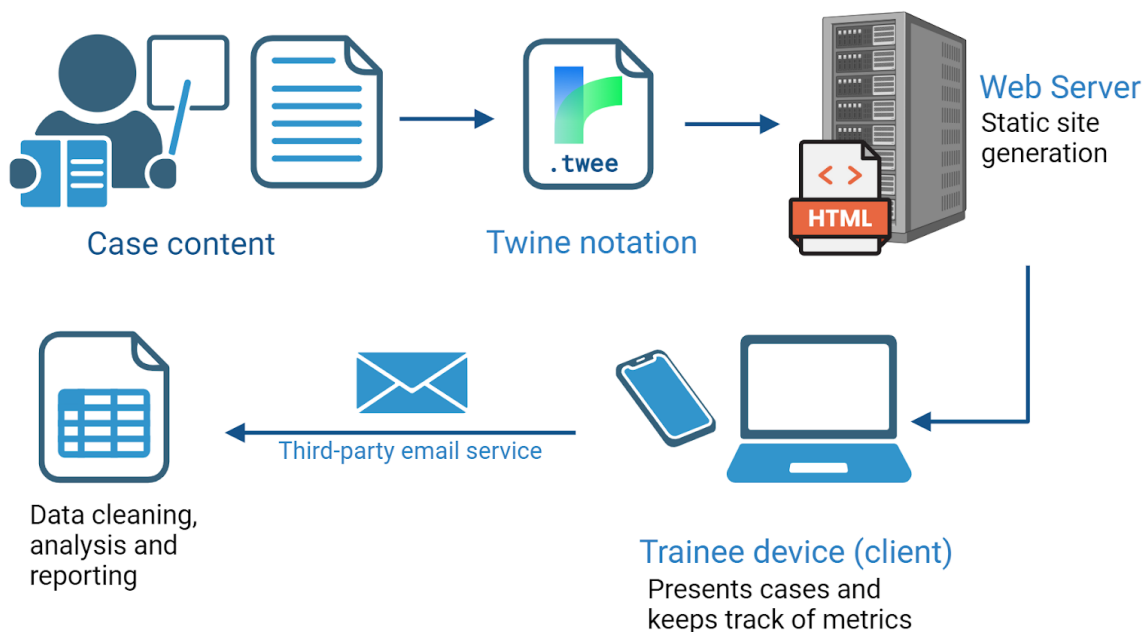


Figure 2: GMENEcho Technical Workflow: After case content creation, the scenarios were translated into Twine notation and uploaded as a static website (single .html file) on a free hosting server. After case completion, results were reported back to the research team via a third-party email service. The data were then imported, cleaned and analyzed manually. Created in BioRender. Koulas, I. (2024) BioRender.com/g96n129

162 testing, and score reporting, as well as CSS style customization and JavaScript for custom en-  
163 gine functionality. All these components were compiled into a browser-readable static website  
164 format using the command line tool Tweego, with continuous deployment during development  
165 and updates managed through GitHub Actions. The static website files generated by Tweego  
166 were then hosted on GitHub Pages under a custom domain. This allowed us to deliver a com-  
167 plete web page in the form of a single HTML file, which could be downloaded to the trainee's  
168 device in one go, just by visiting our webpage URL. This setup avoided the need for expensive  
169 server hosting and minimized server-client interaction.

170 Trainees were identified by entering a unique password of their choosing to start solving the  
171 cases. This was also used to track them between the pre- and posttest. All trainee actions  
172 in a static website are logged and managed client-side, using pre-generated standard HTML.  
173 User feedback and the results of each session were collected via an HTML form submission,  
174 through Static Forms—a service forwarding the data to the developers' email address.<sup>14</sup> This  
175 method was deemed appropriate due to the low expected volume of participants in this pilot.  
176 A graphical representation of this process can be seen in Figure 2.

177 No sensitive patient or trainee information was collected. Prototyping was completed with-  
178 out explicit cybersecurity considerations other than the security inherent in the hosting plat-  
179 form.

