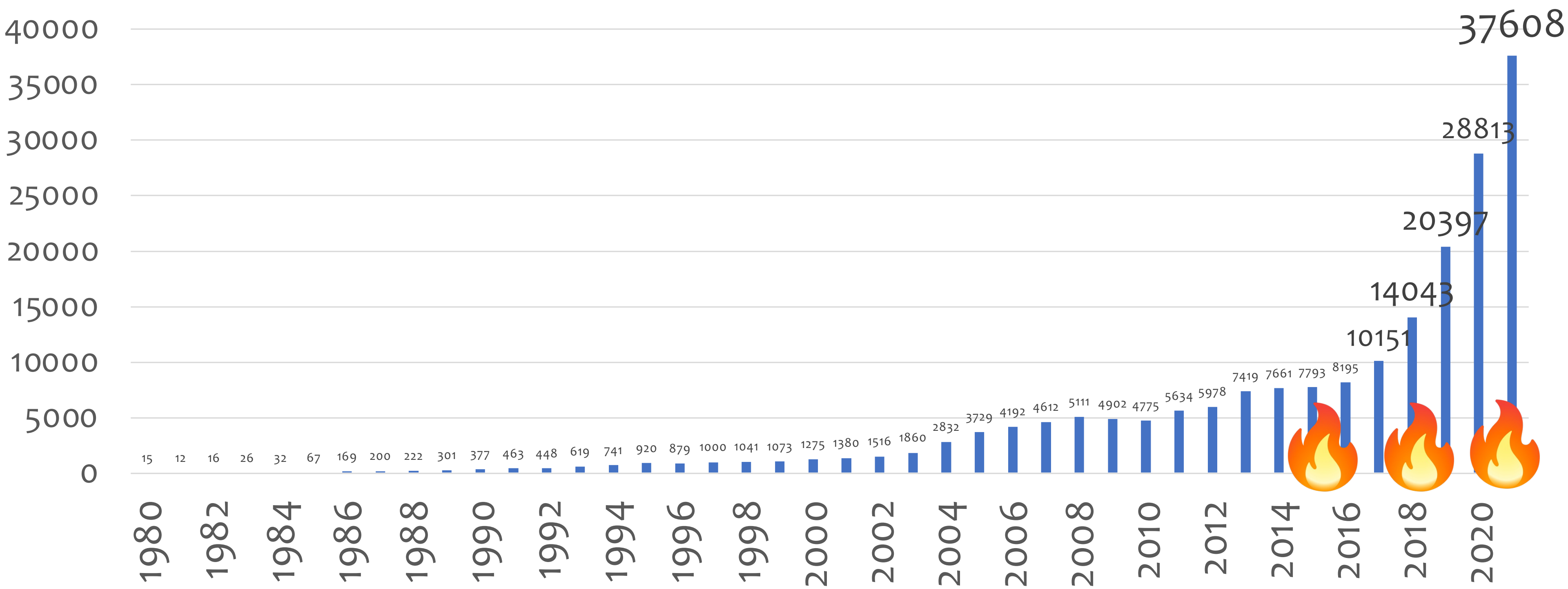
Critical Appraisal of Applying Machine Learning in Healthcare

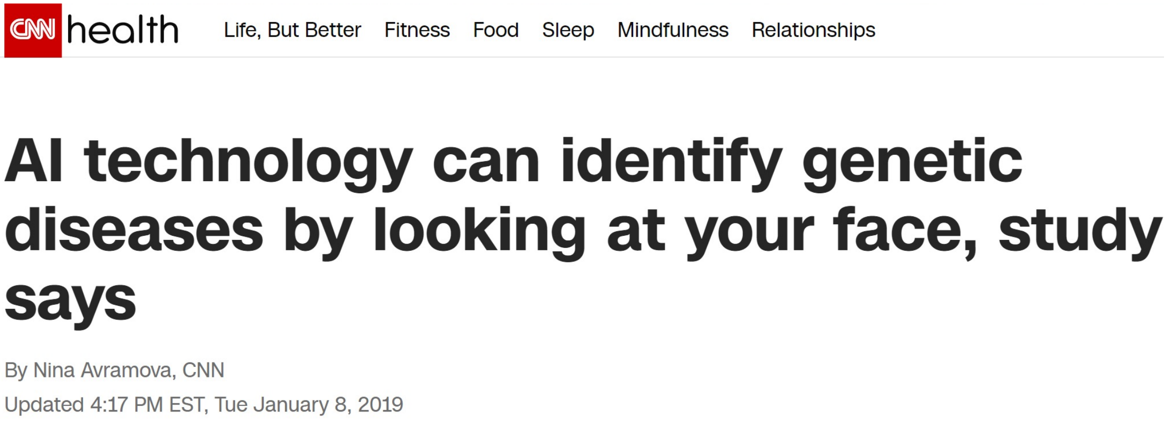
6.871/HST.956: Machine Learning for Healthcare

