Machine Learning and Artificial Intelligence (2/3)

Introduction to Biomedical & Health Informatics
William Hersh

Oregon Health & Science University

Machine learning and artificial intelligence

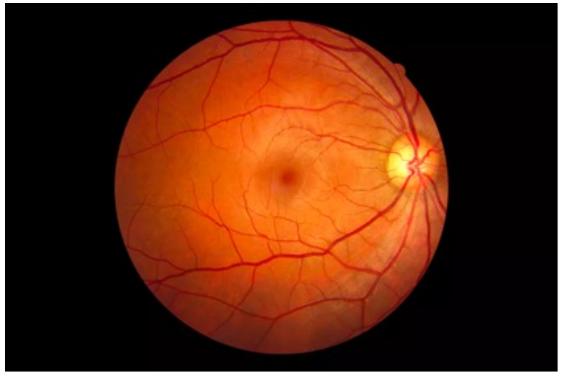
- Overview
- Methods
- Results
- Future directions

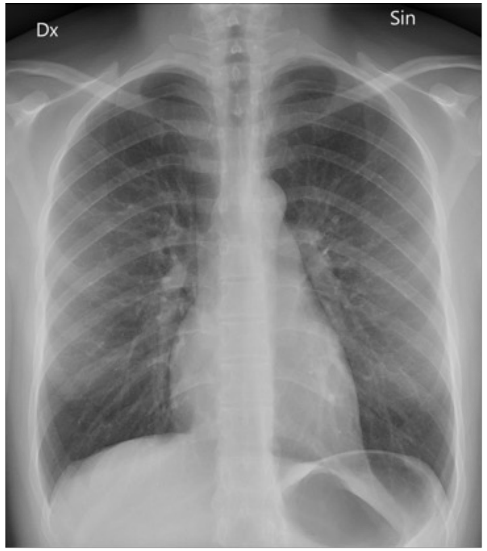
Results: biomedical applications of ML

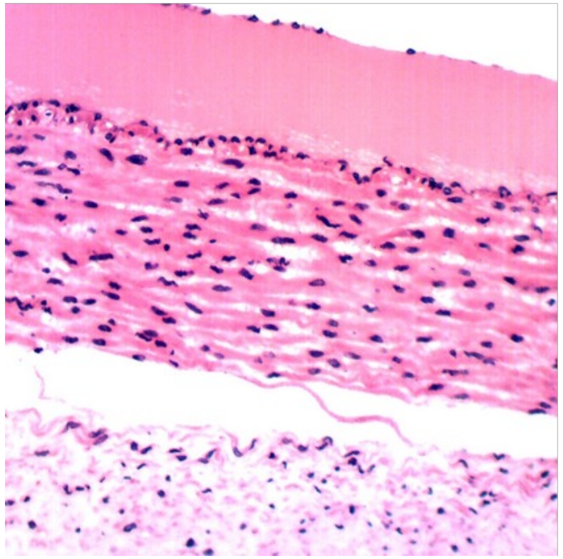
- Specific applications
  - Imaging
  - Clinical prediction
  - Biological processes
  - Assisting humans
- Real-world studies
- Critical appraisal of studies
- Systematic reviews

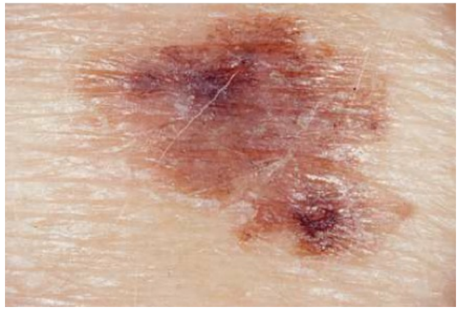
Imaging

- Early studies
  - Diabetic retinopathy (DR) (Gulshan, 2016; Ting, 2017)
  - Histology of cancer (Bejnordi, 2017) and metastases (Veta, 2019)
  - Tuberculosis (Lakhani, 2017) and pneumonia (Rajpurkar, 2018)
  - Skin cancer (Esteva, 2017; Haenssle, 2018; Tschandi, 2019)
- Summarized in systematic review showing deep learning performance approaching or comparable to human experts (Liu, 2019)
- State of the art (Esteva, 2021)







