Daniel J. Morgan*, Laura Scherer, Lisa Pineles, Jon Baghdadi, Larry Magder, Kerri Thom, Christina Koch, Nick Wilkins, Mike LeGrand, Deborah Stevens, Renee Walker, Beth Shirrell, Anthony D. Harris and Deborah Korenstein

Game-based learning to improve diagnostic accuracy: a pilot randomized-controlled trial

https://doi.org/10.1515/dx-2023-0133
Received October 4, 2023; accepted January 9, 2024;
published online January 30, 2024

Abstract

Objectives: Perform a pilot study of online game-based learning (GBL) using natural frequencies and feedback to teach diagnostic reasoning.
Methods: We conducted a multicenter randomized-controlled trial of computer-based training. We enrolled medical students, residents, practicing physicians and nurse practitioners. The intervention was a 45 min online GBL training vs. control education with a primary outcome of score on a scale of diagnostic accuracy (composed of 10 realistic case vignettes, requesting estimates of probability of disease after a test result, 0–100 points total).
Results: Of 90 participants there were 30 students, 30 residents and 30 practicing clinicians. Of these 62 % (56/90) were female and 52 % (47/90) were white. Sixty were randomized to GBL intervention and 30 to control. The primary outcome of diagnostic accuracy immediately after training was better in GBL (mean accuracy score 59.4) vs. control (37.6), p=0.0005. The GBL group was then split evenly (30, 30) into no further intervention or weekly emails with case studies. Both GBL groups performed better than control at one-month and some continued effect at three-month follow up. Scores at one-month GBL (59.2) GBL plus emails (54.2) vs. control (33.9), p=0.024; three-months GBL (56.2), GBL plus emails (42.9) vs. control (35.1), p=0.076. Most participants would recommend GBL to colleagues (73 %), believed it was enjoyable (92 %) and believed it improves test interpretation (95 %).
Conclusions: In this pilot study, a single session with GBL nearly doubled score on a scale of diagnostic accuracy in medical trainees and practicing clinicians. The impact of GBL persisted after three months.

Keywords: medical diagnosis; medical education; online game-based learning

Introduction

Diagnostic error will impact most people during their lifetime [1]. Inappropriate understanding of probability of disease given a positive or negative test result can lead to diagnostic errors and thus is a critical barrier to progress [2–4]. Diagnostic medical-decision making, such as this, is a form of judgement under uncertainty, which is prone to biases if not objective [5–7]. Diagnostic decision making and test interpretation is predominantly taught as a mathematical calculation with formulas or 2×2 tables [8, 9]. Methods for teaching diagnostic probability that have shown promise include using decision analysis tree natural frequencies with or without graphics to make probability more intuitive [10–12]. Better diagnostic education has been called for but tools using these advanced approaches are lacking [8, 9].

Educational games use repetitive, rapid decision-making with immediate feedback to train skills [13]. They have been widely used to improve skills in chess and gambling, and medicine, where applications included simulations in emergency settings [14]. These games are more efficient than problem-based learning and may be superior

***Corresponding author: J. Daniel Morgan**, MD, MS, Department of Epidemiology and Public Health, University of Maryland School of Medicine, 685 West Baltimore Street, MSTF 334, Baltimore 21201, MD, USA; and VA Maryland Healthcare System, Baltimore, MD, USA, Phone: 410-706-1734, Fax: 410-706-0098, E-mail: dmorgan@som.umaryland.edu
Laura Scherer, Adult and Child Consortium of Health Outcomes Research and Delivery Science (ACCORDS), University of Colorado School of Medicine, Aurora, CO, USA; Division of Cardiology, University of Colorado School of Medicine, Aurora, CO, USA; and Center of Innovation for Veteran-Centered and Value-Driven Care, VA Denver, Denver, CO, USA
Lisa Pineles, Jon Baghdadi, Larry Magder, Kerri Thom, Deborah Stevens and Anthony D. Harris, Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD, USA. https://orcid.org/0000-0002-2377-5971 (L. Pineles)
Christina Koch, Division of General Internal Medicine, University of Maryland School of Medicine, Baltimore, MD, USA
Nick Wilkins, Code in the Schools, Baltimore, MD, USA
Mike LeGrand, Firaxis Games, Baltimore, MD, USA
Renee Walker and Beth Shirrell, Visual Communication Design, Thomas Jefferson University, Philadelphia, PA, USA
Deborah Korenstein, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

for training intuitive associations [15]. Games have targeted heuristics to change thinking processes inherent in clinical medicine, suggesting broad future application [14, 16].

