Mass General Brigham

The Growing Demand for AI in Clinical Medicine

William Gordon, MD, MBI, FAMIA

Director, Solution & Experience, Digital Care Transformation - Mass General Brigham
Associate Physician, Hospital Medicine Unit - Brigham & Women's Hospital
Assistant Professor of Medicine, Harvard Medical School

1

Disclosures

Research funding from IBM and Pew Charitable Trusts, unrelated to this topic

Consultant for ONC/HHS, Novocardia, unrelated to this topic

2

2

William Gordon, MD, MBI

- Software Developer turned Physician
- Asst Professor, Harvard Medical School, Brigham and Women's Hospital, by way of Weill Cornell Medical College and Massachusetts General Hospital
- Internal Medicine & Clinical Informatics
- Digital Health, Information Security, Data Interoperability, ML
- Currently lead product development at MGB Personalized Medicine, see patients at Brigham and Women's Hospital, research faculty at Harvard Medical School

3

3

Last Mile

The expression "last mile" describes challenges in delivering telecommunication connectivity to people's homes--for example, the "last mile" of wiring required to connect an actual landline phone to a central exchange.

It is also applied in supply-chain management settings--getting a good, like a medication or perishable, to its destination.

The last mile is usually disproportionately expensive and inefficient

4

4

5

5

Last Mile

The Last Mile problem is seen in medicine as well

We usually know what needs to happen. Getting it done is much harder.

6

6

Diabetic Retinopathy

- Diabetic Retinopathy is a major source of morbidity worldwide – and the largest source of impaired vision globally
- Physiologically, it is one of the numerous microvascular complications of diabetes – chronic hyperglycemia leads to damage within the vessels supplying the retina, causing structural, anatomic, and vascular changes, which ultimately leads to vision loss
- Diabetic retinopathy is **widespread, usually asymptomatic, and leads to significant morbidity**—it is an ideal condition to "screen" for, and annual screening is widely recommended for patients with diabetes
- Early intervention – along with better glucose and BP control – can prevent severe visual loss

7

7

Diabetic Retinopathy – How to Screen?

Fortunately, screening for diabetic retinopathy is straightforward…

… You just need an ophthalmologist or optometrist to look at your eyes (or a picture of your eyes)

8

8

Diabetic Retinopathy

Original Investigation | Innovations in Health Care Delivery FREE

December 13, 2016

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD[1]; Lily Peng, MD, PhD[1]; Marc Coram, PhD[1]; Martin C. Stumpe, PhD[1]; Derek Wu, BS[1]; Arunachalam Narayanaswamy, PhD[1]; Subhashini Venugopalan, MS[1,2]; Kasumi Widner, MS[1]; Tom Madams, MEng[1]; Jorge Cuadros, OD, PhD[3,4]; Ramasamy Kim, OD, DNB[5]; Rajiv Raman, MS, DNB[6]; Philip C. Nelson, BS[1]; Jessica L. Mega, MD, MPH[7,8]; Dale R. Webster, PhD[1]

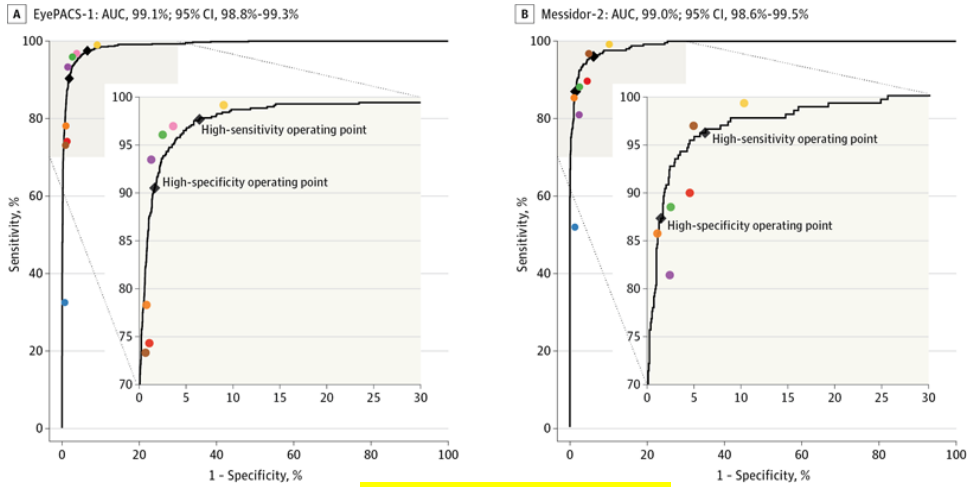
Question How does the performance of an automated deep learning algorithm compare with manual grading by ophthalmologists for identifying diabetic retinopathy in retinal fundus photographs?