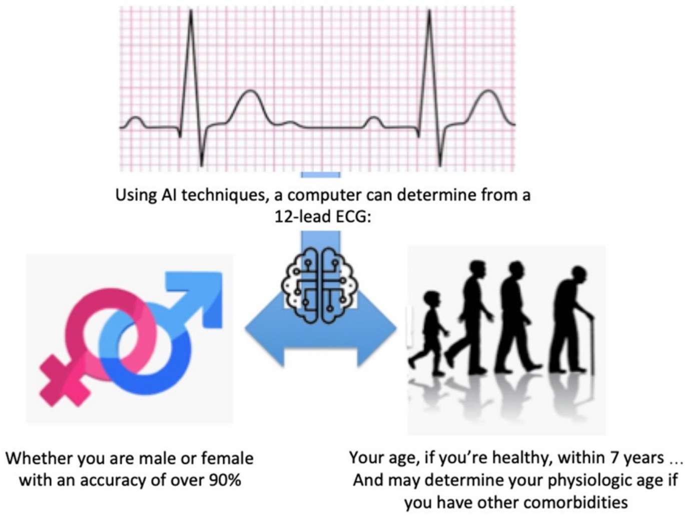


Beyond basic image diagnostic classification

- Clinically acceptable performance in African setting for detecting referable DR, vision-threatening DR, and diabetic macular edema in population-based DR screening (Bellemo, 2019)
- Deep learning predicted cardiovascular disease risks from lung cancer screening low-dose computed tomography (Chao, 2021)
- Gleason grading for prostate cancer comparable to pathologists (Bulten, 2022)
- Cryosectioned images transformed to formalin-fixed and paraffin-embedded views (Ozyoruk, 2022)
- Retinal vasculometry to predict circulatory mortality, myocardial infarction, and stroke (Rudnicka, 2022)
- Detect proximal femoral fractures in emergency department (Oakden-Rayner, 2022)
- Arrhythmic sudden death survival prediction from scarring in heart (Poposecu, 2022)
- Augment efficiency of pathology whole-slide searching (Chen, 2022)
- Self-supervised learning of chest x-ray pathologies (Tiu, 2022)
- Pathology slides predict colorectal cancer biopsy results and outcomes (Wagner, 2023)

Other types of pattern recognition

- Wave forms – use of ECGs
  - Age and sex determination (Attia, 2019)
  - Cardiac arrhythmia detection comparable to cardiologists (Hannun, 2019)
  - Interpretation better than conventional algorithm (Smith, 2019; Hughes, 2021)
  - Detecting hyperkalemia from 2 (of 12) leads (Galloway, 2019)
  - Early diagnosis of low ejection fraction in patients in setting of routine primary care (Yao, 2021)
- Sounds
  - Detecting pathological breath sounds in children with digital stethoscopes (Kevat, 2020; Zhang, 2021)
- Mobile devices
  - Detect anemia from smartphone pictures (Mannino, 2018)



Using AI techniques, a computer can determine from a 12-lead ECG:

Whether you are male or female with an accuracy of over 90%

Your age, if you're healthy, within 7 years … And may determine your physiologic age if you have other comorbidities

Clinical prediction

- Length of stay, mortality, readmission, and diagnosis at two large medical centers (Rajkomar, 2018)
- ML-selected variables outperformed expert-selected variables in predicting patient mortality from coronary artery disease (Steele, 2018)
- Age and sex determination from retinal images (Poplin, 2018)
- Early risk of chronic kidney disease in patients with diabetes (Ravizza, 2019)
- Wide variety of pediatric diagnoses from EHR data at major referral center (Liang, 2019)
- Dementia from EHR data up to two years before clinical diagnosis (Wang, 2019)
- Predict childhood lead poisoning (Potash, 2020)
- Improve accuracy of patient deterioration predictions (Romero-Brufau, 2021)
- Predict need for mechanical ventilation, renal replacement therapy, and readmission in COVID-19 (Rodriguez, 2021)
- Predict coronary artery disease and its outcomes from EHR data (Forrest, 2022)

Biological processes

- Genomics
  - Predicting clinical outcomes from cancer genomic profiles (Yousefi, 2017)
  - Calling gene variants in sequencing data (Poplin, 2018)
  - Identifying facial phenotypes of genetic disorders (Gurovich, 2019)
  - Prioritizing and classifying gene variants in sequencing data (Nicora, 2022)
- Drug discovery
  - Discovery of existing and new drugs effective as antibiotics (Das, 2021)
  - Other discovery of new drugs (Xiong, 2021; Jayatunga, 2022)
- Protein folding prediction
  - Revolutionary success by Google AlphaFold in predicting protein structure from amino acid sequence (Jumper, 2021; Al-Janabi, 2022)
  - Has led to catalog of likely structures of over 200M proteins based on DNA (Varadi, 2022; Travis, 2022)

