AD_________________

Award Number: W81XWH-16-1-0797

TITLE: An Adaptive Tutor for Improving Visual Diagnosis

PRINCIPAL INVESTIGATOR: Martin V. Pusic MD PhD

CONTRACTING ORGANIZATION: New York University
New York, NY 10016

REPORT DATE: October 2018

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE
October 2018

2. REPORT TYPE
Annual

3. DATES COVERED
30 Sep 2017 - 29 Sep 2018

4. TITLE AND SUBTITLE
An Adaptive Tutor for Improving Visual Diagnosis

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-16-1-0797

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Martin V. Pusic – NYUSoM; David Cook – Mayo Clinic; Rose Hatala – UBC; Matt Lineberry – KUMC.

E-Mail: Martin.Pusic@nyumc.org; Cook.David33@mayo.edu; rhatala@mac.com; mlineberry@kumc.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
NYU School of Medicine

550 First Ave.,
New York, NY 10016

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT Online cognitive trainers for visual diagnosis can transcend institutional barriers to enable broad distribution of learning material. However, most current examples are based on declarative knowledge instructional designs that deliver outcomes that are only indirectly connected to patient care. Our key contention is that cognitive learning platforms, using evidence-based instructional designs, can facilitate efficient and effective visual diagnosis skill development and maintenance. **Progress to date:** We have had success in three key components of the eventual adaptive tutor: 1) we have assembled a corpus of 80,000 ECGs with their associated clinical information and have organized that corpus into 2) a prototype presentation database that allows any stakeholder to search and download ECGs according to any of the 120 American Heart Association diagnostic labels. 3) We have completed three pilot studies designed to inform the design of the adaptive tutor including a) focus groups to develop a relative "importance" ranking, b) pairwise comparisons by cardiologists to determine the feasibility of complexity ranking of ECGs and c) exploring the degree to which two overlapping ECG phenotypes can be confused and how this can be statistically modeled. We are now well positioned to use these materials and methods to carry out the next phase, a large prospective data collection and subsequent impact trial of adaptive learning.

15. SUBJECT TERMS
Electrocardiogram (ECG), ECG standards, medical education, visual pattern recognition, educational models, computer-assisted instructions, clinical competence, continuing medical education, psychometrics, e-learning, simulation, and instructional design.

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
Unclassified	Unclassified	Unclassified	Unclassified	26	USAMRMC
					19b. TELEPHONE NUMBER *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

10. INTRODUCTION:

"An Adaptive Tutor for Visual Diagnosis". In combination with co-investigators from the Mayo Clinic, the University of B.C. and University of Illinois (Chicago), the NYU team seeks to optimize the learning of visual diagnosis by developing a demonstration project using 20,000 consecutive ECG downloaded from the NYULMC Electronic Medical Record. The ECGs will be categorized and calibrated using advanced informatics techniques including natural language processing in order to develop a generalizable method of presenting visual diagnosis cases in the exact quantity and order that ensures efficient and effective learning. Target learners in the study are a wide range of healthcare professionals including nurses, paramedics, emergency physicians, internists and cardiologists.

KEYWORDS:

Cardiology, Electrocardiogram, Adaptive Learning, Instructional Design, Readiness Training

ACCOMPLISHMENTS:

What were the major goals of the project?

Specific Aims timeline includes:
SA 1: To assemble a massive online ECG library from authentic field cases collected from a representative clinical setting: the Emergency Department
SA 1: Task 1) USAMRMC ORP approval Obtained – (Completed - December 19, 2016)
SA 1: Task 2) Preliminary labeling of full set of ECGs – (Completed)
SA 1: Task 3) Download ECG cases from MUSE and EPIC systems – (Completed)
SA 1: Task 4) Focus Group for assigning cognitive task scores, difficulty ranking of diagnostic labels – (Completed) SA 1: Task 5) Programing of presentation software – (Completed)
SA 1: Task 6) IRB approvals at all 4 site (NYU, KUMC, UBC, and Mayo Clinic) – (Completed)
SA 2: To develop both ontologic and statistical models of the ECG cases so as to inform the rational design of the adaptive learning system
SA 2: Task 1) Natural language processing across the ECG set to identify full diagnostic labeling – (Completed); Linking to clinical data (Completed) Rare case collection (Ongoing)
SA 2: Task 2) Pilot Data Collection – recruit 100 clinicians – (Recruited 40/60; Completed 40/40)
SA 2: Task 3) Pairwise Comparisons – recruit 20 cardiology fellows/residents – (Completed 20/20)
SA2: Task 4) Prospective Data collection – recruit 500 participants; (Completed
SA 3: To develop an evidence-based "learning adaptation algorithm" that can ensure efficient and reliable development of skill at scale (Development completed; Recruitment begun)

What was accomplished under these goals?