We report a pilot study evaluating an online game designed to improve diagnostic testing skills in medical students, residents and practicing clinicians. The intervention consisted of watching a short video and playing a game teaching estimates of probability of disease before and after testing. The game sought to train clinician intuitive gestalt without formal calculations. A control intervention consisted of traditional education materials. We evaluated the effect immediately after training, at one- and three-month follow-up.

Methods

Participants

We enrolled medical students, internal medicine residents and practicing clinicians at two hospitals. Potential participants were contacted through a group email to medical students, another email to internal medicine residents and emails and direct recruitment of practicing clinicians (physicians and nurse practitioners (NP)).

Participants were randomized to intervention (GBL; 30 participants), intervention with cases to remind participants of game-based learning (30 participants) or control (30 participants). In total, participants were assigned to three groups in a 1:1:1 fashion. These groups were GBL intervention, GBL intervention plus emails, and control (see Figure 1, study overview). Randomization was stratified by type of participant (student, resident, clinician). After enrollment, participants were given a link to a Qualtrics survey that contained links to individual interventions and follow up questionnaires. We provided gift cards at three points of engagement: a $100 gift card after completing the initial assessment and $50 gift cards after the one- and three-month post-intervention assessments.

Intervention

The game was developed iteratively with clinician feedback and is publicly available (https://bird.testingwisely.com/). It requires users to estimate probability of diagnosis based on 10 simplified case presentations. Playing one game takes 5 min or less. Immediate feedback on each question is delivered using natural frequency methods with a score, tips, and comparisons to other users. The game has a short tutorial and different play options.

The entire GBL intervention instructed participants to (1) visit the testingwisely.com website and review materials, (2) watch a 4 min video explaining natural frequencies to determine post-test probability (Episode 4, https://www.testingwisely.com/educational-videos); (3) play the game tutorial and two games. The initial GBL intervention was expected to take 45 min or less. Participants attested to completing each step of the intervention.

After assessing the primary outcome, half of the GBL intervention group was randomized to receive weekly emails containing a clinical case emphasizing points from GBL.

Control exposure

The control condition required a similar amount of time, and standard materials taught diagnosis using traditional calculations and 2×2 table methods. Materials were provided in the control Qualtrics survey and included reading the UpToDate online medical textbook testing chapter [17] and two highly-viewed YouTube videos on diagnostic testing (https://youtu.be/Z5TtopYX1Gc and https://youtu.be/dHj7ygeqelw). Participants attested to completing each step of the control.

Measures

Primary outcome: A diagnostic accuracy score based on participant responses to 10 case vignettes. Each vignette asked participants to estimate the post-test probability of disease based on results of real tests and diagnoses (see Supplemental Material, Appendix 1). Responses were multiple choice; 10 points were awarded if correct and four points for the closest to the correct answer among the incorrect options. The total possible score was 100 points.

Secondary outcomes: Comparisons of diagnostic accuracy score at one-month and three-month follow-up.

In addition to a diagnostic accuracy score, participants provided demographic information and open-ended feedback.

Statistical methods, analyses and expected outcomes

The primary outcome compared the continuous accuracy score in GBL intervention vs. control subjects. The distribution of this score was

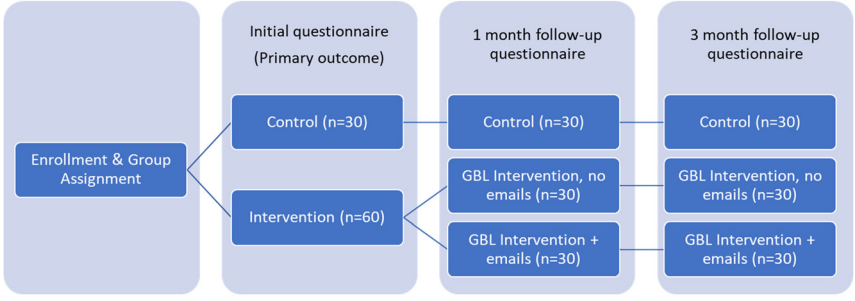

Figure 1: Overview of participant assignment and frequency of assessments during the study.

compared between the GBL intervention or control at the initial assessment, and all three groups at the one-month and three-month assessments. Point estimates and p-values were based on a longitudinal regression model fit by maximum likelihood.

Sample size considerations

This was a pilot study with the goal of making estimates around effect size, optimal intervention and feasibility of enrolling different groups.

This study was deemed exempt by the Institutional Review Board (IRB) of note. Because this was a pilot study, it was not registered in clinicaltrials.gov consistent with guidance for pilot studies [18].