Meaning Deep learning algorithms had high sensitivity and specificity for detecting diabetic retinopathy and macular edema in retinal fundus photographs.

9

9

Diabetic Retinopathy



AUC = .991 and .990

10

10

Diabetic Retinopathy

Conclusions and Relevance In this evaluation of retinal fundus photographs from adults with diabetes, an algorithm based on deep machine learning had high sensitivity and specificity for detecting referable diabetic retinopathy. Further research is necessary to determine the feasibility of applying this algorithm in the clinical setting and to determine whether use of the algorithm could lead to improved care and outcomes compared with current ophthalmologic assessment.

11

11

Real World Study

RESEARCH-ARTICLE OPEN ACCESS

A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy

Authors: Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, Laura M. Vardoulakis Authors Info & Affiliations

How would this algorithm work in a real world, resource-constraint setting?

- Thailand has an increasing prevalence of diabetes and complications related to diabetes including diabetic retinopathy
- However, there are only 1500 ophthalmologists, and 200 retinal specialists in the entire country – for 4.5m patients with diabetes
  - Backlog to screen
  - Screening backlog causes backlog for delays in treatment

12

12

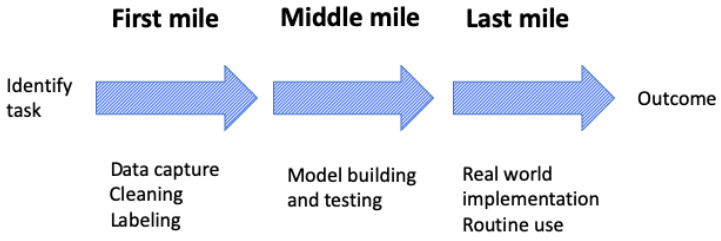
Real World Study

What did the researchers discover?

- Workflow varied significantly between sites
- Lighting conditions made getting images more challenging
  - Of the 1838 initial images, 21% could not be automatically screened due to quality
  - Challenging to fix – how to make a room darker if it's the same room used for other purposes?
  - Adjusted the protocol so that ungraded images required ophthalmologist, reverting to existing processes
- Bandwidth issues – in some clinics the back and forth took 1-2 minutes, which slowed the screening queue
  - Internet went down for 2 hours in one clinic, patients had to wait

13

13

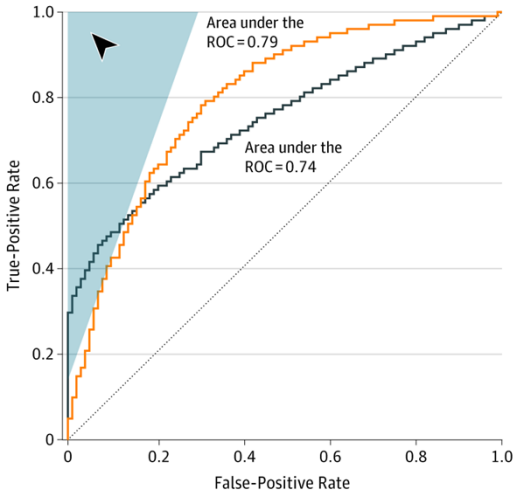


14

Source: Coiera E. The Last Mile: Where Artificial Intelligence Meets Reality. J Med Internet Res 2019;21(11):e16323

14

How do we make models clinically useful?