Specific Aims timeline for Yr. 2 Q4 (NCE) includes:
SA 1: To assemble a massive online ECG library from authentic field cases collected from a representative clinical setting: the Emergency Department – (Ongoing aim)
SA 1: Task 1) Submit documents for USAMRMC ORP approval and IRB applications – (Completed. USAMRMC ORP approved NYU site with all the necessary documentation. Approvals for studies taking place to end of grant are in place. Applications for IRB were submitted for all the participating sites for studies. NYU IRB was fully approved on October 14, 2016; University of British Columbia IRB was fully approved on March 9, 2017; KUMC

site fully approved April 7th; and Mayo Clinic IRB was fully approval on May 1, 2017. IRB continuations are in place at all sites.
SA 1: Milestone 1) USAMRMC ORP approval Obtained – (After review of the necessary documents USAMRMC ORP approval was obtained on December 19, 2016 for Part I of the protocol: Part II approved on May 16, 2017)
SA 1: Task 2) Download ECG cases from MUSE and EPIC Electronic Health Records, initial labeling using reporting functions – After a few technical issues 95,000 candidate ECG cases are downloaded; the EPIC collateral data has also been downloaded. We have accurately merged the two datasets resulting in a bank of 80,000. Functionality that we have added this quarter includes:

- Link of ECG to past/"future" ECGs at the patient level, within the dataset time period
- Visit level Troponin and Potassium levels available for clinical correlation
- Text of any cardiologic procedures associated with that patient, including timestamps
- "Track change" representation of cardiologist over-read of machine interpretation
- Rhythm strip animation for every lead of every ECG
- Manuscript in preparation (MS#1). "Population-based Item Sampling for Online Learning"
    - Submission envisioned before grant end

SA 1: Task 3) Focus groups for assigning cognitive task scores, difficulty ranking of 120 American Heart Association (AHA) diagnostic labels. We divided this into two sessions, one with cardiologist and other with emergency medicine and internal medicine clinicians. First focus group with cardiologist was held on March 23, 2016. While, second was held April 6, 2016. Transcription was completed end of May. Importance ranking has been captured and entered for each case in the bank. Qualitative coding of transcription is complete.

- Manuscript in preparation (MS#2). "Focus Group Rankings of ECG Importance".
    - Submission envisioned before grant end.

SA 2: To develop both ontologic and statistical models of the ECG cases so as to inform the rational design of the adaptive learning system
SA 2: Task 1) Natural language processing across the ECG set to identify full diagnostic labeling – all 80,000 cases have been parsed by AHA diagnostic category. Set is fully categorized.
SA 2: Task 2) Pilot Data Collection – recruit 100 clinicians – recruitment largely complete

- Confusable diagnoses –recruited 36 clinicians (16 beyond the planned 20) completed the study maneuver. Results show the anticipated pattern of experts being able to distinguish 5 thresholds of certainty for the Pericarditis/STEMI distinction; clear distinction between experts and non-experts; uncovers ability and bias variations.
    - Obtained IRB approval for additional chart review for the 20 ECG cases to further delineate/assure clinical outcome associated with the given ECG
    - Analyzed data leading to an accurate psychometric model for this clinical decision.
    - Model informs new Prospective Cohort study – details below.
- Manuscript in preparation (MS#3). "Graded Response Models for Assessing Diagnostic Discernment: An Example Using Electrocardiogram Interpretation."
    - Submission envisioned before grant end.
- Reaction Time – plan is to recruit 60 subjects; this study had been on hold pending the pairwise comparison data analysis. We are in data collection, having 40/60 recruited. Recruit paused to coordinate with LAA recruitment.

SA 2: Task 3) Pairwise Comparisons – recruit 20 cardiology fellows/residents – recruitment is completed (20/20 have completed; data analysis ongoing).

- Manuscript in preparation (MS#4). "The Difficulty with ECG Difficulty: Statistical Models of ECG Interpretation" Manuscript in progress

SA2: Task 4) Case characteristics model initial calibration is completed with Importance rankings and complexity rankings. Completed.
SA2: Task 5) Programming Study Database – this has been completed with full ECG and EMR data on 80,000 ECGs and a search interface.
SA2: Task 6) Main Cohort Study

- Milestone #5 Study Presentation Mechanism and Database Programming complete. Release before Study End.
- Milestone #4 and #6: 426 participants completed 100 ECGs. Recruitment stopped because study goals met and resources redirected to the Learning Adaptation Algorithm trial (SA3).
    - We have therefore proposed anchoring our learning intervention in one of three clinically relevant "Yes or No" (Binary) decisions:
        - Is this ECG abnormal? (Data Collection complete)
        - Should I call a STEMI Alert for possible cardiac catheterization? (Data collection complete for SA2; this clinical decision is basis for LAA in SA3; see below)
        - Should I shock this patient? (Animated set or rhythm strips in development)
- Manuscript in preparation (MS#5): Hierarchical Linear Modelling for ECG Skill Development.
    - Submission in preparation pending other manuscripts.

SA3: Task 1: Initial Algorithm Specification – Algorithm fully specified based on Prospective Cohort data.
SA3: Task 2: Programming of Algorithm and Feedback: Programming completed. Usability testing completed. SA3: Task 3: RCT has begun – recruitment in progress
SA3: Task 4: LAA analysis: Awaits data collection
SA3: Task 5: Programming of Demonstration Prototype – Prototype developed; Usability Testing ongoing
SA3: Task 6: Final Data Analysis and Manuscripts – See writing plan
SA3: Milestone #7 – Initial algorithm specified –Completed
SA3: Milestone #8 – LAA RCT recruitment – In Progress
SA3: Milestone #9 – Final Data Analyses and one manuscript completed – In Progress – See writing plan

What opportunities for training and professional development has the project provided?