Results

Demographics

Among the 90 participants, 30 were medical students, 30 residents and 30 practicing clinicians. Demographics are provided in Table 1.

Follow-up

All participants completed the initial evaluation. At one month, 71/90 completed follow-up (79 %) and at three months, 66/90 completed follow-up (73 %). Notably, follow-up was worse in those receiving intervention plus emails than intervention alone (19/30 vs. 24/30), likely as a participant noted, because of routinely ignoring study emails in the group that received additional emails.

Table 1: Demographic information on the 90 participants participating in the study.

	Control (30)	GBL intervention (60)
Female gender	17 (56.7 %)	39 (65.0 %)
Race		
Asian	8 (26.7 %)	27 (45 %)
Black	3 (10 %)	4 (6.7 %)
White	17 (56.7 %)	30 (50.0 %)
Other race	3 (10 %)	3 (5.0 %)
Hispanic or Latino	3 (10 %)	3 (5.0 %)
Practitioner type		
Medical student	10 (33.3 %)	20 (33.3 %)
Resident	10 (33.3 %)	20 (33.3 %)
Practicing clinician	10 (33.3 %)	20 (33.3 %)
MD/DO	10 (33.3 %)	17 (28.3 %)
NP	0	3 (5.0 %)
Report current use of online resources	28 (93.3 %)	54 (90.0 %)

NP, nurse practitioners.

Primary outcome

The game-based learning intervention improved mean diagnostic accuracy; mean scores were 59.4 (of 100) in those who underwent game-based learning vs. 37.6 in controls. The estimated initial difference in score based on the longitudinal regression model was found to be 21.8 points (95 % confidence interval 9.8 to 33.9 points, p=0.0005). This is based on a longitudinal regression model accounting for the correlation between repeated measures on the same participant (see Figure 2).

Secondary outcomes

At one-month and three-month follow up, there were three groups: GBL intervention, GBL intervention plus emails and control. At one-month follow up, GBL (mean score 59.2) and GBL plus emails (mean score 54.2) performed better than control (33.9), p=0.024. There was no difference between GBL with or without emails (p=0.99). At three-month follow up, game-based learning (mean score 56.2) and GBL plus emails (mean score 42.9) were not significantly different than control (35.1), p=0.077. There was no significant difference between GBL with and GBL without emails (p=0.091) (see Figure 2).

Medical students statistically performed no differently than residents or clinicians in practice, although the effect of GBL appeared potentially less in practicing clinicians

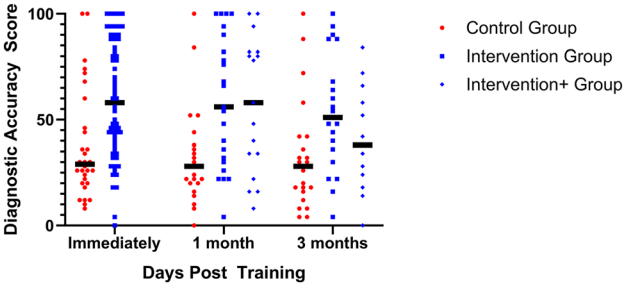

Figure 2: Impact of game-based learning with or without follow-up emails (intervention) vs. a standard control education immediately after game-based learning (primary outcome) and at one- and three-month follow-up. Horizontal bars represent group means.

than other groups. Initial performance after GBL intervention (mean 68.0 residents, vs. 62.1 students, vs. 48.2 practicing clinicians) (p=0.067).

Experience of game-based learning

Participants were asked if they would recommend GBL or use GBL again (possible answers yes/maybe/no) and their degree of support for statements about GBL (possible answers not at all/slightly/moderately/very/extremely). Participants were generally positive about GBL being enjoyable and improving test interpretation and most would recommend or use again (see Figure 3).

Qualitative responses to game-based learning

Intervention group participants were asked to give optional feedback on GBL. All participants who completed the intervention provided feedback (60/60). Themes and examples are described in Table 2.

Discussion

In this pilot study, an online GBL intervention nearly doubled diagnostic accuracy in medical trainees and practicing clinicians and persisted for three months. Follow up emails with cases did not improve accuracy. Participants found the training generally easy, enjoyable, and reported that GBL improves diagnostic test interpretation.