Assisting humans

- Automatically charting symptoms from patient-physician conversations (Rajkomar, 2019)
- "Weakly supervised" (using clinical diagnoses) interpretation of pathology slides would allow pathologists to exclude 65–75% of slides while retaining 100% sensitivity (Campanella, 2019)
- Learning outlier clinical alerts to reduce drug prescribing errors and adverse events (Segal, 2019)
  - 85% confirmed clinically valid, 80% considered clinically useful
  - Alert burden low – 0.4% of all medication orders
- Assisting dermatologists improved accuracy but poor ML worsened human performance (Tschandl, 2020)
- Commercially available AI algorithm assessed screening mammograms with sufficient diagnostic performance to be further evaluated as an independent reader in prospective clinical trials (Salim, 2020)
  - Combining first readers with best algorithm identified more cases positive for cancer than combining first readers with second readers

Assisting humans (cont.)

- Aiding radiologists
  - In breast ultrasound, reduced false-positive rates by 37.3% and requested biopsies by 27.8% while maintaining same level of sensitivity (Shen, 2021)
  - In interpreting CXRs, increased sensitivity for junior radiologists and specificity for senior radiologists (Homayounieh, 2021)
  - In fracture assessment, improved sensitivity without increasing reading time (Guermazi, 2022)
- AI system helped physicians extract relevant patient information in a shorter time while maintaining high accuracy (Chi, 2021)
- AI assistance associated with improved dermatology diagnoses by PCPs and NPs for 1 in every 8 to 10 cases (Jain, 2021)
- Identify features in CDS medication alerts to reduce volume by half while still maintaining 99% sensitivity (Liu, 2022)
- Feedback by AI tutoring system led to better surgical training for medical students than virtual expert instruction (Fazlollahi, 2022)

Real-world studies

- Eye diseases
  - Diagnosis and treatment decisions for congenital cataracts
    - High accuracy for diagnosis (98%), risk stratification (93-100%), and treatment suggestions (93%) (Long, 2017)
    - Accuracy for diagnosis and treatment determination were 87.4% and 70.8%, which were significantly lower than 99.1% and 96.7% than senior consultants but took less time (2.79 min vs. 8.53 min) (Lin, 2019)
  - Detect previously undiagnosed DR at primary care clinics (Abràmoff, 2018)
    - Sensitivity 87.2%, specificity 90.7%, imageability rate 96.1%
  - Use for DR in rural India (Gulshan, 2019)
    - Sensitivity 88.9%, specificity 92.2%, comparable to manual grading
  - Use for DR in smartphone (Natarajan, 2019)
    - Images from 18 of 231 were deemed ungradable
    - For rest, sensitivity and specificity of referable DR were 100.0% and 88.4%

Real-world studies (cont.)

- In pathology
  - Algorithm-assisted pathologists demonstrated higher accuracy than either the deep learning algorithm or pathologist alone (Steiner, 2018)
    - Assistance significantly increased sensitivity of detection for micrometastases (91% vs. 83% alone)
    - Reduced time compared to pathologist alone for positive (61 vs. 116 sec) and negative images (111 vs. 137 sec)
  - In simulated study, decisions with AI decision-aid improved performance, even when it was not used (Meyer, 2022)
- Mixed results in GI endoscopy
  - Predicted pathology of detected diminutive colonic polyps (≤5 mm) on basis of real-time comparison with pathologic diagnosis of resected specimen (gold standard) to “detect and leave” (Mori, 2018)
    - Negative predictive value 94%
  - Colonic adenoma detection rate improved from 20-30% to 50%, although additional polyps mostly small and benign (Wang, 2019)
  - ML system better able to detect blind spots in upper endoscopy (EGD) than human endoscopists (Wu, 2019)
  - Improved adenoma detection in screening colonoscopy without increased resection of non-neoplastic polyps (Shaukat, 2022)
  - In multi-country study, no improved diagnosis of colonic polyps (Barua, 2022)
  - Improved performance later in half-day sessions when human performance degrades (Lu, 2023)

Real-world studies (cont.)