- *The Investigators have benefited in terms of their professional development by learning from each other during weekly study phone calls, consultation with the advisory group (in particular Dr. David Swanson of American Board of Medical Specialties) as well as with Co-Investigator Dr. Jennifer Hill who is a premier educational statistician.*
- *The Investigators consulted internationally with experts in psychometrics (Dr. David Andrich of the University of Western Australia, expert in psychometrics and specifically Graded Response Models used for the main finding of the study; Dr. Chris Gibbons of Cambridge University, expert in Computer Adaptive Testing).*
- *The project has enhanced the professional development of our research staff including:*
    - *Ms. So Young Oh, Instructional Designer who has done the work of a Systematic Review which informs the grant work; on the basis of the grant work, she is pursuing a PhD in Medical Education where the dissertation work will involve the learning of ECG interpretation.*
    - *Mr. Eric Feng, Programmer has extended his skill with the representation of electrical signals into visual representations and, the use of Python and SQL databases*
    - *Mr. Ilan Reinstein and Ms. Jacqueline Gutman, the data scientists affiliated with the project have performed important stretch work at the interface between machine learning and psychometrics*

How were the results disseminated to communities of interest?

- To date:
    - Presentation at AMSUS meeting: "An Adaptive Tutor for Visual Diagnosis of ECGs". Booth organized by Dr. Joseph Loprieato of USU. National Harbor USA
    - Presentation at International Objective Measurement Workshop: "A Graded Response Model for Diagnostic Discernment in Visual Diagnosis". Chicago, Ill.
    - Presentation at Military Health System Research Symposium: "Capturing the entire learning space: a massive online ECG library as an advanced example of adaptive learning". Orlando FL

- Manuscripts in Preparation:

- See writing plan below

What do you plan to do during the next reporting period to accomplish the goals?

- We will complete the participant recruitment
- Finish the prototype development and release of the Database
- Data analysis will continue on the datasets in support of the planned manuscripts.

IMPACT:

What was the impact on the development of the principal discipline(s) of the project?

The present study has the potential to have a game-changing impact on the learning of visual diagnosis in that it points the way forward to developing adaptive tutors that allow precise calibration of clinicians in service of both patient safety and optimal clinician learning. In our systematic review of educational interventions for learning ECG interpretation, we found that research into the best way to teach this important skill is based on research performed on only 5000 subjects, of which fewer than 1000 are the advanced trainees or physicians who would interpret ECGs in their work. Even worse, this small research base is plagued by poor study design with only half of the studies making sound methodological comparisons. Finally, the outcome instruments used to determine skill are inadequate being made up of an insufficient number of questions (median number of questions per test: 11) and are often of sub-optimal psychometric validity and reproducibility.

Phase 1: A large ECG Database. We were able to collect 80,000 ECGs in their atomic XML format and parse the cardiologist interpretations such that they are fully labelled with one or more of the 120 American Heart Association Diagnostic Statements. In addition, the cases are ranked with an importance scale derived from an expert-consensus process. The planned open access database can be searched on any ECG numerical parameter, by AHA diagnostic statement and by key laboratory values. Search results yield not only a full pixel-for-pixel ECG representation but also prior and future ECGs for that (deidentified) patient, associated laboratory results and electronic health record data such as the final discharge diagnosis and the results of procedures such as cardiac catheterizations. ECGs can be downloaded either individually or as sets. The size of this open access teaching database dwarfs the currently available databases which number only a few hundred ECGs.

Phase 2: Case Characteristics Model. Our work has uncovered flaws in current assessment models for visual diagnosis. Standard assessment theories require a unidimensional factor structure in which all items are measuring the same thing. Visual diagnosis, however, will often be *multidimensional,* as evidenced by our prospective cohort study where 175 clinicians rated at least 100 ECGs each across 25 diagnoses representative of important diagnoses for emergency care. An exploratory factor analysis revealed a high level of dimensionality within the dataset, with 37 factors having eigenvalues > 1.0. On a concurrent systematic reivew of ECG learning interventions we found that, in the identified 53 studies, no two studies used the same assessment instrument, nor did any of the studies use assessment instruments with more than an ordinal scale, lacking rigorous psychometric evaluation. This situation would be akin to 53 studies of fever each using an unvalidated thermometer with a different scale.

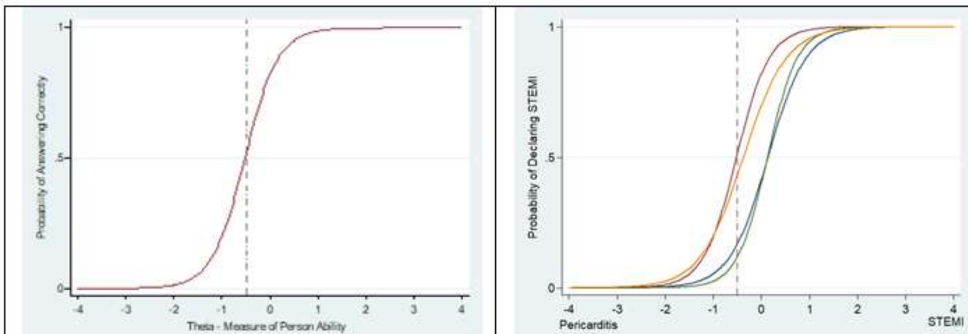
We have successfully shown that a unidimensional scale can be constructed by narrowing the diagnostic consideration to two frequently-confused diagnoses, such as Pericarditis and ST elevation myocardial infarction (STEMI). We created a scale (i.e., a set of cases and an underlying psychometric model) that ranged from “Definitely STEMI”, a diagnosis that demands a costly intervention (cardiac catheterization) to “Definitely Pericarditis”, which is treated with anti-inflammatory medications on an outpatient basis. This scale is continuous and unidimensional, which supports psychometric estimation and thus learner-adaptive item selection. It is also highly clinically relevant: the discrimination task leads to dramatically different treatment plans, and yet the stimuli are often confused for one another.