180 For technical details, please refer to the project's code repository.<sup>15</sup>

### 181 **3 Results**

182 The trainees were able to fully access the cases via web browsers either in their mobile devices  
183 or their personal computers. The material was delivered independently of client screen dimen-  
184 sions ensuring minimal issues with responsiveness. No data were received or sent from and to  
185 the server once the scenarios were loaded.

186 Out of a total of 16 trainees, 12 completed the pretest and 15 completed the posttest (Table  
187 1). There was a statistically significant improvement in overall scores and diagnostic scores be-  
188 tween the pretest and posttest, while management scores closely missed statistical significance  
189 (Table 2). Large effect sizes were observed for all three scores.

190 The overall score in the posttest was significantly higher than the pretest, (mean absolute score  
191 increase 15.26%,  $p < 0.001$ ) and a large effect size (Cohen's  $d: 1.398$ ). The diagnostics score re-  
192 ports showed a similarly significant and large improvement (mean absolute difference 16.09%,  
193  $p < 0.001$  and Cohen's  $d: 1.402$ ). Management scores also improved, though statistical signifi-  
194 cance was not reached (mean absolute difference 0.98,  $p : 0.012$ , Cohen's  $d: 0.805$ ). Confidence  
195 intervals for the effect size are included in the supplement.

196 In terms of time efficiency, median time per page decreased from 16.63 seconds to 8.10 sec-  
197 onds between the two assessments, with diagnostic-specific time decreasing from 11.04 to 5.7  
198 minutes. There also was reduced variability across metrics in the posttest, including the SUS  
199 score.

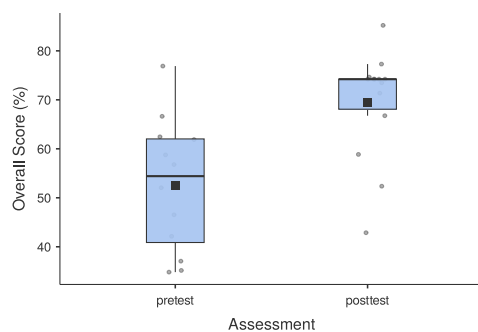
200 Qualitative feedback from participants indicated a generally positive reception of the virtual  
201 patient cases. Trainees found the cases useful for understanding key concepts, prioritizing di-  
202 agnostic procedures, and summarizing clinical scenarios. The cases were perceived as realistic,  
203 with participants noting that they reflected common clinical presentations seen in their hospital.  
204 Some suggestions for improvement included gradually increasing the difficulty of echocardi-  
205 ographic interpretations and highlighting specific diagnostic criteria for each clinical vignette.



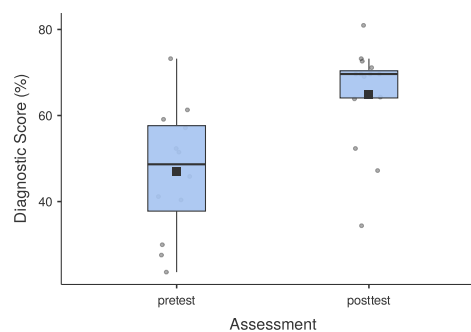
Table 1: Descriptive Statistics for the Pretest and Posttest Assessments

Measure	Time	N	Missing	Mean	Median	SD	Minimum	Maximum
Overall Score (%)	Pre	12	0	52.6	54.4	13.60	34.8	76.9
	Post	15	0	69.6	74.2	10.65	42.9	85.2
Diagnostic Score (%)	Pre	12	0	46.9	48.7	15.03	23.6	73.2
	Post	15	0	64.8	69.6	11.76	34.4	81.0
Management Score (%)	Pre	12	0	67.1	69.7	13.03	36.2	83.5
	Post	15	0	78.0	80.5	8.12	58.6	90.4
SUS Score (%)	Pre	12	0	74.6	78.8	14.61	47.5	90.0
	Post	15	0	81.0	87.5	16.47	47.5	100.0
Median Time per Node (min)	Pre	120	0	16.63	16.00	11.60	3.00	39.00
	Post	150	0	8.10	6.00	4.48	3.00	19.00
Median Time per Node, Diagnostics (min)	Pre	120	0	11.04	10.00	7.26	3.00	25.50
	Post	150	0	5.70	5.00	2.54	2.00	11.00

