Journal Club: Clinical Impact and Quality of Randomized Controlled Trials Involving Interventions Evaluating Artificial Intelligence Prediction Tools

William Hersh, MD
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
School of Medicine
Oregon Health & Science University
Portland, OR, USA
http://www.ohsu.edu/informatics
Email: hersh@ohsu.edu
Web: www.billhersh.info
Blog: https://informaticsprofessor.blogspot.com/
Twitter: @williamhersh

References

Donoho, D., 2017. 50 Years of Data Science. Journal of Computational and Graphical Statistics 26, 745–766. https://doi.org/10.1080/10618600.2017.1384734

Hersh, W., 2021. Translational Artificial Intelligence: A Grand Challenge for AI. Informatics Professor. URL https://informaticsprofessor.blogspot.com/2021/07/translational-artificial-intelligence.html (accessed 9.28.21).

McCarthy, J., Feigenbaum, E.A., 1990. In Memoriam: Arthur Samuel: Pioneer in Machine Learning. AIMag 11, 10–10. https://doi.org/10.1609/aimag.v11i3.840

Payne, P.R.O., Bernstam, E.V., Starren, J.B., 2018. Biomedical informatics meets data science: current state and future directions for interaction. JAMIA open 1, 136–141. https://doi.org/10.1093/jamiaopen/ooy032

Rajpurkar, P., Chen, E., Banerjee, O., Topol, E.J., 2022. AI in health and medicine. Nat Med 1–8. https://doi.org/10.1038/s41591-021-01614-0

Shortliffe, E.H., 2019. Artificial Intelligence in Medicine: Weighing the Accomplishments, Hype, and Promise. Yearb Med Inform 28, 257–262. https://doi.org/10.1055/s-0039-1677891

Straus, S.E., Glasziou, P., Richardson, W.S., Haynes, R.B., 2018. Evidence-Based Medicine E-Book: How to Practice and Teach EBM, 5th edition. ed. Elsevier.

Topol, E., 2019. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, Illustrated Edition. ed. Basic Books, New York.

Zhou, Q., Chen, Z.-H., Cao, Y.-H., Peng, S., 2021. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. NPJ Digit Med 4, 154. https://doi.org/10.1038/s41746-021-00524-2

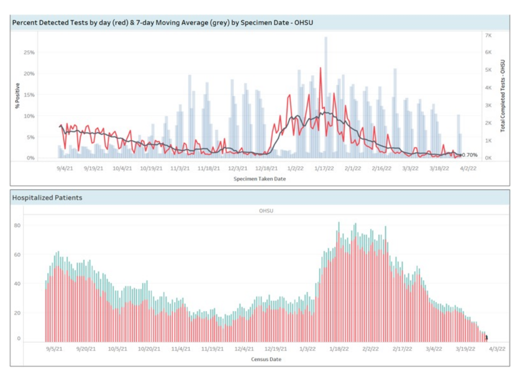
Journal Club: Clinical Impact and Quality of Randomized Controlled Trials Involving Interventions Evaluating Artificial Intelligence Prediction Tools

William Hersh
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Conference – March 31, 2022 – PDF of slides and references at www.billhersh.info or from @williamhersh

1

We're back to the office … somewhat

- In-person classes and fellows' meeting will be in person
- Conference will offer presenters to speak in-person if they desire – will continue to stream as always
- Save the dates – in-person graduation and DMICE banquet weekend of June 4-5
- Faculty returning to office mostly 1-2 days per week but still very accessible via email and WebEx
- Staff returning to office later
- Good news of late for COVID-19 locally – hopefully will stay
- Wearing of masks optional but low-threshold