A transformational insight from this work is that, using a twist on standard IRT modelling, we are able to go beyond measuring each clinician's "skill" in the abstract, instead derving a nuanced picture of their decision-making and the degree of correspondence between their mental "scale" and the true scale of variability across patient cases.

We note that item-response-theory based graded response models (GRM) are well suited for representing a biological continuum so as to describe practitioner behavior for important clinical decisions. We will describe how item-response models are typically used to create measures of ability and then show a modification with transformational implications for modeling practitioner diagnosis and decision-making. We provisionally call this new method Decision Aligned Graded-Response Modelling (DA-GRM).

From abstract measures of ability to a clinically relevant scale describing decision-making. In item-response theory, which is the basis of computer adaptive testing, an item is represented by how difficult it is for a learner of a given ability (page4-top-left panel). However, unit-less item "difficulty" translates poorly to clinically meaningful metrics. We developed a set of 20 items that had differing levels of confusability between two ECG diagnoses: pericarditis and ST-elevation MI (STEMI). This diagnosis is important, as false positive declarations of STEMI waste considerable resources while false negatives have potentially life-threatening morbidity. While the biology is entirely different (pericarditis results in inflammation of the heart lining, while STEMI results from thrombotic occlusion of a coronary vessel), actual patient ECGs range from cases that are clearly one diagnosis or the other, along with a middle zone where even expert cardiologists are not able to make a confident declaration. In our pilot test, we had 36 clinicians of varying ability rank the 20 cases according to a 5-point graded scale ("Definitely Pericarditis" – "Probably Pericarditis" – "Either" – "Probably STEMI" – "Definitely STEMI"). See Guttman Grid.

A biologically consistent unidimensional scale. We modeled the dataset using a graded response model; however, we *inverted the focus of measurement*: rather than modeling cases' "difficulty" for given clinicians, we modeled clinicians' *propensities* for various responses to given cases. This Decision-Aligned Graded Response Modeling (DA-GRM) yielded a unidimensional scale with acceptable factor structure. Items ranged from -1.11 to +2.40 logits, suggesting appropriate coverage of the biological spectrum. Using a chart review blinded to the ECG diagnoses, we were able to verify that the factor score correlated with the clinical diagnosis in all but one case, a patient whose primary diagnosis was neither pericarditis nor STEMI but rather Major Gastrointestinal Hemorrhage, Row "IR" in Guttman grid below. Note that by modeling the biological spectrum, we have converted the scale from a unit-less representation of ability (left panel) to one where the location of a case is determined by the degree to which an ECG visually resembles STEMI vs. pericarditis (right panel).



Typical IRT Item Characteristic Curve in which the x-scale is a measure of a person's ability; a person of -0.5 logits ability is 50% likely to correctly diagnose the case.	DA-GRM Person Characteristic Curve. The x-scale is a measure of how typical of STEMI or Pericarditis a case is. For a case ranked at -0.5 logits, these 4 individuals would have different probabilities of diagnosing it a STEMI.