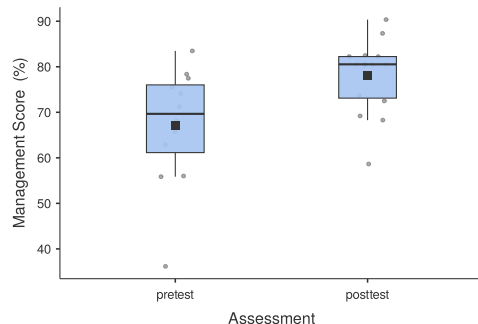
Note. SUS = System Usability Scale



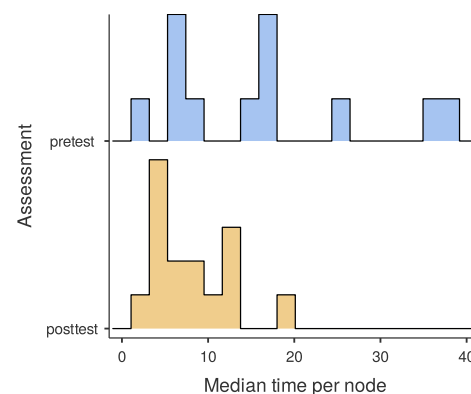
(a) Overall Score



(b) Diagnostic Score



(c) Management Score



(d) Median Time per Node

Figure 3: Average overall score, diagnostic score and management score box plots. Median time per node/step through all cases per trainee (histogram).

Table 2: Paired Samples t-test Results

			statistic	df	p	Mean diff	Cohen's d
Overall Score Post	Overall Score Pre	Student's t	4.64	10.0	<0.001	15.26	1.398
Diagnostic Report Post	Diagnostic Report Pre	Student's t	4.65	10.0	<0.001	16.09	1.402
Management Score Post	Management Score Pre	Student's t	2.67	10.0	0.012	9.84	0.805

Note.  $H_a: \mu_{\text{Measure 1}} - \mu_{\text{Measure 2}} > 0$ ; df = degrees of freedom; Mean diff = Mean difference, (%)

206 Participants also appreciated the inclusion of culturally relevant terms and patient-physician  
207 interactions, though they recommended addressing issues like treatment interruptions due to  
208 financial constraints.

## 209 4 Discussion

210 This study presents a low-cost framework for the development and deployment of virtual patient  
211 cases tailored specifically to enhance echocardiography training in resource-limited settings.  
212 Our approach leverages open-source technology to provide scalable and cost-effective educa-  
213 tional tools.

### 214 4.1 Outcomes

215 In terms of the results, three general knowledge domains were assessed: cardiovascular disease  
216 diagnosis, cardiovascular disease management, and an aggregate score combining both. The  
217 questions were primarily weighted towards diagnosis, as the curriculum emphasized the ability  
218 to diagnose and characterize different types of heart failure based on clinical presentation and  
219 echocardiography. This emphasis may have underpowered the management component of the  
220 assessment, which likely explains an improvement in management post-test scores and a large  
221 effect size that did not reach statistical significance.

222 The results, which showed statistically significant large effect sizes in two of the three main  
223 metrics (overall score and diagnostic scores), highlight the feasibility and practicality of using  
224 custom patient cases to evaluate knowledge improvement over time as part of a broader ed-  
225 ucational intervention in resource-constrained settings. The same questions were included in  
226 both the pre- and post-tests, though presented in different orders with varied multiple-choice  
227 options. Additionally, the six-month gap between pre- and post-tests likely minimized recall  
228 bias. Some participants submitted perfect scores (100%) on the post-test SUS evaluation, and  
229 there was an observed increase in both score improvement and efficiency (reduced time to

230 completion).