- Predicting hospital readmissions – "orange" model seems better than the "blue" model
- But—given the costs/benefits of actually preventing a readmission, the "triangle" is where the actual benefit is
- "Blue" curve has better synergy with the real-world potential for impact

Source: Shah, N et al. JAMA. 2019;322(14):1351-1352.

15

15

ML + Real World Implementation

How do we do this right? How do we pair a "good" model with the right environment so that it can be successful?

Can a machine learning model that predicts patient deterioration actually improve mortality?

SPECIAL ARTICLE

Automated Identification of Adults at Risk for In-Hospital Clinical Deterioration

Gabriel J. Escobar, M.D., Vincent X. Liu, M.D., Alejandro Schuler, Ph.D., Brian Lawson, Ph.D., John D. Greene, M.A., and Patricia Kipnis, Ph.D.

- Based on a previously validated model ("Advanced Alert Monitor") that identifies full-code inpatients at high risk of deterioration and predicts risk of ICU transfer or death (Kipnis et al., PMID 27658885)
- Logistic regression model
  - No deep learning!
- Dozens of predictors – including labs, vital signs, time of day, season, admission type
- Model had performed well previously (AUC=.82)
- A score of 5 indicates a 12-hour risk of clinical deterioration of 8% or more.

16

16

ML + Real World Implementation

How was this implemented / studied?

Intervention:

- Specially trained remote nurses with no other concurrent patient care responsibilities were alerted to a patient's risk of deterioration (24/7) based on the ML model
  - Clinical team was *not* alerted
- Study RNs would review the patient chart and alert the clinical team if appropriate
- Main outcome was 30d mortality; also looked at LOS, ICU admission, favorable status at 30 days

Methods:

- Monitored across more than 500k (!!) hospitalizations and 21 hospitals from 2015-2019
- Eligibility: Patients admitted to inpt/step down (no ICU)
- ~43k (7%) of study population "alerted" based on the AAM model
- Randomization: 15k received intervention, 28k did not

17

17

ML + Real World Implementation

What did they find?

Table 2. Adjusted Outcomes in the Eligible Population, with Comparison between the Intervention Cohort and Comparison Cohort.*

Variable	Study Population	Adjusted Relative Risk or Hazard Rate Ratio (95% CI)
Target population		
No. of hospitalizations	43,949	
No. of patients	35,669	
ICU admission within 30 days after alert		0.91 (0.84–0.98)
Death within 30 days after alert		0.84 (0.78–0.90)
Favorable status at 30 days after alert†		1.04 (1.02–1.06)
Hospital discharge, as assessed by proportional-hazards analysis		1.07 (1.03–1.11)
Survival, as assessed by proportional-hazards analysis		0.83 (0.78–0.89)
Nontarget population		
No. of hospitalizations	504,889	
No. of patients	313,115	
ICU admission within 30 days after admission		0.94 (0.89–0.99)
Death within 30 days after admission		0.97 (0.93–1.02)
Favorable status 30 days after admission†		1.00 (0.99–1.00)
Hospital discharge, as assessed by proportional-hazards analysis		0.98 (0.97–0.99)
Survival, as assessed by proportional-hazards analysis		0.99 (0.96–1.03)