As of Wednesday, March 30

Patients hospitalized with COVID-19

- OHSU: 4
- Hillsboro Medical Center: 6

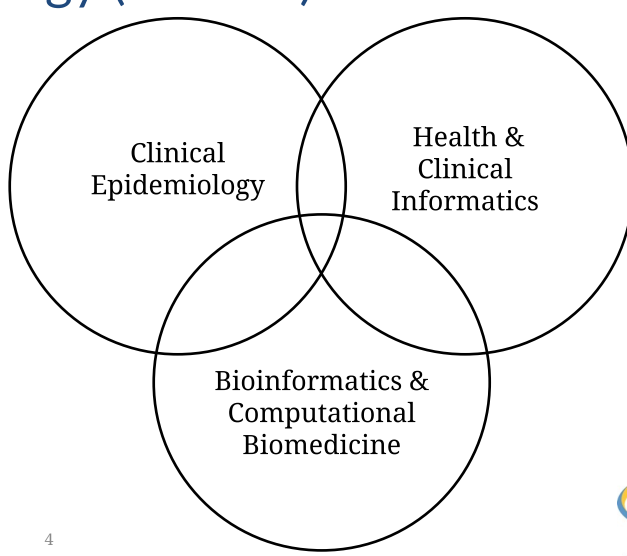
2

AI meets EBM – beyond data wrangling and modeling

- Some background on evidence-based medicine (EBM), clinical informatics, and machine learning
- Systematic review of clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence (AI) prediction tools
- Discussion on clinical evaluation of AI, including at OHSU

Department of Medical Informatics & Clinical Epidemiology (DMICE)

- Clinical Epidemiology
  - Evidence-based medicine
  - Systematic reviews
- Health & Clinical Informatics
  - Clinical informatics systems
  - Applied AI
- Bioinformatics & Computational Medicine
  - Omics
  - Data science



This talk will address a topic at the overlap of the three areas of DMICE

- A systematic review
  - Clinical Epidemiology
- Of the clinical predictive AI tools
  - Health & Clinical Informatics
- Applying data science and machine learning
  - Bioinformatics & Computational Medicine

Some level-setting – clinical epidemiology and evidence-based medicine (EBM)

- EBM applies the best evidence for making clinical decisions (Straus, 2018)
  - Prefer experimental studies but can use observational studies when appropriate
- Most clinical questions fall into four categories, each of which have best study types
  - Treatment – randomized controlled trial (RCT)
  - Diagnosis – comparison vs. gold standard
  - Harm – cohort and case-control studies when RCT not possible
  - Prognosis – prospective cohort studies
- For all study types, when sufficient number have been done
  - Can carry out a systematic review
  - If data across studies homogeneous, can perform meta-analysis

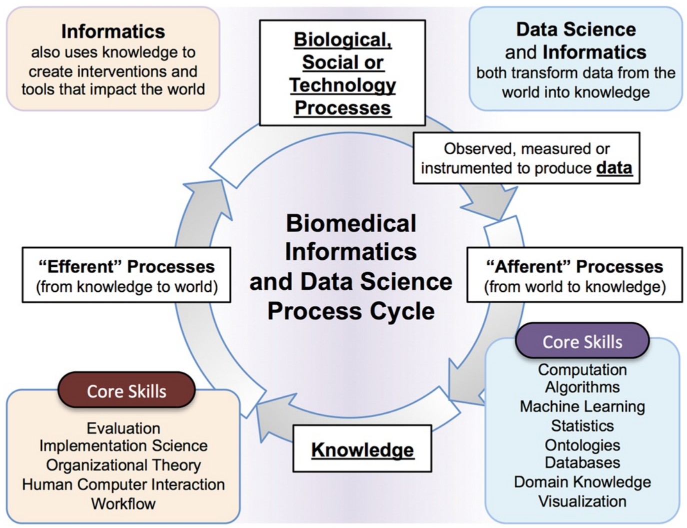
More level-setting – informatics

- A major activity of clinical informatics has been application of AI to improving patient care (Shortliffe, 2019)
- First generation in 20th century
  - Focus on hand-crafted knowledge bases
  - Computers lacking power, GUIs, Internet, etc.
  - Led to "AI winter" in late 1980s and beyond
- Resurgence in 21st century
  - Driven by advances in machine learning, especially deep learning
  - Based on large amounts of data and plentiful computer power and networks
  - Modest impact (as of 2022) in clinical care

More level-setting – data science

- Data science – "science of learning from data" (Donoho, 2017)
  - A data scientist is a "person who is better at statistics than any software engineer and better at software engineering than any statistician"
- Recent achievements driven by advances in machine learning (Arthur Samuel in 1959: "field of study that gives computers the ability to learn without being explicitly programmed" McCarthy, 1990)
  - Especially deep learning (Topol, 2019; Rajpurkar, 2022)

Final level-setting – informatics and data science (Payne, 2018)



DMICE Conference 3/31/22

9

Let us now ask: what is the evidence of clinical benefit of AI?