- Sepsis surveillance reduced in-hospital mortality and length of stay (Shimabukuro, 2017; Adams, 2022)
- Prospective validation of predicting 180-day mortality in outpatients with cancer (Manz, 2020), shown to improve number of serious illness conversations with patients (Manz, 2020)
- Low agreement among clinicians and ML system in outpatient triage decisions (Entezarjou, 2020)
- Reduced readmissions at a community hospital (Romero-Brufau, 2020)
- Improved diagnosis of COVID-19 in chest x-rays (Rangarajan, 2021)
- In patients undergoing health checkups, RCT of AI assisting chest x-ray interpretation found improved detection of actionable nodules (Nam, 2023)

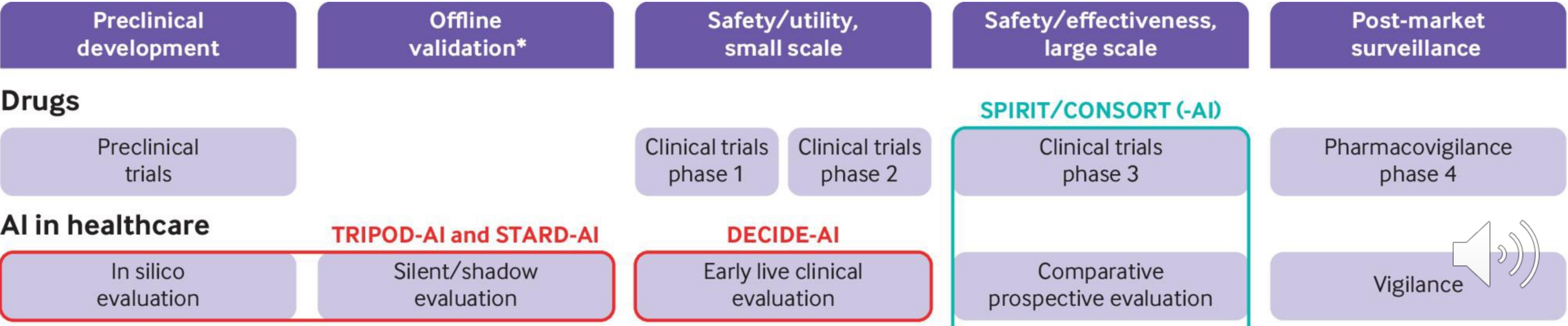
Critical appraisal of ML/AI studies

- As with all evidence-based medicine (EBM), major questions that guide patient care to ask about ML/AI include
  - Diagnosis – “test” for a “disease”
    - Ideally from comparison with ”gold standard”
    - Predictive models are type of diagnostic test
  - Treatment – therapy or intervention to prevent or treat disease; ideally from randomized controlled trial (RCT)
- Studies of both questions can be aggregated into systematic reviews that may (if data allow) be combined via meta-analysis
- Much detail to follow present as reference – keep focus on big picture

Instruments for assessing reporting of studies adapted for AI (Ibrahim, 2021)

- Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) (adapted for AI by Collins, 2019)
- Standards for Reporting of Diagnostic Accuracy Study (STARD) (adapted for AI by Sounderajah, 2021)
- Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence (DECIDE-AI) (Vasey, 2022)
- Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) (adapted for AI by Rivera, 2020)
- Consolidated Standards of Reporting Trials (CONSORT) (adapted for AI by Liu, 2020)

- Prediction model Risk Of Bias ASsessment Tool (PROBAST) (Wolff, 2019)
- Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (adapted for AI by Cacciamani, 2023)



Preclinical development

Offline validation*

Safety/utility, small scale

Safety/effectiveness, large scale

Post-market surveillance

Drugs

Preclinical trials

Clinical trials phase 1

Clinical trials phase 2

SPIRIT/CONSORT (-AI)