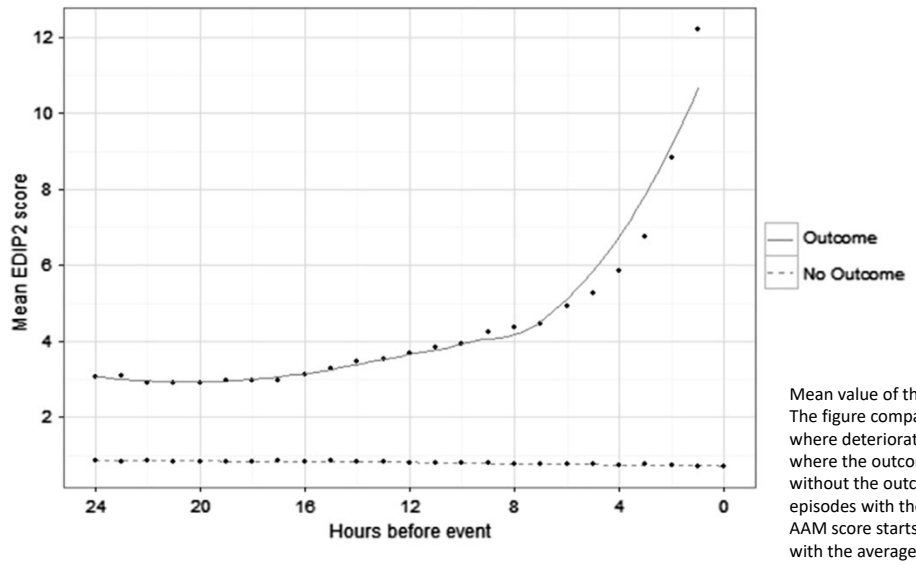
Patients in intervention were 17% more likely to survive (statistically significant)

Other outcomes (ICU admission, clinical status at 30d) also improved, though LOS was increased

18

18

Kipnis et al. Page 15



Mean value of the AAM score in the 24 h prior to event. The figure compares score trajectories among episodes where deterioration did and did not occur. Episodes where the outcome occurred are shown at. Episodes without the outcome have generally lower scores than episodes with the outcome. The figure shows that the AAM score starts increasing about 8 h prior to the event with the average score being close to 12 at the time of the event

19

19

Evolving and Scaling ML

One of the core challenges for any predictive model is understanding how it was initially developed

- What was the care setting? AMC, community?
- What types of patients? Gender, race, age?
- When was it developed? Has clinical practice changed?

Ultimately, we need a clinician who is looking at the output of a machine learning model to ask, "Was this model developed against patients similar to the patient in front of me right now?"

Ex: Framingham study – representative of one place (Framingham) with minimal diversity

Table 1

Geographic and ethnic identities of participants in the FHS population

Geographic diversity		Geographic diversity	
Birth place	*Percent*	*Ethnicity*	*Percent*
Framingham	19.15	England, Scotland and Wales	19.86
Other regions of Massachusetts	40.31	Ireland	14.95
Other regions of New England	9.79	Italy	19.00
Other US regions	9.81	French Canadian	2.26
Canada	5.46	Other Canadian	2.63
England, Scotland, Wales	1.28	Eastern Europe	5.93
Ireland	1.37	Western Europe	31. 77
Italy	7.26	Other	2.67
Other	3.32	Unknown	0.94
Unknown	2.26		
Total	100.00	Total	100.00

Source: Govindaraju, et al., 2008: PMID 19010253

20

20

Evolving and Scaling ML

This "problem" is particularly pronounced for machine learning, where models are being developed extremely rapidly, implemented even more rapidly, without the statistical rigor and awareness of historical precedents

JAMA Internal Medicine | Original Investigation

External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients

Andrew Wong, MD; Erkin Otles, MEng; John P. Donnelly, PhD; Andrew Krumm, PhD; Jeffrey McCullough, PhD; Olivia DeTroyer-Cooley, BSE; Justin Pestrue, MEcon; Marie Phillips, BA; Judy Konye, MSN, RN; Carleen Penoza, MHSA, RN; Muhammad Ghous, MBBS; Karandeep Singh, MD, MMSc

A recent study looking at a popular sepsis prediction model in the Epic EHR is a good example of where we can see the hazards of ML play out in real clinical settings

Source: Govindaraju, et al., 2008: PMID 19010253

21

21

Evolving and Scaling ML

JAMA Internal Medicine | Original Investigation

External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients

Andrew Wong, MD; Erkin Otles, MEng; John P. Donnelly, PhD; Andrew Krumm, PhD; Jeffrey McCullough, PhD; Olivia DeTroyer-Cooley, BSE; Justin Pestrue, MEcon; Marie Phillips, BA; Judy Konye, MSN, RN; Carleen Penoza, MHSA, RN; Muhammad Ghous, MBBS; Karandeep Singh, MD, MMSc