- Best evidence for interventions (treatment or prevention) comes from RCTs
  - Ideally RCTs that are well-conducted, generalizable, and well-reported
- Although there are other clinical questions that can be answered about AI
  - Diagnosis – can AI methods improve ability to diagnose disease?
  - Harm – can AI identify harms from environment, medical care, etc.?
  - Prognosis – can AI inform the prognosis of health and disease?
- Ultimately, however, AI interventions must be demonstrated experimentally to benefit patients, clinicians, and populations
  - Some instances when RCTs are infeasible so observational studies may be justified

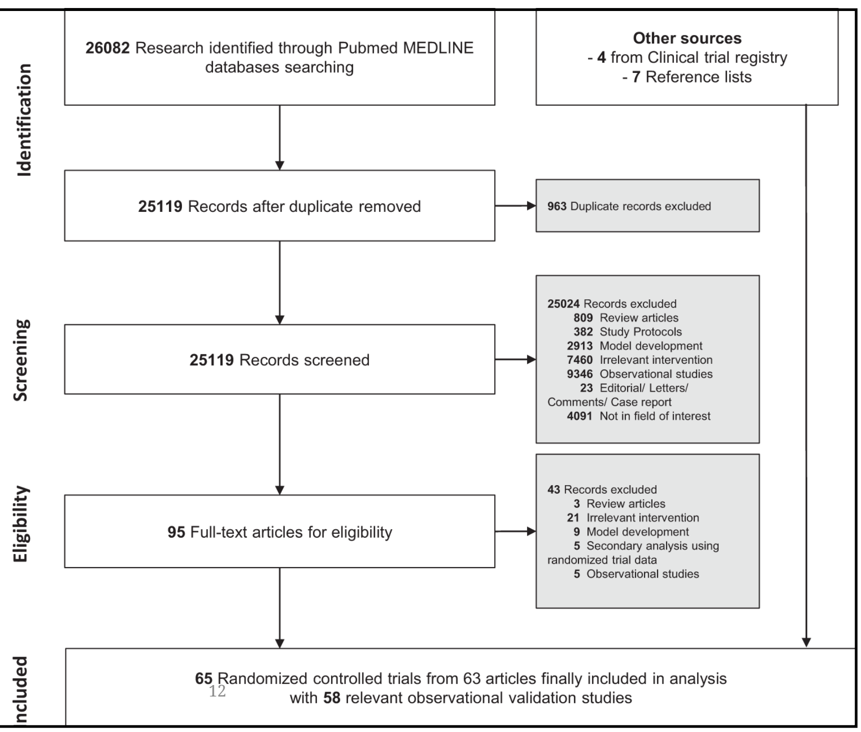
DMICE Conference 3/31/22 10

10

Systematic review of interventions using AI clinical prediction tools (Zhou, 2021)

- Zhou, Q., Chen, Z.-H., Cao, Y.-H., Peng, S., 2021. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. *NPJ Digit Med* 4, 154. https://doi.org/10.1038/s41746-021-00524-2
- Review of all randomized controlled trials (RCTs) using
  - Traditional statistical (TS) – mostly regression
  - Machine learning (ML) – all but deep learning
  - Deep learning (DL) – neural networks
- TS and ML tools focused on assistive treatment decisions, assistive diagnosis, and risk stratification, whereas DL tools only focused on assistive diagnosis

65 RCTs from 26K+ publications – not uncommon



Identified 65 RCTs with following characteristics
- 61.5% positive results
- Variety of disease categories – cancer, other chronic disease, acute disease, and primary care
- Types of algorithms – TS > ML > DL
- Predictive tool function – assistive treatment decisions > assistive diagnosis > risk stratification

Some concerns of bias in studies
- One-third no sample size estimation
- Three-fourths no masking (open-label)
- Majority did not reference CONSORT, use intent-to-treat analysis, or provide study protocol

- Caveat: number of positive studies does not necessarily indicate general superiority of methods

Table 1. General characteristics of the 65 randomized controlled trials.