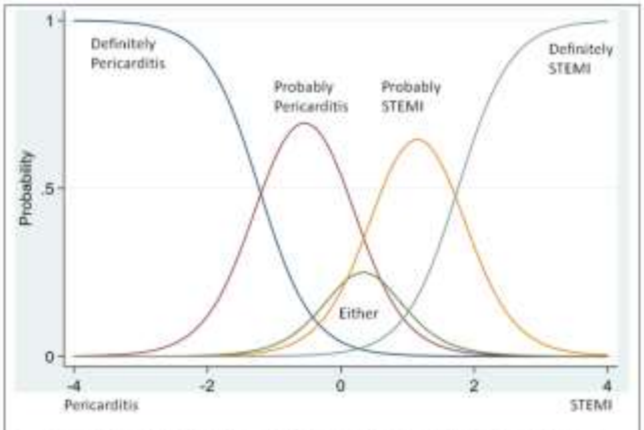
item	Logit	Std Error	Chart Review	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
BP	-1.11	0.13	Pericarditis	DP	EPS	EPS	DP	PP	DP	PP	PS	PS	EPS
IR	-1.03	0.13	STEMI	DP	PS	PP	PP	PP	DP	EPS	PS	PP	PP
BK	-0.87	0.12	Pericarditis	DP	DP	PP	PP	PP	DP	PP	PS	DS	EPS
IT	-0.82	0.12	Pericarditis	DP	DP	PP	PP	DS	PP	PP	DP	EPS	EPS
JW	-0.78	0.12	Pericarditis	DP	PP	PS	PP	PS	DP	PP	PP	PS	EPS
JO	-0.74	0.12	Pericarditis	DP	EPS	DS	DP	DS	PP	PS	PS	PS	PP
KN	-0.74	0.11	Pericarditis	DP	PP	PP	PP	PP	DP	PS	PP	PP	PP
CT	-0.73	0.13	Pericarditis	DP	PP	PP	PP	PP	PP	PP	DP	EPS	PP
IS	-0.48	0.11	Pericarditis	DP	PP	PP	PP	PP	PP	DP	PS	EPS	EPS
AX	-0.46	0.12	Pericarditis	DP	PP	DS	PS	PS	DP	EPS	PS	PP	PP
CI	-0.26	0.11	Pericarditis	DP	EPS	DS	PP	PS	PP	PS	DS	EPS	PP
BM	0.20	0.09	Pericarditis	EPS	DS	PS	PS	PP	PP	PS	PP	PS	EPS
LC	0.40	0.09	Pericarditis	PS	PS	PS	PP	PS	PP	DS	DS	DS	EPS
CL	0.53	0.10	STEMI	PS	DS	PS	PP	PS	DS	PS	DS	PS	EPS
ER	0.55	0.11	Pericarditis	DS	DS	DS	PS	PS	PP	PS	DS	DS	DS
EZ	0.63	0.10	STEMI	DS	DS	PS	PS	DS	DS	EPS	PS	PS	DS
CY	0.74	0.12	STEMI	PS	PS	DS	PS	DS	DS	PS	DS	PS	EPS
EE	1.19	0.17	STEMI	DS	DS	DS	PS	PS	DS	DS	DS	DS	PS
ED	2.09	0.23	STEMI	DS	DS	DS	DS	DS	DS	DS	DS	DS	EPS
LE	2.40	0.28	STEMI	DS	DS	DS	DS	DS	DS	DS	DS	DS	PS

The Guttman grid is an information rich basis for determining person ability. Following the example of Baldwin and colleagues (2009), we emphasized the degree to which an individual can discriminate amongst items, as opposed to standard practice, in which items discriminate between individuals. This DA-GRM considers an individual's skill as their ability to accurately reproduce the biological scale (from Pericarditis to STEMI) and situate items on it. This conceptual advance allows us to model a clinician's ability in terms of their demonstrated behavior classifying cases, rather than as an abstract person-level "ability". We will demonstrate the implications of this below. Consider the modified Guttman grid shown in the Figure. Items (cases) are rows and persons are columns (R1-R10; i.e., flipped from the traditional). We have ordered the cases based on each item's logit score, with the most pericarditis-representative case at top. The variability between individuals and amongst cases is readily apparent.

The Guttman data is the basis of a Decision Aligned Graded Response Model that can demonstrate an individual's ability to discriminate amongst items at differing levels of a biological continuum.

The Person Characteristic Curve (see figure at right) summarizes features to note for any one clinician's profile.

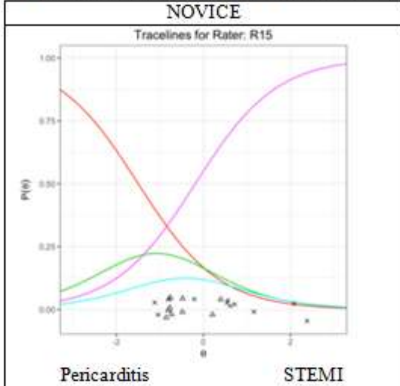
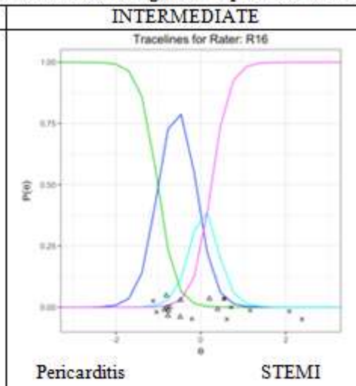
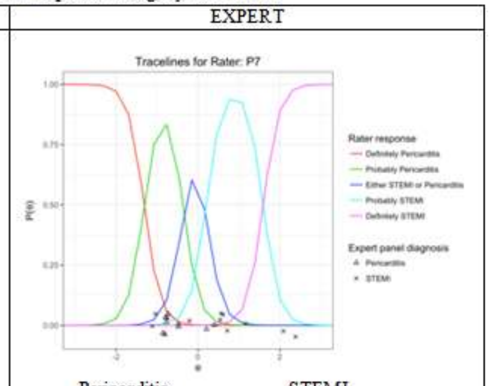
- The mode of any of the category curves is the case that would be exactly representative of that category, and least likely to be assigned an adjoining category.
- Confusion of two categories, i.e. overlap of the category curves, is minimal among experts and large among novices.
- For a given case anywhere along the spectrum, a vertical plumb line predicts how often each category would be chosen if the person were to encounter 100 such cases (i.e. proportions total to 1.0). Where the characteristic curves cross, those categories would be equally likely to be chosen.
- Differences between individuals in terms of the right- or leftward shift of a given characteristic curve (and the associated thresholds) indicate how that individual balances sensitivity with specificity, analogous to the bias parameter in signal detection theory. These metrics would be calculated.



Representative DA-GRM Person Characteristic Curve. X-axis is logit scale. Tracelines show this individual's use of the 5 response categories.