## 231 **4.2 Lessons Learned**

232 In terms of methodology, at the outset, we defined precise objectives aimed at evaluating the  
233 efficacy of the echocardiography training program through contextual clinical scenarios. Unlike  
234 traditional teaching and testing tools, our virtual patient cases were designed to simulate local  
235 clinical environments, fostering comprehensive clinical skill evaluation.

236 We utilized an iterative integration and deployment pipeline which enabled us to have a flex-  
237 ible and collaborative case development platform, allowing us to modify the scenarios based  
238 on test user feedback. Twine's integration with multimedia elements was crucial for reflecting  
239 complex clinical scenarios. The platform's compatibility with various foundational technologies  
240 ensured accessibility across multiple devices—a vital feature in resource-constrained environ-  
241 ments. This adaptability is reflected in our impressive System Usability Scale score of 74.6  
242 (“Excellent”), affirming the tool's intuitive and user-friendly design, thereby enhancing the edu-  
243 cational experience. With this SUS score our application was verified to be easy to navigate, so  
244 that the evaluation of the users' performance can be mainly attributed to their medical knowl-  
245 edge and clinical reasoning. This is an important, yet often neglected, aspect of modern edu-  
246 cational applications, where the technical aspects of the application itself can potentially affect  
247 academic performance.

248 Our framework's scalability is another noteworthy strength. Relying on foundational web tech-  
249 nologies like HTML and JavaScript guarantees long-term viability and adaptability across di-  
250 verse medical specialties and geographic contexts. Additionally, Twine's potential for proce-  
251 dural generation of scenarios can introduce unique and diverse learning experiences, expos-  
252 ing trainees to a wide array of clinical presentations.<sup>16</sup> The introduction of elements such as  
253 achievements or progress tracking in future iterations could further enhance learner engage-  
254 ment and motivation, potentially improving knowledge retention and classroom dynamics.

255 The framework does also have limitations, including limited server interactivity and analytics ca-  
256 pabilities, along with a reliance on third-party data sharing integrations, which may pose privacy  
257 concerns. Compared to platforms like DecisionSim and OpenLabyrinth, our approach empha-  
258 sizes cost-efficiency and customization but provides fewer advanced features. Nonetheless,  
259 its adaptability and minimal infrastructure requirements make it invaluable for institutions or

260 individuals constrained by budgets and the lack of technical capacity.

261 Our use of AI-generated images via DALL-E<sup>17</sup> to create culturally specific visuals free from  
262 copyright issues marked a significant enhancement. AI-generated images were selected as an  
263 option for their ease of adaptability, lack of copyright liability and time efficiency. The authors  
264 want to acknowledge the concerns that arise given that these models were trained on artistic  
265 works without permission or remuneration and suggest that future and/or larger projects take  
266 this into account.

267 Addressing AI bias was also critical, as initial outputs did not reflect the Haitian context accu-  
268 rately. Even though “Haiti” - where the White race represents a minority of the population - was  
269 mentioned in the prompt, doctors were shown in the generated images to be White more often  
270 than not. This highlights the need for ongoing ethical vigilance when employing AI applications,  
271 ensuring educational content is both representative and unbiased.

272 In conclusion, the GMENEcho framework addressed a critical need in cardiology training tools,  
273 particularly in resource-limited settings. By prioritizing accessibility, rapid customization, and  
274 cost-effectiveness, our model offers a flexible template for adoption in rapid prototyping and  
275 small-scale projects.

276 **Ethics and Oversight** The Ethics Committee of the University Hospital of La Paix gave ethi-  
277 cal approval for this work. Yale University’s institutional review board (ID: 2000034601) gave  
278 ethical approval for this work. Compliance with both local regulations and international ethical  
279 standards was ensured. Informed consent was obtained from all participants, who were assured  
280 of their right to withdraw from the study at any time without penalty.

281 **Data Availability Statement** All data produced in the present study are available upon reason-  
282 able request to the authors.

283 **Conflicts and Disclosures** Norrisa Haynes is the president and co-founder of GlobalMedEd  
284 Network. Veauthyelau Saint-Joy is a co-founder of the GlobalMedEd Network. The other authors  
285 have no conflicts to declare.