Wong, et al. look at the Epic Sepsis Model (ESM), a popular, proprietary prediction model used nationally

Set out to externally the model by retrospectively looking at close to 30,000 patients "exposed" to the ESM score (single-center), which typically runs every 15 minutes. Compared ESM to patients with confirmed sepsis and contemporary clinical practice (e.g. early administration of antibiotics)

ESM "missed" 67% of patients with sepsis; falsely alerted for 18%, and "identified" sepsis in 7% who did not receive early antibiotics

22

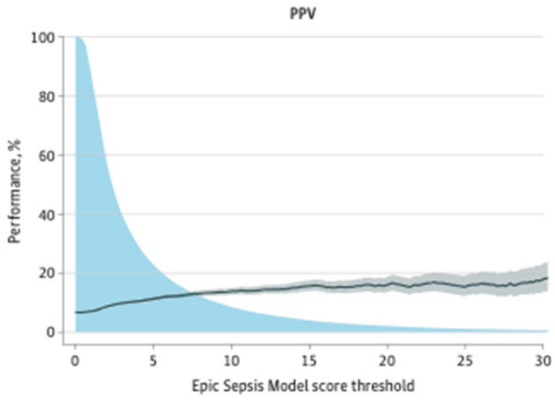
22

Evolving and Scaling ML

JAMA Internal Medicine | Original Investigation

External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients

Andrew Wong, MD; Erkin Otles, MEng; John P. Donnelly, PhD; Andrew Krumm, PhD; Jeffrey McCullough, PhD; Olivia DeTroyer-Cooley, BSE; Justin Pestrue, MEcon; Marie Phillips, BA; Judy Konye, MSN, RN; Carleen Penoza, MHSA, RN; Muhammad Ghous, MBBS; Karandeep Singh, MD, MMSc



Models can run over a variety of "scores" or "thresholds" – how did the ESM perform as the score was adjusted?

Here, blue-shaded is % of patients classified as positive based on the ESM score

23

23

Evolving and Scaling ML

What are the issues at play?

- Epic's model is proprietary – nobody knows how it was developed, on which patient population, how it was validated, how it is maintained / updated
- Yet it is so well implemented into clinical workflow – significantly lowers implementation barriers; no new vendors, falls into existing governance, etc.
- Are these EHR-based models simply encoding clinical knowledge already known?
  - e.g. if a stat order for vancomycin is a predictor of sepsis, doesn't that suggest that clinician already knows the patient is sick?
- Will EHR-based ML models ever stand alone, or will they need to be tightly coupled with an implementation like the Stanford study?
- Out of scope: how do we *regulate* machine learning models? Should they be regulated like other devices? Or is just fancy "Clinical Decision Support" ?

24

24

Evolving and Scaling ML

Data Drift

CORRESPONDENCE

The Clinician and Dataset Shift in Artificial Intelligence

Dataset shift / drift occurs when a ML model underperforms because there is a mismatch between the data set with which it was developed, and the data ultimately used for runtime/deployment

Can have many causes:

- Change in underlying technology
- Change in the population & setting (new demographics, new hospital)
- Changes in behavior
- Changes in clinical practice

Source: Finlayson, et al. NEJM – Correspondence – 7.15.2021

25

25

Digital Medicine in the Real World

Conclusions:

- Tremendous opportunity to improve clinical care
- Important to do “basic science” aspect of this work
- Critical to understand real-world workflow implications and practicalities

26

26

Where do we go from here?

27

27

Conclusion

How do we deliver the promise of "digital medicine" to the "last mile"?

Workflow & Collaboration: Those building *must* be integrated into the clinical environment.

Study *clinical outcomes*, not just statistical performance measures. Build evidence.

Monitor models over time: things change – models are living systems and need constant monitoring and assessment.

The "last mile" is hard – requires multidisciplinary teams, organizational focus, collaboration, and consensus.

28

28

Thank you!

wjgordon@partners.org