In the 3-paned figure below are DA-GRM Person Characteristic Curves that show three individuals' ability to use 5 categories to distinguish amongst cases. Ideal category curves show distinct separation with minimal overlap. The novice shows poorly separated categories with considerable overlap. While these data are cross-sectional, we noted a developmental progression between novices and experts to be exploited by the cognitive trainer.

Person Characteristic Curves by Level of Expertise for ECG Distinction of Pericarditis from ST Elevation MI
Three individuals who showed differences in their ability and willingness to use the five categories. Raters with more experience were more likely to use all response categories and to demonstrate greater separation between response category thresholds.

NOVICE	INTERMEDIATE	EXPERT
		
The **novice** really has only the two extreme categories and even those overlap considerably.	The **intermediate** uses 4 of the categories but only 3 with meaning as the Probably STEMI (teal) is barely distinguished from Definitely STEMI (pink).	The **expert** uses all 5 of the categories and they each have the intended meaning. Clear separation between Definitely STEMI (pink-rightmost) and Definitely Pericarditis (orange-leftmost) tracelines

DA-GRM can be the basis of decision-threshold modelling. Where the response categories are directly tied to an implicit decision option, the DA-GRM model allows us to examine practice variation between individuals. For example, in deciding whether to activate a "STEMI Alert', which involves mobilizing a team of professionals and preparing the cardiac catheterization laboratory, the threshold between categories 2 ("Probably Pericarditis") and 3 ("Either" – still risk of STEMI) is relevant. In the figure, we show the Boundary Characteristic Curves of 4 participants in our pilot study. The drawn plumb-line represents the proportion of times a STEMI Alert would be called for a case at -0.5 logits (e.g. row "IS" in Gutman Grid). Note that for this particular case, some individuals would theoretically activate 50% of the time, while others never. This represents a considerable training opportunity.

DA-GRM can be the basis of learning adaptation. In the current RCT, we are comparing a control (non-adapted group) to two groups with varying levels of adaptation based on the DA-GRM. The program determines and keeps a running estimate of the participants decision thresholds and compares them to an ideal practitioner. Deviations in thresholds result in increased practice by the individual at those thresholds.

This conceptualization is likely to extend to other domains of visual diagnosis. In the proposal we will describe, we propose to extend this technique to other considerations in ECG diagnosis and to prostate cancer histology. Using a dataset from the International Society of University Pathologists (ISUP) where 23 pathologists rated 50 Prostate Cancer slides according to a Gleason-based scale, we were again able to successfully model both discernment and decision-relevant practice variation. In the figure we show a resident physician (left) compared with an ISUP expert (right). Note that the two physicians would handle a -1.0 logit case completely differently, one with routine followup (expert) and the other with surgery (resident).

Summary. The key innovation is to begin considering the learning and practice task not only in terms of answering individual cases correctly or incorrectly but also to consider that the important transactional unit is actually the full *scale*. In the follow-on proposal, we will formalize the development of the scales and incorporate them into cognitive interventions.

<u>What was the impact on other disciplines?</u>

The DAGR model conceptualization can apply to other domains of visual diagnosis. We cannot speak of impact yet as we have not fully disseminated the results.

What was the impact on technology transfer?

The ECG Database and the process by which it was developed will allow open source access to ECGs and other such visual artifacts. The technology itself will be available open source on a public server.

What was the impact on society beyond science and technology?

The correct interpretation of ECGs and other forms of visual diagnosis will have a beneficial effect on the process of diagnosis. In turn, this has the potential to improve both the quality of healthcare and its safety.

CHANGES/PROBLEMS: *xx*

Changes in approach and reasons for change

Our initial approach had been to develop an adaptive tutor based on traditional forms of psychometric analysis based on item-response theory. As described above, this was suboptimal because of the multi-dimensionality of ECG interpretation. The DA-GR model approach was adopted instead which required a different form of data collection, slowing down the procedures of the grant, but with the benefit of a more impactful outcome.

Actual or anticipated problems or delays and actions or plans to resolve them

The ECG Database and the process by which it was developed will allow open source access to ECGs and other such visual artifacts. The technology itself will be available open source on a public server.

Changes that had a significant impact on expenditures

None to report

Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents

None to report

Significant changes in use or care of human subjects

None to report

Significant changes in use or care of vertebrate animals.

Not applicable

Significant changes in use of biohazards and/or select agents

Not applicable

PRODUCTS:

Publications, conference papers, and presentations

Journal publications. Writing Plan

Working Title	First Author	Senior Author	Other authors	Description	Type	Target Journal		Status
Systematic Review of ECG Teaching Interventions	Oh	Cook	Nicholson, Fairbrother, Smeenk, Van Gerven	SR of teaching interventions	Literature review	Medical Education		Manuscri
Population-based Item Sampling for Online Learning	Cook	Gelman	RH, ML, MP	Population-based sampling of ECGs	MUSE-Dataset, Prospective Cohort study	General Education Literature		Pre-Manuscri Gelman Masters Thesis
Focus Group Rankings of Importance by Experts	Penalo	Pusic	Elysee,	Summary of successive focus groups of experts.	Focus Groups	Interdisciplinary journal		Pre-Manuscri
Graded Response Models for Assessing Diagnostic Discernment: An Example Using Electrocardiogram Interpretation.	Pusic	Hatala	Cook, Friedman, Rosenzweig, Lorin, Line berry, Tong		Hip psychometrics			Manuscri