## 286 **References**

287 [1] Lyons, J.; Miller, M.; Milton, J. *Contemporary Nurse* **1998**, *7*, 98–102, Publisher: Routledge  
288 \_eprint: <https://doi.org/10.5172/conu.1998.7.2.98>.

- 289 [2] Thistlethwaite, J. E.; Davies, D.; Ekeocha, S.; Kidd, J. M.; MacDougall, C.; Matthews, P.;  
290 Purkis, J.; Clay, D. *Medical Teacher* **2012**, *34*, e421–e444, Publisher: Taylor & Francis  
291 \_eprint: <https://doi.org/10.3109/0142159X.2012.680939>.
- 292 [3] McCoy, L.; Lewis, J. H.; Dalton, D. *Journal of Osteopathic Medicine* **2016**, *116*, 22–34, Pub-  
293 lisher: De Gruyter.
- 294 [4] Kononowicz, A. A.; Woodham, L. A.; Edelbring, S.; Stathakarou, N.; Davies, D.; Saxena, N.;  
295 Car, L. T.; Carlstedt-Duke, J.; Car, J.; Zary, N. *Journal of Medical Internet Research* **2019**,  
296 *21*, e14676, Company: Journal of Medical Internet Research Distributor: Journal of Medi-  
297 cal Internet Research Institution: Journal of Medical Internet Research Label: Journal of  
298 Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.
- 299 [5] olab/Open-Labyrinth. 2024; <https://github.com/olab/Open-Labyrinth>, original-date:  
300 2012-02-15T18:11:53Z.
- 301 [6] Frehywot, S.; Vovides, Y.; Talib, Z.; Mikhail, N.; Ross, H.; Wohltjen, H.; Bedada, S.; Korhumel, K.;  
302 Koumare, A. K.; Scott, J. *Human Resources for Health* **2013**, *11*, 4.
- 303 [7] Seethamraju, R. R.; Stone, K. P.; Shepherd, M. *Simulation in Healthcare* **2022**, *17*, e113.
- 304 [8] Haynes, N.; Saint, J. V.; Swain, J. *Journal of the American College of Cardiology* **2021**, *77*,  
305 2749–2753, Publisher: American College of Cardiology Foundation.
- 306 [9] DALL·E 2. <https://openai.com/index/dall-e-2/>.
- 307 [10] Normal Cardiac Anatomy — TPA. [https://www.thepocusatlas.com/](https://www.thepocusatlas.com/normal-cardiac-anatomy)  
308 [normal-cardiac-anatomy](https://www.thepocusatlas.com/normal-cardiac-anatomy).
- 309 [11] Univeristy of Michigan Medical School | Professional Skill Builder. [https://med.umich.](https://med.umich.edu/lrc/psb_open/html/menu/index.html)  
310 [edu/lrc/psb\\_open/html/menu/index.html](https://med.umich.edu/lrc/psb_open/html/menu/index.html).
- 311 [12] Everyday Ultrasound. 2020; [https://everydayultrasound.com/blog/category/Acute+](https://everydayultrasound.com/blog/category/Acute+Heart+Failure)  
312 [Heart+Failure](https://everydayultrasound.com/blog/category/Acute+Heart+Failure).
- 313 [13] Wikimedia Commons. [https://commons.wikimedia.org/wiki/Main\\_Page](https://commons.wikimedia.org/wiki/Main_Page).
- 314 [14] HTML Forms for you static websites. <https://www.staticforms.xyz/>.

315 [15] tsaftaridis Tsaftaridis/VirtuaEcho. 2024; <https://github.com/Tsaftaridis/VirtuaEcho>,  
316 [echo](https://github.com/Tsaftaridis/VirtuaEcho), original-date: 2023-04-09T06:18:04Z.

317 [16] Quail, N. P. A.; Boyle, J. G. *BMC Medical Education* **2023**, 23, 417.

318 [17] Miftahul Amri, M.; Khairatun Hisan, U. *Journal of Novel Engineering Science and Technol-*  
319 *ogy* **2023**, 2, 34–39.