The primary goal of this pilot study was to determine the effect size and optimal intervention with GBL. We found an unexpectedly large effect size on a case-based score of diagnostic accuracy. GBL led to a near doubling of diagnostic accuracy. The effect of GBL persisted at one and three months without additional training. This large effect size means that in future RCTs, sample sizes required to study this accuracy outcome will be relatively small raising the possibility of more clinically focused outcomes, such as clinician practice patterns. Follow up emails did not improve performance, likely, as noted by one participant because of ignoring emails. The large effect size also suggests that this online, experiential approach to teaching diagnosis by estimation with natural frequencies [10–12, 19] outperforms traditional calculations and should be considered for medical education.

Most participants reported GBL was enjoyable, improved test interpretation and that they would recommend it to a colleague. Estimating the probability of a diagnosis after testing uses Bayesian updating, a skill that is developed with repetitive practice. Achieving adequate practice on a game, requires that the game be acceptable and moderately entertaining [20, 21]. Some trainees expressed opinions that GBL would have been helpful for USMLE board review. The goal the game, to require gestalt estimates in a timeframe similar to that encountered in patient care was experienced as stressful for some participants.

Limitations

Limitations to this pilot study include that the sample size was moderate and within a single medical system and may not be generalizable. The outcome scale of diagnostic accuracy has not been validated. Although GBL greatly improved mean accuracy score, there was variability in clinician scores in all groups. We do not know how lack of follow-up

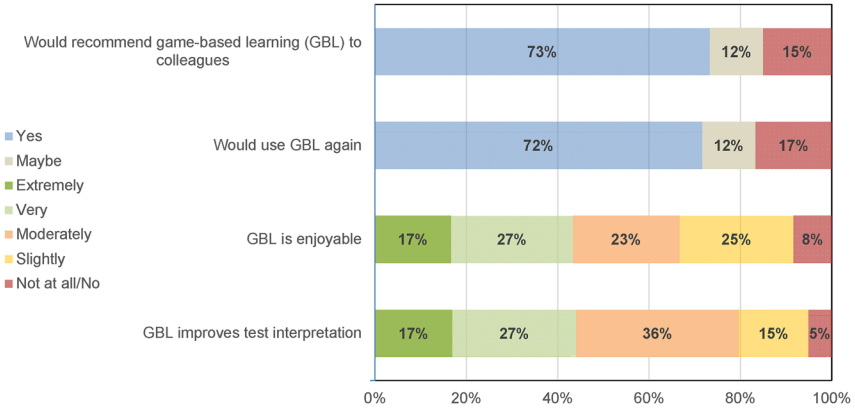

Figure 3: Participants perceptions of game-based learning (GBL). Recommending or using GBL could be answered yes, maybe or no. Opinion of enjoyability and improving test interpretation could be extremely, very, moderately, slightly, not at all/no.

Table 2: Responses of participants randomized to GBL interventions to optional, open-ended questions, organized by most common themes.

Question (# responses)	Theme	Example
Describe the best part of GBL (58/60 GBL intervention participants responding)	GBL improves number sense for probabilities	"The game really helped me develop a quick method to evaluate the meaning of a test given a pre-test probability, sensitivity, specificity, and test result"
	GBL is easy and intuitive to use	"Very intuitive interface that was easy to navigate. Feedback was helpful to gauge patterns with right/wrong answers"
	GBL would help with board exams	"I wish it was something I looked at when studying for step 3"
What was the worst part of or what you would change about GBL? (53/60)	The game was hard	"It was frustrating when I got the wrong answers, but I started to pick up on patterns as I continued to play,"
	A timer was challenging	"Trying to calculate everything as fast as possible – 30 s is very short when there is a timer in your face!"
	Moving from formulas to gestalt estimates was different	"I am used to calculating PPV, NPV, so it felt weird learning to intuit it"
Has GBL changed the way you think of diagnosis? (36/60 positive)	GBL emphasized the role of medical decision making for diagnosis	"It made me realize how important assessing the patient thoroughly and using the entire presentation of the patient to take into account the potential diagnosis vs. a particular aspect greatly impacts how to interpret tests,"
	Avoid unnecessary testing	Instead of just ordering something directly just because I have a suspicion of something, it's also important to consider the probabilities behind this and avoid unnecessary testing"
	Understand patterns around use of tests	"This helped me pick up patterns quicker to understand what it means when a test is a certain % sensitive or specific. This will help me think quicker and on the fly in the future"
Do you believe you will practice any differently when performing diagnostic testing based on this training? (38/60 positive)	Will consider pretest probability when making a diagnosis	"Will be factoring in pre-test probabilities more often and whether a positive test would change my management"
	Will order tests more carefully	"It is a reminder to think more critically about the tests I order and how they will actually help develop a diagnosis"
	Will change how tests are interpreted	"I will use testing as a tool to aid in diagnoses, as opposed to jumping straight to conclusions based on results"

GBL, game based learning.

may have impacted results at one and three months. Finally, participants were reimbursed for completion of the study and attention may not be replicated with standard voluntary participation.