The Difficulty with ECG Difficulty: Statistical Models of ECG Interpretation	Pusic	Line berry	Cook, Hatala, Oh, Lorin, Smith	Multi-faceted description of how the "item difficulty" conceptualizaiton does not work for ECGs.	Combination of UBC and DoD datasets. Cardiology Fellow Pairwise Comparisons	Journal that represents colour properly ? JGIM		Manuscri
Hierarchical Linear Modelling for ECG Skill Development	Line berry	Pusic	Jennifer Hill, DC, RH	Using three datasets, will demonstrate the additive value of HLM in individual person estimates	.elbow radiology; ECG prospective; RQI	AHSE		Manuscri
Systematic Review of ECG Assessment Instruments				Complementary SR that focuses on how poorly we assess ECG expertise	Prospective cohort -- Factor Analysis.	AHSE		Pre-Manuscri

Books or other non-periodical, one-time publications.

Not applicable

Other publications, conference papers, and presentations.

- Presentation at AMSUS meeting: "An Adaptive Tutor for Visual Diagnosis of ECGs". Booth organized by Dr. Joseph Loprieato of USU. National Harbor USA
- Presentation at International Objective Measurement Workshop: "A Graded Response Model for Diagnostic Discernment in Visual Diagnosis". Chicago, Ill.
- Presentation at Military Health System Research Symposium: "Capturing the entire learning space: a massive online ECG library as an advanced example of adaptive learning". Orlando FL

Website(s) or other Internet site(s)

In development is the ECG Database which will be placed on a public server. Screen capture below.

Technologies or techniques

Not applicable

Inventions, patent applications, and/or licenses

Not applicable

Other Products

Not applicable

PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

What individuals have worked on the project?

Name:	Marc Triola, MD (NYU)
Project Role:	Co-Investigator
Contribution to Project:	No Change
Name:	Silas Smith, MD (NYU)
Project Role:	Co-Investigator
Contribution to Project:	No Change
Name:	Jennifer Hill, PhD, (NYU PRIISM)
Project Role:	Co-Investigator
Contribution to Project:	No Change
Name:	Barry Rosenzweig, MD,
Project Role:	(NYU) Co-Investigator
Contribution to Project:	No Change
Name:	Greta Elysee, (NYU)
Project Role:	Program Coordinator
Contribution to Project:	No Change
Name:	Laura Penalo RN, (NYU)
Project Role:	Project Manager
Contribution to Project:	No Change
Name:	Eric Feng, (NYU) Programmer
Project Role:	No Change
Contribution to Project:	
	So–Young Oh, (NYU)
Name:	Instructional Designer
Project Role:	No Change
Contribution to Project:	
Name:	J. Gutman – replaced by Ilan Reinstein, (NYU)
Project Role:	Statistical Analyses
Researcher Identifier (e.g. ORCID	none
ID): Nearest person month worked:	1
Contribution to Project:	No change

Name:	Martin Pusic, MD, (NYU) Principal
Project Role:	Investigator
Contribution to Project:	No Change
Name:	David Cook, MD, (MAYO
Project Role:	CLINIC) Co-Investigator
Contribution to Project:	No Change
Name:	Rose Hatala, MD, (UBC)
Project Role:	Co-Investigator
Contribution to Project:	No Change
Name:	Matthew Lineberry, PhD, (KUMC)
Project Role:	Co-Investigator
Contribution to Project:	No Change

Xx

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?

No

What other organizations were involved as partners?

Not applicable

SPECIAL REPORTING REQUIREMENTS

COLLABORATIVE AWARDS: *Not applicable*

QUAD CHARTS: *Previously supplied.*

APPENDICES: MHSRS Abstract.

Capturing the entire learning space: a massive online ECG library as a forward-looking example of adaptive learning.

BACKGROUND:

As healthcare is increasingly carried out in a digital environment, where large amounts of process data are captured, we are increasingly able to use that data to increase the depth and breadth of simulations in service of education and quality improvement. Here we consider the example of electrocardiograms (ECGs) as but one type of digital artifact, like radiographs, clinical photographs, and laboratory results, which can be captured via standardized protocol as digital representation from the patient. These authentic clinical artifacts are no longer scarce teaching resources to be preserved and presented in small curated personal sets but rather can be delivered at scale to serve the full breadth and depth of the military learner's needs.

Our Current Approach to Teaching & Learning Visual Diagnosis of ECGs…….is Not Effective. Electrocardiograms are read over 300 million times in the U.S. annually. A systematic review by Salerno shows that in cases where the ECG interpretation is central to the case, errors can lead to management changes in >8%, inappropriate admission to hospital in ~3% and, most worrisome, inappropriate discharge in 0.6%. **…. is not Evidence-Based –** A recent IOM report explicitly calls for use of "educational approaches that are aligned with evidence from the learning sciences."; **…and our current use of teaching files is inadequate** -- Current ECG learning collections fail for a number of reasons. First, their breadth can be insufficient. The American Heart Association (AHA) has generated a list of potential ECG diagnoses numbering 120 statements. Second, the depth of the collection is often insufficient. For each of the 120 diagnoses, there can be variability in appearance such that trainees might need to see 10-20 examples to have a sense of the range of presentation. Third, in the same way a butterfly collection bears little resemblance to the their distribution in the wild, ECG collections need to consider the spectrum of presentation to a clinician in terms of the base rate of abnormality, the variability of normal presentations and the effects of presentation sequence.