Variables	Levels	Total (*n* − 65)
Results (%)	Negative	25 (38.5)
	Positive	40 (61.5)
Duration of study (n^a − 59, months, median [IQR])		12 [6, 24]
Sample size (median [IQR])		435 [192, 999]
Sample size estimation (%)	Larger or equal than expected	37 (56.9)
	Less than expected	7 (10.8)
	Not performed	21 (32.3)
Publication year (%)	2010–2015	21 (32.3)
	2016–2020	44 (67.7)
Study design (%)	RCT superiority (individualized)	48 (73.8)
	RCT superiority with crossover (individualized)	1 (1.5)
	RCT non-inferiority (individualized)	2 (3.1)
	Clustered RCT superiority (clustered)	7 (10.8)
	Stepped-wedge design (clustered)	7 (10.8)
Allocation ratio (%)	1:1 parallel	55 (84.6)
	Others	10 (15.4)
Masking (%)	Open-label	49 (75.4)
	Single-blinded	12 (18.5)
	Double-blinded	4 (6.2)
Centers (%)	Single	33 (50.8)
	Multi	32 (49.2)
Disease category (%)	Cancer	11 (16.9)
	Chronic disease not included cancer	18 (27.7)
	Acute disease	19 (29.2)
	Primary care	9 (13.8)
	Others	8 (12.3)
Types of algorithms (%)	Traditional statistical model	37 (56.9)
	Machine learning	17 (26.2)
	Deep learning	11 (16.9)
Prediction tools function (%)	Assistive treatment decision	35 (53.8)
	Assistive diagnosis	16 (24.6)
	Risk stratification	12 (18.5)
	Others	2 (3.1)
Referenced CONSORT (%)	No	47 (72.3)
	Yes	18 (27.7)
Intent-to-treat analysis (%)	No	39 (60.0)
	Yes	26 (40.0)
Study protocol available	No	49 (75.4)
	Yes	16 (24.6)
Model development (%)	No	7 (10.8)
	Yes—independent publication	49 (75.4)
	Yes—published in the same article with RCT	9 (13.8)
Internal validation (%)	No	23 (35.4)
	Yes	42 (64.6)
External validation (%)	No	25 (38.5)
	Yes	40 (61.5)
AUC in model development (n^a − 21, median [IQR])		0.81 [0.75, 0.90]
AUC in internal validation (n^a − 18, median [IQR])		0.78 [0.73, 0.78]
AUC in external validation (n^a − 20, median [IQR])		0.83 [0.79, 0.97]

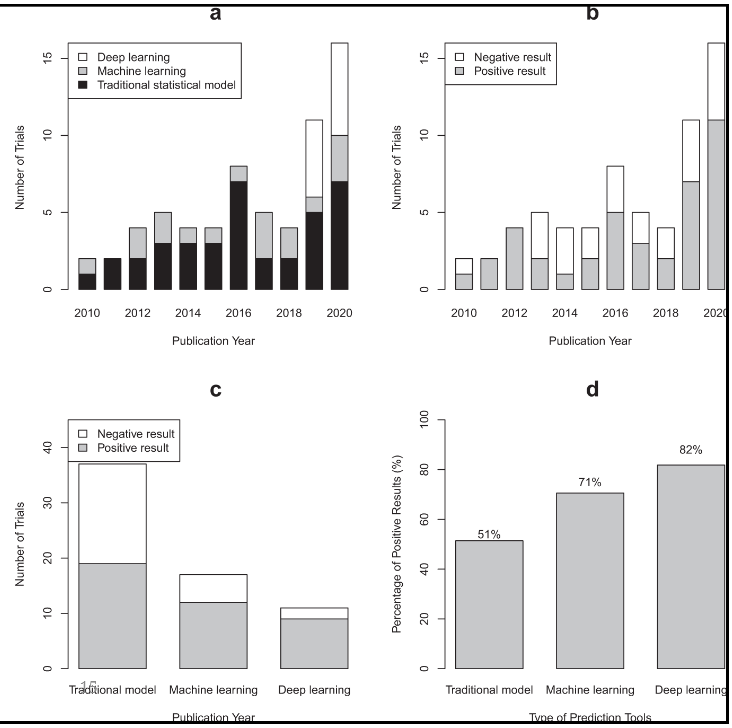
IQR interquartile range, *AUC* area under the receiver operating characteristic curve.
[a]Available numbers used for description