Clinical trials phase 3

Pharmacovigilance phase 4

AI in healthcare

TRIPOD-AI and STARD-AI

In silico evaluation

Silent/shadow evaluation

DECIDE-AI

Early live clinical evaluation

Comparative prospective evaluation

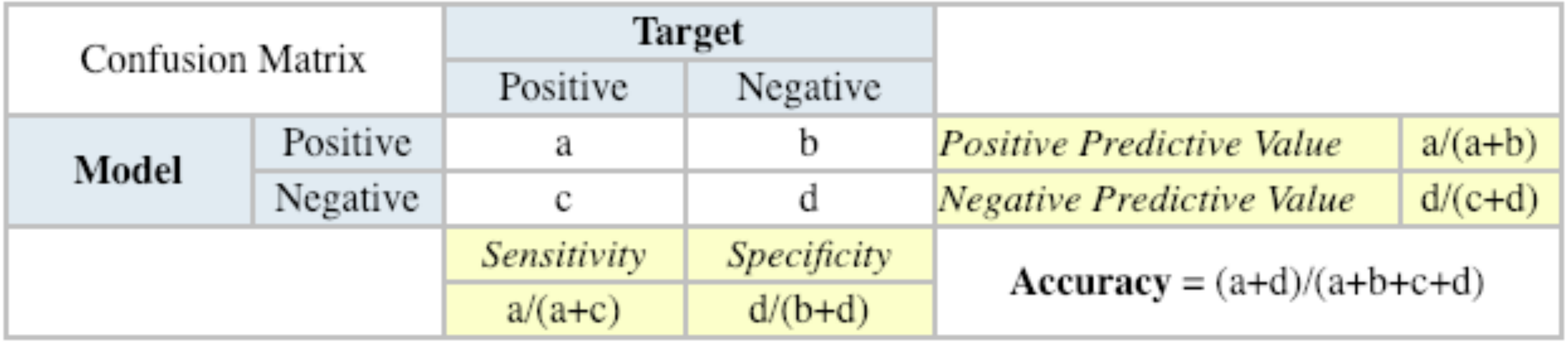
Vigilance

Critical appraisal of ML (and other predictive) models

- Validity of any study of diagnosis (Straus, 2018)
  - Was there an independent, blind comparison with a reference standard?
  - Did the patient sample include an appropriate spectrum of patients to whom the diagnostic test will be applied in clinical practice?
  - Was there a completely independent validation set?
  - Did the results of the test being evaluated influence the decision to perform the reference standard?
  - Were the methods for performing the test described in sufficient detail to permit replication?
- Specific to ML and AI, additional questions about data, model, and evaluation (Liu, 2019)
  - Other frameworks (Vollmer, 2020; Hernandez-Boussard, 2020; Vinny, 2021; van Smeden, 2022)

Key considerations for predictive model evaluation (Singh, 2019)

- To avoid overfitting of model, data sets should include
  - Training set to build predictive model
  - (Optional) validation set to assess performance and fine-tune model’s parameters to selecting best-performing model
  - Test set to assess likely future performance of model on unseen data
- When only limited amount of data is available, use k-fold cross-validation
  - In the k-fold cross-validation, divide the data into k subsets of equal size and each time leave out one of subsets from training to use in test set
  - If k equals sample size, this is “leave-one-out” method



Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

Additional questions about ML models (Liu, 2019)

- Data
  - Input – are these data easily obtainable in routine clinical workflows?
  - Output – is the prediction clinically relevant? Is this prediction at the right granularity, such as a meaningful number of patient risk categories?
  - Label – is this an accepted grading scale and is the reference standard reliable?
  - Patient population – what are the patient population and inclusion/exclusion criteria? What are the numbers of patients with each label and other characteristics?
  - Development/validation split – do any patients appear in both development and validation sets? Is this a strong study design with respect to evaluation of generalization?
  - Amount of data in the development set – is this sufficient to develop the ML model, given its complexity?
  - Amount of data in validation set – is this sufficient to have confidence in the generalizability of the results and any clinically important subgroups?
- Machine learning
  - Method – is this a standard or customized method? What are the parameters and how many are there?
  - Training process - Was transfer learning (e.g., pre-initialization), or multi-task learning (e.g., multiple predictions) used to help training?
  - Data augmentation – data augmentation typically helps performance and generalization. Was this done and appropriate for this data type?
  - Hyperparameters that were optimized – what were the hyperparameters?
  - Use of tuning set – was a separate tuning set (independent of the final validation set) used for hyperparameter tuning?
  - Time taken to apply model per data-point – is this amount of time feasible in the context of routine clinical workflows?

Additional questions about ML models (cont.)

- Evaluation
  - Dataset inclusion/exclusion criteria – were any data excluded based on ML predictions? If so, why and does this skew the results?
  - Performance metric – is this standard?
  - What is “random” performance (e.g., 0 or 0.5), and what is perfect (e.g., 1)? Is this appropriate given the incidence or prevalence of the predicted label? How were the operating points selected?
  - Was this an independent validation set? – was this final validation set used to make any ML model development decisions? Including but not limited to hyperparameter tuning, neural network “checkpoint” selection, method selection, etc.
  - Human comparison metric – is this a fair comparison: e.g., does the human grader have sufficient training (e.g., years post residency), information (e.g., other clinical variables) and time (e.g., comparable with routine practice)? Are the statistics (confidence intervals etc.) present and appropriate?
  - Human comparator – were there any deviations from standard practice, e.g., were the human graders provided sufficient time and with the full-resolution image?
  - Human vs ML model performance – based on prior literature, is the human performance sensible? Do they over- or under-represent human performance?
  - Performance gap in development – validation – What was the gap between tuning and validation, and does this suggest good generalizability?
  - Subgroup/sensitivity analysis – are there any potential confounding factors that should be examined more closely (e.g., image capturing device manufacturer or model, etc.)?