Conclusions

Game-based learning is a promising intervention to improve diagnosis that was well received and is easily disseminated to medical training programs.

Research ethics: The local Institutional Review Board deemed the study exempt from review.
Informed consent: Not applicable.
Author contributions: The authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: The authors state no conflict of interest.

Research funding: National Institutes of Health (http://dx.doi.org/10.13039/100000002) 1DP2LM012890-01 (DJM).
Data availability: The raw quantitative data can be obtained on request from the corresponding author.

References

1. Balogh E, Miller BT, Ball J, Institute of Medicine (U.S.), editors. Improving diagnosis in health care. Washington, DC: The National Academies Press; 2015:444 p.
2. Morgan DJ, Pineles L, Owczarzak J, Magder L, Scherer L, Brown JP, et al. Accuracy of practitioner estimates of probability of diagnosis before and after testing. JAMA Intern Med 2021;181:747–55.
3. Casscells W, Schoenberger A, Graboys TB. Interpretation by physicians of clinical laboratory results. N Engl J Med 1978;299:999–1001.
4. Morgan DJ, Meyer AND, Korenstein D. Improved diagnostic accuracy through probability-based diagnosis [Internet]. Report No.: AHRQ Publication No. 22-0026-3-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2022. https://www.ahrq.gov/patient-safety/reports/issue-briefs/probabilistic-thinking.html. [Accessed 9 Nov 2022].

5. Korenstein D, Scherer LD, Foy A, Pineles L, Lydecker AD, Owczarzak J, et al. Clinician attitudes and beliefs associated with more aggressive diagnostic testing. Am J Med 2022;135:e182–93.
6. Kahneman D, editor. Judgment under uncertainty: heuristics and biases. Cambridge: Cambridge Univ. Press; 2008. 555 p.
7. Sox HC, Higgins MC, Owens DK. Medical decision making, 2nd ed. Chichester, West Sussex: Wiley-Blackwell; 2013:347 p.
8. Graber ML, Holmboe E, Stanley J, Danielson J, Schoenbaum S, Olson APJ. A call to action: next steps to advance diagnosis education in the health professions. Diagnosis 2021;9:166–75.
9. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med 1980;302:1109–17.
10. Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. Acad Med 1998;73:538–40.
11. Hoffrage U, Krauss S, Martignon L, Gigerenzer G. Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. Front Psychol 2015;6:1473.
12. Garcia-Retamero R, Hoffrage U. Visual representation of statistical information improves diagnostic inferences in doctors and their patients. Soc Sci Med 2013;83:27–33.
13. Girard C, Ecalle J, Magnan A. Serious games as new educational tools: how effective are they? A meta-analysis of recent studies. J Comput Assist Learn 2013;29:207–19.
14. Mohan D, Farris C, Fischhoff B, Rosengart MR, Angus DC, Yealy DM, et al. Efficacy of educational video game versus traditional educational apps at improving physician decision making in trauma triage: randomized controlled trial. BMJ 2017;359:j5416.
15. Graafland M, Dankbaar M, Mert A, Lagro J, De Wit-Zuurendonk L, Schuit S, et al. How to systematically assess serious games applied to health care. JMIR Serious Games 2014;2:e11.
16. Mohan D, Schell J, Angus DC. Not thinking clearly? Play a game seriously! JAMA 2016;316:1867–8.
17. Mahutte NG, Duleba AJ. Evaluating diagnostic tests; 2022. Available from: https://www.uptodate.com/contents/evaluating-diagnostic-tests?search=Evaluating%20diagnostic%20tests&source=search_result&selectedTitle=1~150&usage_type=default&display_rank=1.
18. Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, et al. A tutorial on pilot studies: the what, why and how. BMC Med Res Methodol 2010;10:1.
19. Brush JE, Lee M, Sherbino J, Taylor-Fishwick JC, Norman G. Effect of teaching Bayesian methods using learning by concept vs. learning by example on medical students' ability to estimate probability of a diagnosis: a randomized clinical trial. JAMA Netw Open 2019;2:e1918023.
20. Poindexter O. 8 great chess apps for beginners and grand masters. Wired [Internet]; 2022. Available from: https://www.wired.com/story/best-chess-apps/.
21. How A.I. conquered poker – The New York Times [Internet]; 2022. Available from: https://www.nytimes.com/2022/01/18/magazine/ai-technology-poker.html.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/dx-2023-0133).