Characteristics by tool type varied
- Model input – clinical quantitative data for TS/ML, images for DL
- Disease category – varied for TS, chronic disease for ML, cancer for DL
- Tool function – risk stratification and treatment for TS, treatment for ML, diagnosis for DL
- Results – mixed for TS, more positive for ML/DL

Variables	Levels	TS (*n* = 37)	ML (*n* = 17)	DL (*n* = 11)	*P* value
Duration of study (*n* = 59, months, median [IQR])		17 [8, 32]	7 [4, 19]	6 [4, 9]	0.005
Sample size (median [IQR])		435 [194, 999]	258 [90, 537]	700 [548, 994]	0.122
Clinical settings (%)	Outpatients	19 (51.4)	6 (35.3)	1 (9.1)	0.015
	Inpatients	17 (45.9)	8 (47.1)	10 (90.9)	
	Home	1 (2.7)	3 (17.6)	0 (0.0)	
Publication year (%)	2010–2015	14 (37.8)	7 (41.2)	0 (0.0)	0.041
	2016–2020	23 (62.2)	10 (58.8)	11 (100.0)	
Model input (%)	Clinical quantitative data	36 (97.3)	16 (94.1)	0 (0.0)	<0.001
	Images or videos	1 (2.7)	0 (0.0)	10 (90.9)	
	Natural language	0 (0.0)	1 (5.9)	1 (9.1)	
Disease category (%)	Cancer	2 (5.4)	0 (0.0)	9 (81.8)	<0.001
	Chronic disease	4 (10.8)	13 (76.5)	1 (9.1)	
	Acute disease	16 (43.2)	2 (11.8)	1 (9.1)	
	Primary care	9 (24.3)	0 (0.0)	0 (0.0)	
	Others	6 (16.2)	2 (11.8)	0 (0.0)	
Prediction tools function (%)	Assistive diagnosis	3 (8.1)	2 (11.8)	11 (100.0)	<0.001
	Risk stratification	11 (29.7)	1 (5.9)	0 (0.0)	
	Assistive treatment decision	22 (59.5)	13 (76.5)	0 (0.0)	
	Others	1 (2.7)	1 (5.9)	0 (0.0)	
Results (%)	Negative	18 (48.6)	5 (29.4)	2 (18.2)	0.136
	Positive	19 (51.4)	12 (70.6)	9 (81.8)	0.044 (*P* for trend)

By publication year
- Increasing per year
- Increasing DL per year

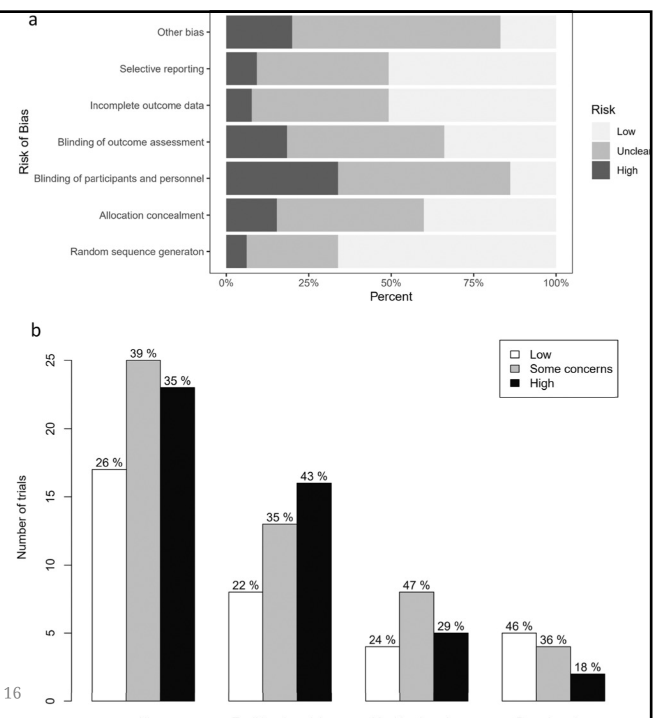
By tool type, more positive for DL > ML > TS