Critical appraisal of ML (and other) RCTs

- Validity of any type of RCT (Straus, 2018)
  - Did experimental and control groups begin study with similar prognosis?
    - Were patients randomized?
    - Was randomization concealed (blinded or masked)?
    - Were patients analyzed in groups to which they were randomized?
    - Were patients in treatment and control groups similar with respect to known prognosis?
  - Did experimental and control groups retain a similar prognosis after the study started?
    - Were patients aware of group allocation?
    - Were clinicians aware of group allocation?
    - Were assessors aware of group allocation?
    - Was follow-up complete?
- Risk of bias tool from Cochrane Collaboration indicates risk that RCT methods might lead to biased results (Sterne, 2019)

Additional questions to ask about RCTs of ML/AI interventions (Liu, 2020)

Title and abstract	• Indicate that the intervention involves AI/ML in the title and/or abstract and specify the type of model • State the intended use of the AI intervention within the trial in the title and/or abstract
Background and objectives	• Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (e.g., healthcare professionals, patients, public)
Eligibility criteria for participants	• State the inclusion and exclusion criteria at the level of participants • State the inclusion and exclusion criteria at the level of the input data
Settings and locations where data collected	• Describe how the AI intervention was integrated into the trial setting, including any onsite or offsite requirements
Interventions for each group with sufficient details to allow replication, including how and when administered	• State which version of the AI algorithm was used • Describe how the input data were acquired and selected for the AI intervention • Describe how poor quality or unavailable input data were assessed and handled • Specify whether there was human-AI interaction in the handling of the input data, and what level of expertise was required of users • Specify the output of the AI intervention • Explain how the AI intervention's outputs contributed to decision-making or other elements of clinical practice
Harms	• Describe results of any analysis of performance errors and how errors were identified, where applicable – if no such analysis was planned or done, justify why not
Funding	• State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use

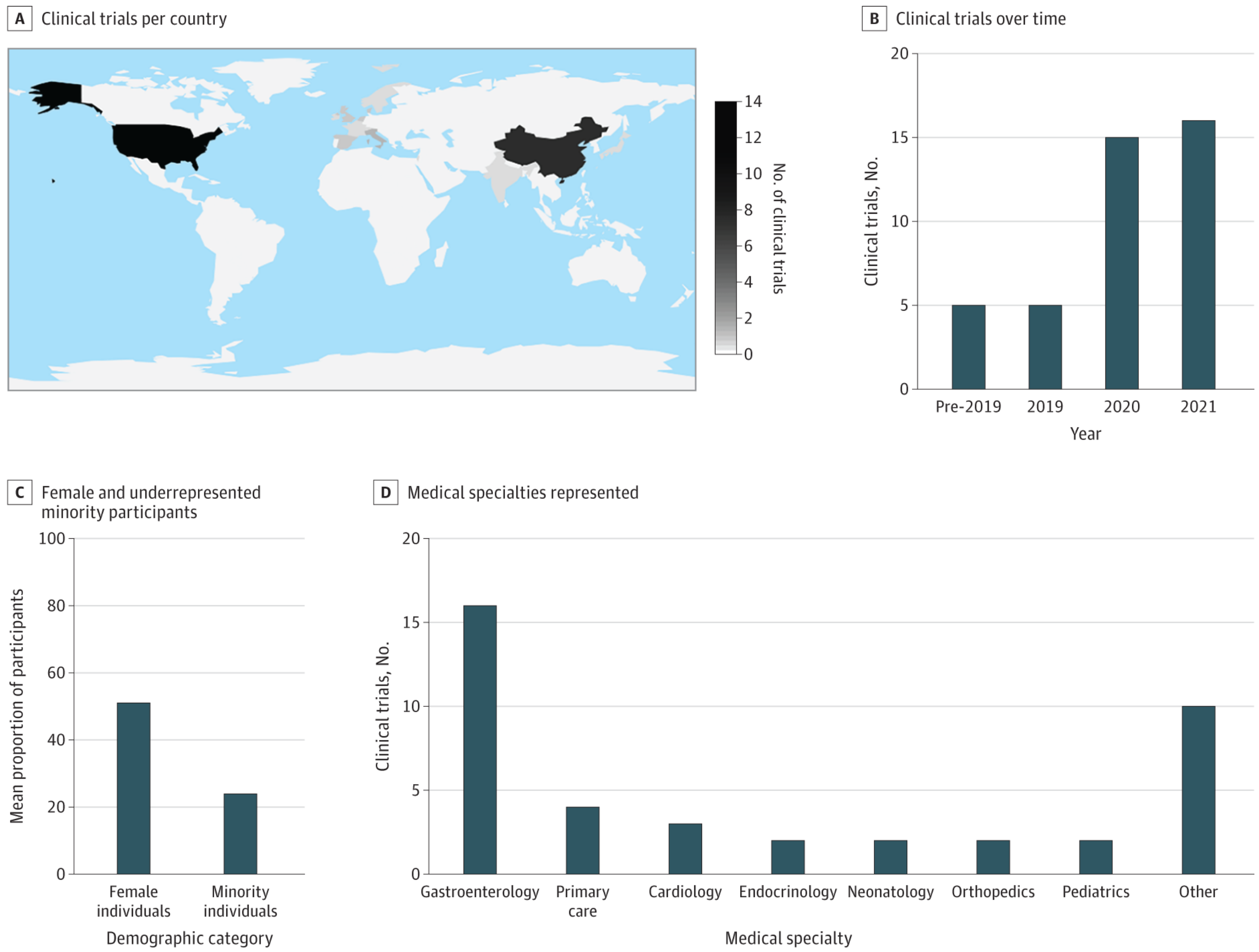
Systematic review of ML/AI models and clinical applications