## 320 5 Supplemental Information

321 The Shapiro-Wilk Test Results for Normality validates the appropriateness of the paramet-  
322 ric tests utilized by confirming a normal distribution of the differences between the pre- and  
323 posttests. Completion rates were high, with 12/16 completing the pretest and 14/16 completing  
324 the posttest.

Table 3: Shapiro-Wilk Test Results for Normality

		W	p
Overall Score Post	- Overall Score Pre	0.954	0.691
Diagnostic Report Post	- Diagnostic Report Pre	0.959	0.761
Management Score Post	- Management Score Pre	0.917	0.296

Note. A low p-value suggests a violation of the assumption of normality

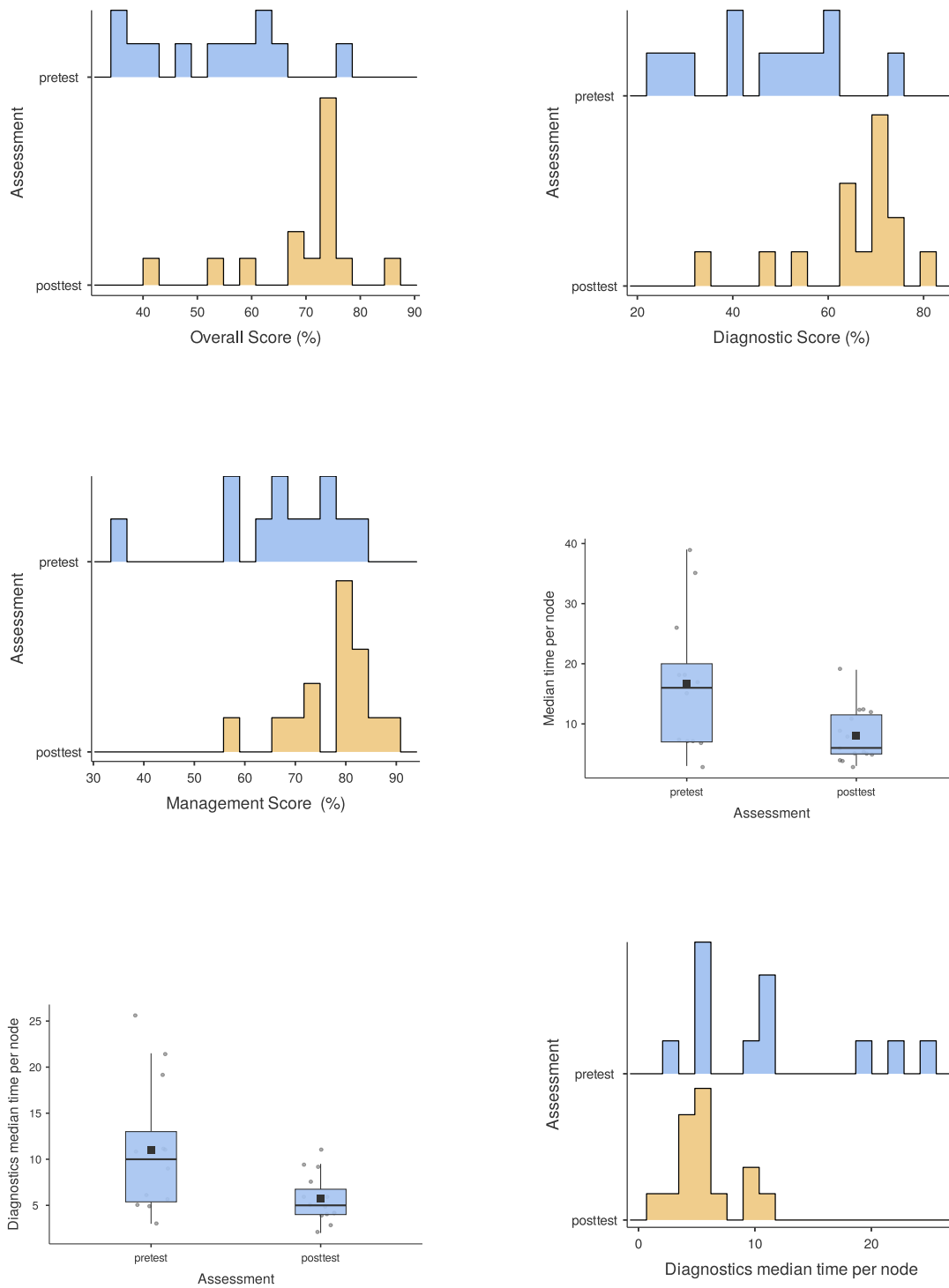


Figure 4: Complimentary histograms and box plots for Figure 3

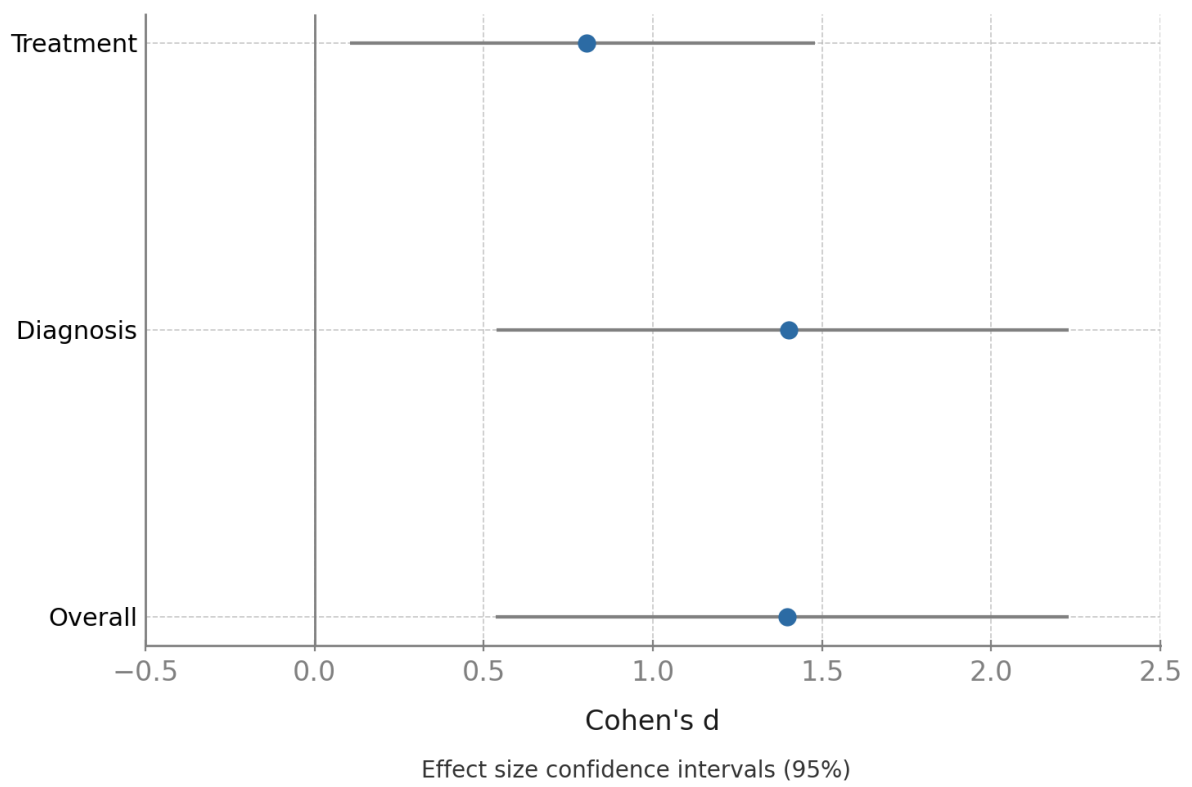
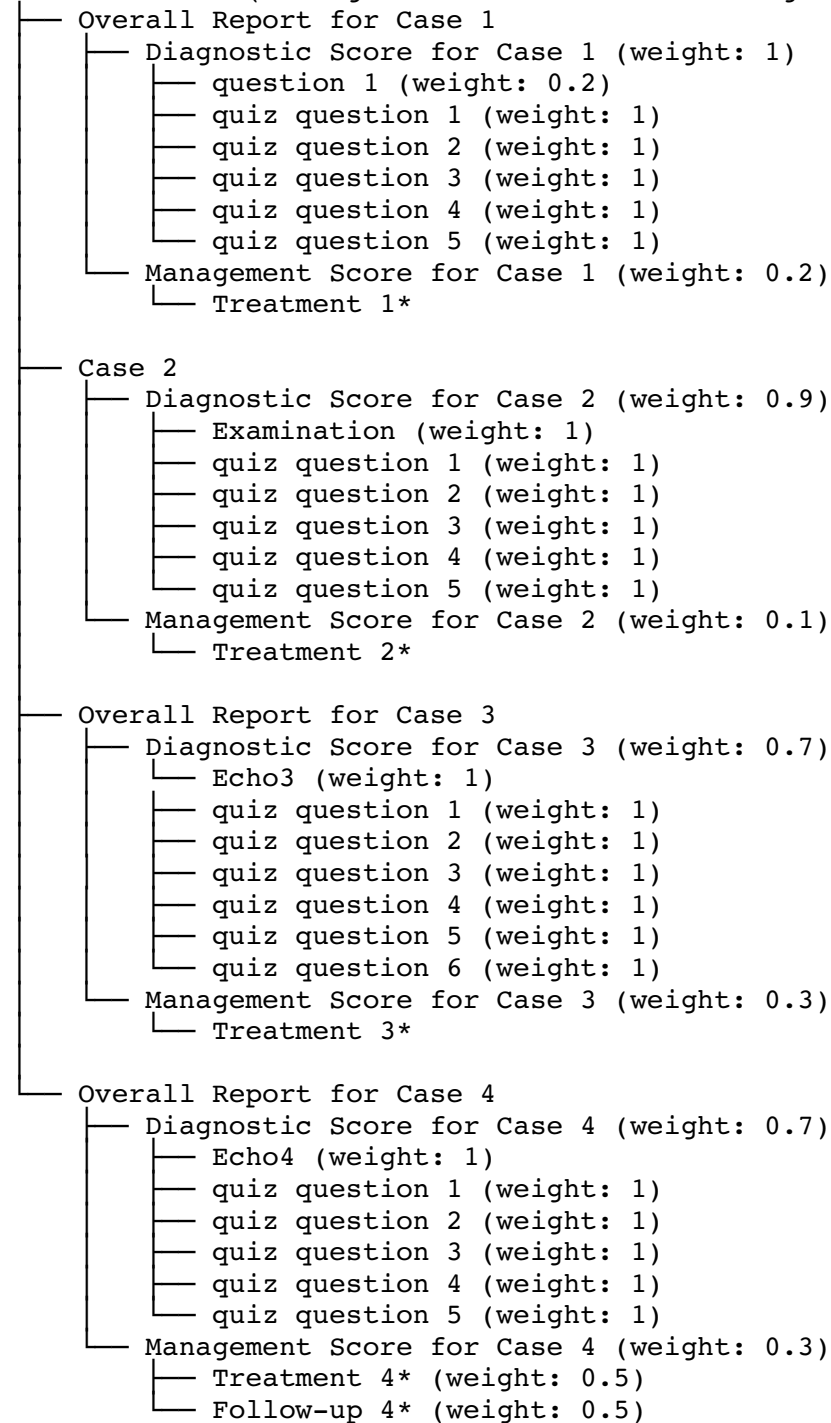


Figure 5: 95% confidence intervals for the three comparisons.



Overall Score (Average of Overall Scores through cases 1 to 4)



\* Variable integrated in narrative and auto-graded, then normalized between 0 and 1

Final Calculations:

- \*\*Overall Score (%)\*\*: Average of Overall Reports 1 to 4 multiplied by 100

Figure 6: Schematic representation of weights and contributions of assessment data to the reported scores.