Only 17 of 65 trials with low risk of bias

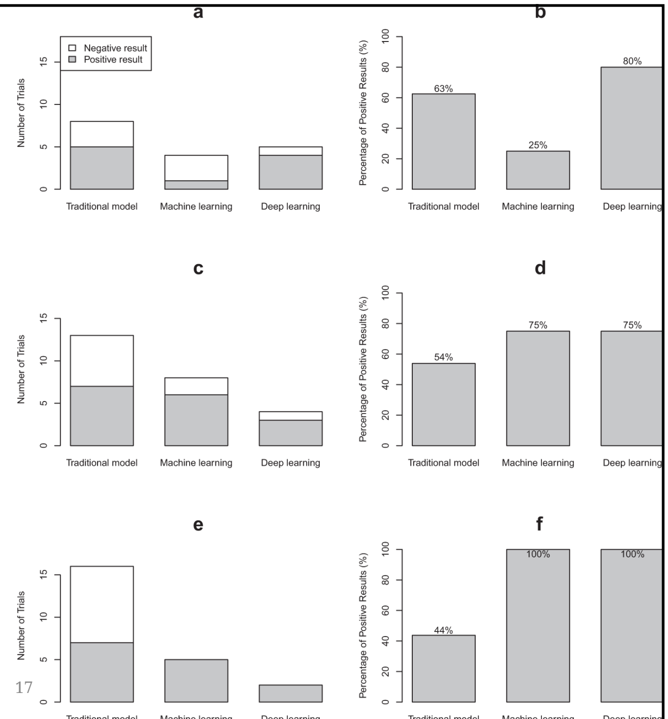
Risk of bias high or unclear for most studies – higher for TS > ML > DL

Suboptimal use of CONSORT, sample size pre-estimation, randomization, and intent-to-treat analysis



Proportion of trials and results for
- Low risk of bias – a-b
- Some concerns – c-d
- High risk of bias – e-f

For low risk of bias trials, positive outcomes in TS 63%, ML 25%, DL 80%



Characteristics of DL trials
- Of 11 RCTs, 9 evaluate assisting endoscopy – all positive results
- 2 other RCTs have negative results

Table 2. Procedures of predictive tool interventions in the eleven randomized controlled trials involving interventions evaluating deeplearning tools