- From pubmed.gov
  - Tens of thousands of studies applying ML or AI
  - Hundreds of systematic reviews of ML and AI studies – mostly of models applied to clinical topics
- How many studies assessing ML/AI interventions using gold standard, RCT?
  - Collated in systematic reviews – earlier ones did not assess as rigorously, e.g., (Zhou, 2021; Lam, 2022)
  - Most recent and rigorous (Plana, 2022)

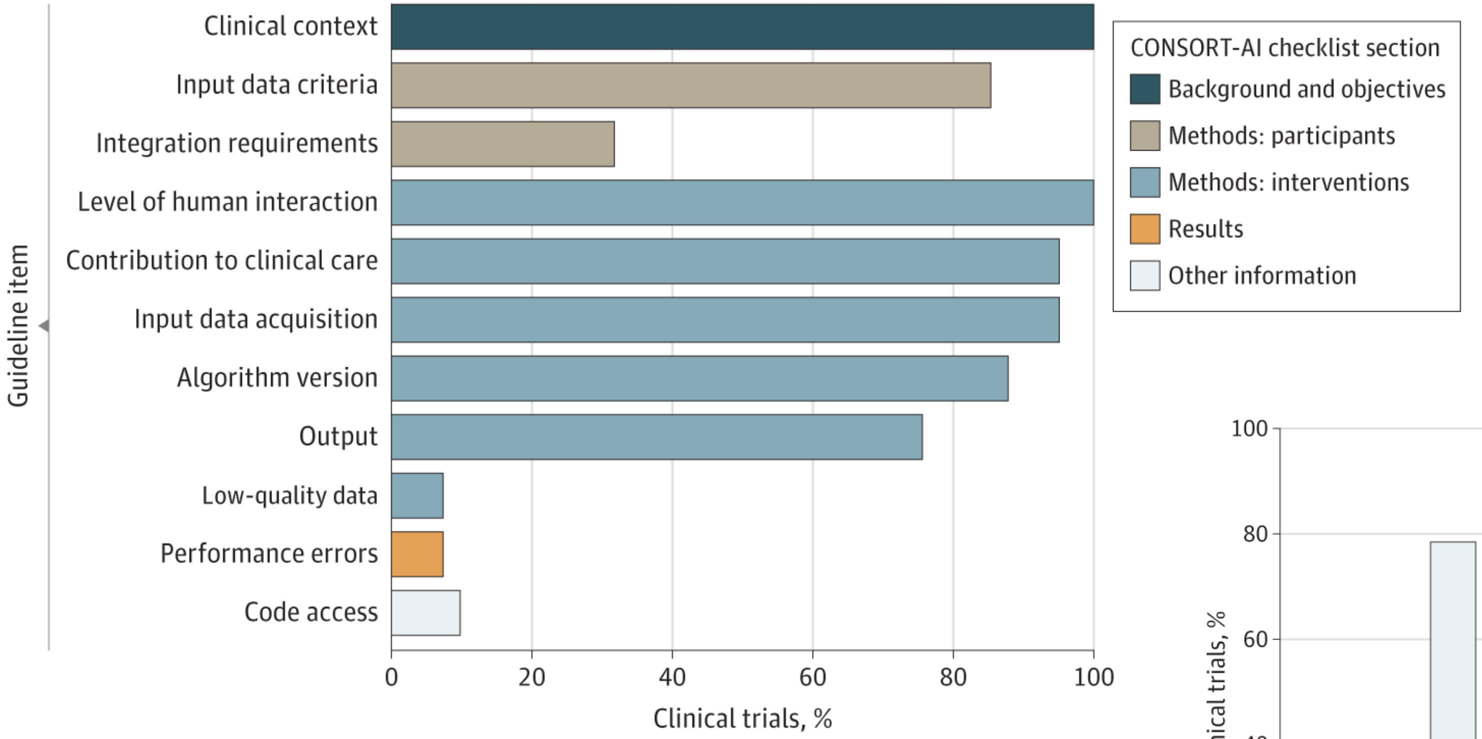
Systematic review of RCTs of ML/AI in healthcare (Plana, 2022)