In this project we describe harvesting over 80,000 digital ECGs in a consecutive manner so as to preserve the unique spectrum of Emergency Department clinical presentation. We collect associated clinical data including the proximate outcome based on available serum and/or advanced diagnostic techniques such as echocardiography, catheterization or electro-physiologic studies. A subset of ECGs was drawn from the set and validated on a wide variety of learners.

METHODS:
This study had two phases: assembly of the ECG database and subsequent validation on a prospective cohort of learners and expert clinicians.
Database Assembly: The initial dataset was comprised of 98,420 ECGs obtained from the General Electric MUSE system of the NYU Langone Medical Center. These records represent all patients treated in the Perelman Department of Emergency Medicine who received at least one ECG during their ED visit, from the 5 year period immediately preceding the study. It is institutional policy that the ECGs performed during an ED visit be over-read by a cardiologist. Only one "most relevant" ECG is formally reported per ED visit. ECGs of patients younger than 18 or older than 80 years were excluded. Fifteen cases were excluded for data quality issues in the structure of the ECG data files. A total of 81,287 resting electrocardiograms remained. The cardiologist interpretations were mapped to a controlled vocabulary (AHA Diagnostic Statement List).

Cohort validation study:
We subsequently recruited participants at three levels of expertise: medical student, resident or attending clinician. In an online browser-based application, each individual interpreted >100 unknown ECG cases drawn from the database. ECGs were selected so as to represent the larger population of cases including one-third normal cases and a range of rhythm and ischemic pathologies, some rare (Brugada) and some common (Atrial arrhythmias). We tested the validity of the ECG answers parsed from the cardiologists' interpretation using both item-response theory 1-PL parameters (mean difficulty and standard error) as well as qualitative analysis of case-by-case user feedback, for which we report dominant themes. Exploratory factor analysis was performed to examine the factor structure.

RESULTS:
The 81,287 ECGs were each successfully coded with at least one diagnosis. The most common diagnosis was normal (N=25,151, 31%) followed by T-wave abnormalities (15%), Sinus Tachycardia (13%), Left atrial enlargement (13%). The median number of diagnostic statements per ECG was 2 but with a broad range (IQR 1, 6) and with the 90%ile having > 6 diagnostic statements per ECG. Of 120 AHA diagnostic categories, 75 were represented with at least 10 examples in the database. As an example, there are over 8000 ECGs with mention of myocardial infarction.

Medical students (N=25), residents (N=126), attending EM physicians (N=9) and nurses (N=13) each completed >100 ECGs within the system resulting in >20,000 ratings in total.

Item analysis showed that normal cases (N=35, Mean proportion correct 0.65, Standard deviation 0.25) were more often correctly specified than cases with pathology (N=65, 0.59 $\pm$ 0.27) but the difference did not achieve statistical significance. Indeed, some of the most difficult cases were in fact ECGs that were classified as normal by the treating cardiologist. Key diagnoses identified incorrectly in >60% of cases included Type II AV Block, Atrial fibrillation and certain locations of myocardial infarction. Well identified were Ventricular arrhythmias as well as most normal ECGs. On IRT analysis the cases had a mean difficulty of -0.95 logits (indicating approximately one standard deviation easier than the mean difficulty for the population of ECGs) with mean standard errors of 0.56 logits. For the 100 items, the exploratory factor analysis showed fully 32 factors with eigenvalues >1.0 suggesting high dimensionality. The main themes in the qualitative comments included: disagreement with diagnosis, negotiation of correctness, need for clinical correlation. A number of the statements could have served as further feedback to learners.

CONCLUSIONS:
We have described an approach to learning a diagnostic skill in which, instead of learning from small samples, recruits would have the opportunity to swim in the entire learning space. Every ECG type of significance to an emergency medicine clinician will be somewhere in the 80,000 curated ECGs in the NYU database. The set properly represents the full range of possible pathologies including the more difficult ECGs that have multiple diagnoses. Additionally, we noted that a number of the normal ECGs can also be relatively difficult.

In the current study we have shown how the ECGs can be calibrated by having a cohort of representative learners interpret the ECGs under a "you make the call" context. Each ECG completion is a vote for the difficulty of that particular ECG and gives information as to which diagnoses are difficult or easy. In addition, we were able to qualitatively derive even finer-grained information as to the case-specific features of each ECG.

While the database is useful pedagogically in its current form, there are a number of issues to work out before it can realize its full potential. The factor analysis revealed just how multi-dimensional the act of interpreting an ECG is. As a result, measurement of a clean latent variable of ECG ability will require the development of more specific subscales summing to a holistic ability estimate. Specific targeting of ECG cases to the ability of the learner will need to await such work.

All of this calibration information can inform algorithmic approaches that have the advantage of quickly adapting to learner needs and to changes in illness patterns. We point out the direct relevance of improved ECG interpretation for military medicine, given the specific utility of the ECG in toxicologic exposures, trauma and health screening of recruits. Future work will include drawing ECG sets with these specific indications in mind.