Reference	Conditions	Sample size	Tools for intervention	Control	Algorithms	Tool function	Tool input	Tool output	How the output being used in clinical settings	Trial primary outcomes	Gold standard	Trial findings
Chen 2019	Upper gastrointestinal lesions	437	Routine EGD examination stratified by three types with the assistance of ENDOANGEL AI system	Routine EGD examination stratified by three types without AI	DCNN (VGG-16)	Assistive diagnosis	EGD images	A virtual stomach model monitoring blind spots; timing; scoring and grading	Experts referenced AI output to make EGD examination and monitor blind spots.	Mean blind spot rate	Experts	Positive
Lin 2019	Childhood cataracts	700	CC-cruiser web diagnosis platform	Regular ophthalmic diagnosis	DCNN (ImageNet)	Assistive diagnosis	Ocular images from slit-lamp photography	Diagnosis outcome; comprehensive evaluation; treatment recommendation	AI made diagnosis independently, and its results would be comparted with experts and not impact clinical decision making.	Accuracy of diagnosis	Experts	Negative
Su 2019	Colorectal cancer	659	Routine colonoscopies with the assistance of an AI automatic quality control system	Routine colonoscopies	DCNN (AlexNet, ZFNet, YOLO V2)	Assistive diagnosis	Colonoscopy images	Location of colorectal polyps; timing; reminding retest and clean	Endoscopists referenced AI output to make endoscopic examination and report of polyps and adenomas.	Adenoma detection rate	Pathology	Positive
Wang 2019	Colorectal cancer	1058	Routine colonoscopies with the assistance of an automatic polyp detection system	Routine colonoscopies	Deep learning architecture	Assistive diagnosis	Colonoscopy images	Location of polyps; alarming	Endoscopists were required to check every polyp location detected by the system and report of polyps and adenomas.	Adenoma detection rate	Pathology	Positive
Wu 2019	Upper gastrointestinal lesions	303	Routine EGD examination with the assistance of WISENSE AI system	Routine EGD examination	DCNN (VGG-16 and DenseNet)	Assistive diagnosis	EGD images	A virtual stomach model monitoring blind spots; timing; scoring and grading; extracting frames with the highest confidence	Experts referenced AI output to make EGD examination and monitor blind spots.	Mean blind spot rate	Experts	Positive
Gong 2020	Colorectal cancer	704	ENDOANGEL-assisted routine colonoscopy	Routine colonoscopy	DCNN and perceptual hash algorithms (VGG-16)	Assistive diagnosis	Colonoscopy images	Timing; safe, alarm, and dangerous ranges of withdrawal speed for real-time monitoring; slipping warning	Operating endoscopists referenced AI output to make endoscopic examination and report of polyps and adenomas.	Adenoma detection rate	Pathology	Positive
Liu 2020	Colorectal cancer	1026	Routine colonoscopy with CADe assistance	Routine colonoscopy	DCNN-3D	Assistive diagnosis	Colonoscopy images	The probability of polyps in each frame; lesions alarming	Endoscopists focused mainly on the main monitor during the examination process, and a voice alarm prompted them to view the system monitor to check the location of each polyp detected by the system.	Detection rate of polyps and adenomas	Pathology	Positive
Luo 2020	Colorectal cancer	157	AI-assisted colonoscopy	Traditional colonoscopy	CNN (YOLO)	Assistive diagnosis	Colonoscopy images	Location of polyps	Endoscopists referenced AI output to make endoscopic examination and report of polyps.	Polyp detection rate	Not reported	Positive
Repici 2020	Colorectal cancer	685	High-definition colonoscopies with the AI-based CADe system	Routine colonoscopy	CNN	Assistive diagnosis	Colonoscopy images	Location of polys	Endoscopists referenced AI output to make endoscopic examination and report of polyps and adenomas.	Adenoma detection rate	Pathology	Positive
Wang 2020	Colorectal cancer	962	White light colonoscopy with assistance from the CADe system	White light colonoscopy with assistance from a sham system	Deep learning architecture	Assistive diagnosis	Colonoscopy images	Location of polyps; alarming	Endoscopists were required to check every polyp location detected by the system and report of polyps and adenomas.	Adenoma detection rate	Pathology	Positive
Blomberg 2021	Out-of-hospital cardiac arrest (OHCA)	5242	Normal protocols with alert	Normal protocols without alert	Speech recognition using deep neural networks	Assistive diagnosis	Emergency calls	OHCA Alert	Dispatchers in the intervention group were alerted when the machine learning model identified out-of-hospital cardiac arrest.	The rate of dispatcher recognition of subsequently confirmed OHCA	Danish Cardiac Arrest Registry	Negative

Abbreviations: AI = Artificial intelligence; DL = Tools using deep learning algorithms; ML = Tools using machine learning algorithms; CNN = Convolutionalneural networks; DCNN = Deep convolutional neural networks; CADe = Computer-aided detection; EGD = Esophagogastroduodenoscopy; OHCA = Out-of-hospital cardiac a

Conclusions about review

- AI predictive tools show great promise in improving clinical decisions for diagnosis, treatment, and risk stratification but comprehensive evidence lacking
  - Number of clinical trials assessing clinical benefit is small
  - Majority of the clinical trials have indeterminate or high risk of bias
  - Trials of deep learning methods highly focused on endoscopic procedures
- Concerns about review
  - Missing column in Table 2 of DL interventions
    - Does not include Yao et al. 2021 – published after review done?
  - Difficult to use data in Supp Table 4 of ML interventions
    - Includes Wijnberge et al. 2020 (62) but not in ML table – considered TS?
  - No data/table for TS interventions

Which OHSU department is best poised to lead clinical implementation and evaluation of AI?

- Wrangling
- Modeling

Bioinformatics & Computational Biomedicine

- Clinical implementation
- Evaluation

Health & Clinical Informatics

- Clinical trials
- Systematic reviews

Clinical Epidemiology

Who can lead “translational AI?” (Hersh, 2021)

Conclusions (from the audience)

- How successful has AI been in improving clinical care and patient outcomes?
- Where might AI have the most benefit in the future, near and far?
- How can we operationalize the implementation and evaluation of AI at OHSU?