- Exhaustive search of literature databases to identify RCTs of ML/AI interventions through October, 2021
- Excluded studies of non-RCT design, absence of original data, and evaluation of nonclinical interventions
- Identified 41 RCTs for further analysis
- Analyzed RCT characteristics, including primary intervention, demographics, adherence to CONSORT-AI reporting guideline, and Cochrane risk of bias

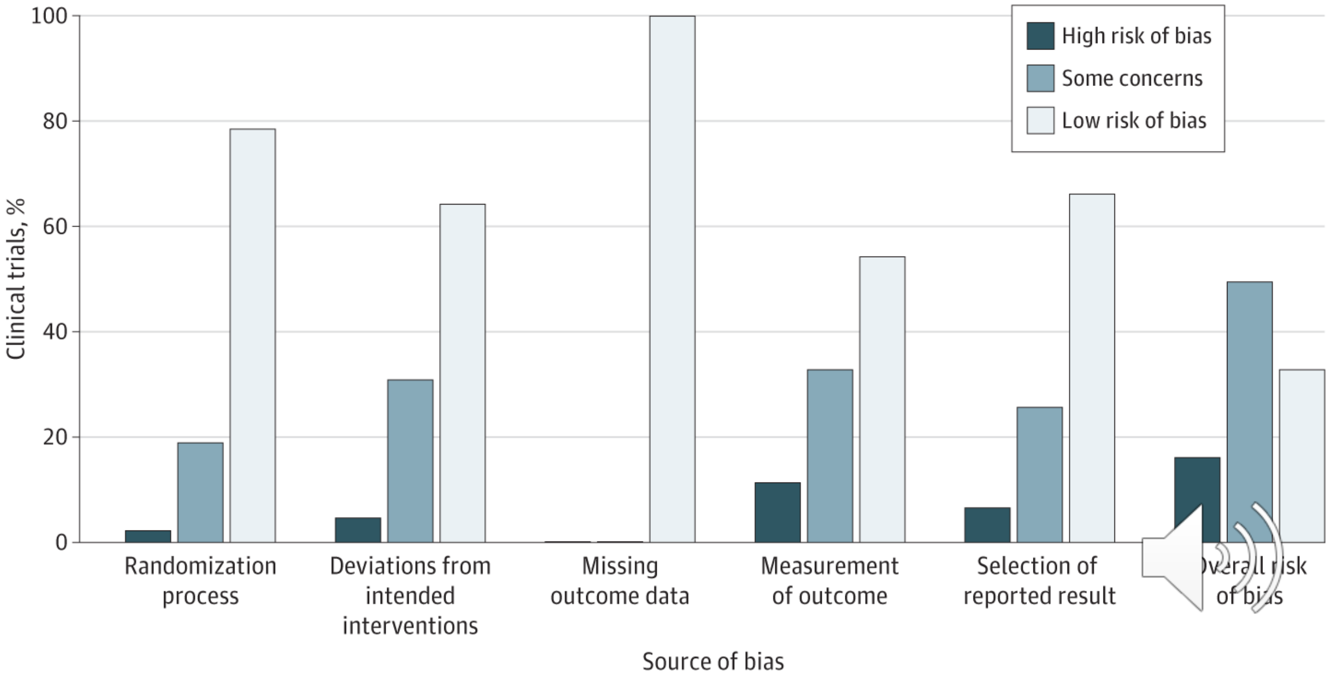
Where, when, who, and what (Plana, 2022)



Adherence to reporting guidelines and risk of bias (Plana, 2022)



- No assessment of study outcomes, unlike Zhou (2021) showing about 60% positive
- No attempted meta-analysis


High risk of bias
Some concerns
Low risk of bias
Clinical trials, %
Randomization process
Deviations from intended interventions
Missing outcome data
Measurement of outcome
Selection of reported result
Overall risk of bias
Source of bias