March 29, 2022

Dr. Madhur Nayan

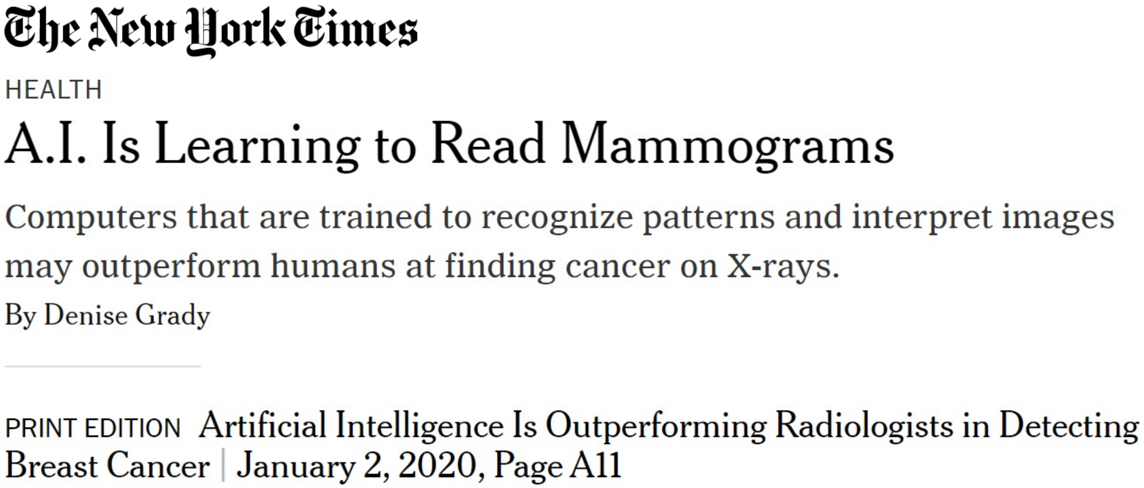
Number of Pubmed citations on Artificial Intelligence or Machine Learning





AI technology can identify genetic diseases by looking at your face, study says

By Nina Avramova, CNN
Updated 4:17 PM EST, Tue January 8, 2019


The New York Times

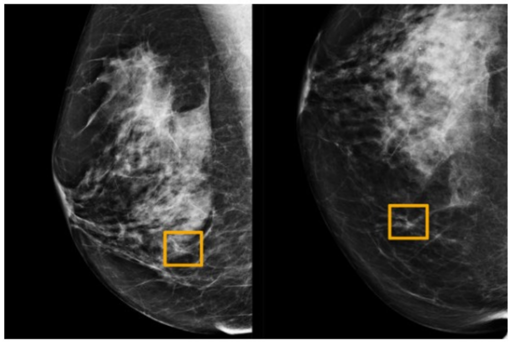
HEALTH

A.I. Is Learning to Read Mammograms

Computers that are trained to recognize patterns and interpret images may outperform humans at finding cancer on X-rays.

By Denise Grady

PRINT EDITION Artificial Intelligence Is Outperforming Radiologists in Detecting Breast Cancer | January 2, 2020, Page A11



https://www.cnn.com/2019/01/08/health/ai-technology-to-identify-genetic-disorder-from-facial-image-intl/index.html
https://www.nytimes.com/2020/01/01/health/breast-cancer-mammogram-artificial-intelligence.html

Application of ML in healthcare

- Despite the rapid proliferation of ML in healthcare research, very little, if any, is currently applied in healthcare
    - Why?

Today’s Outline

- Critical appraisal
- Case 1: Early prediction of Sepsis
- Case 2: Diagnosis of COVID-19 with imaging
- Case 3: Detection of diabetic retinopathy

Today's Outline

- **Critical appraisal**
- Case 1: Early prediction of Sepsis
- Case 2: Diagnosis of COVID-19 with imaging
- Case 3: Detection of diabetic retinopathy

What is Critical Appraisal?

- An analytical framework to evaluate the **quality** and **utility** of a research study
  - Quality relates to methods
  - Utility relates to clinical application

Burls, Amanda. *What is critical appraisal?*. Hayward Medical Communications, 2014.

Reporting Guidelines in Health Research

- Reporting guidelines provide a minimum list of information needed to ensure a manuscript can be:
  - Understood by a reader
  - Replicated by a researcher
  - Used by a physician to make a clinical decision

https://www.equator-network.org/about-us/what-is-a-reporting-guideline/

equator network

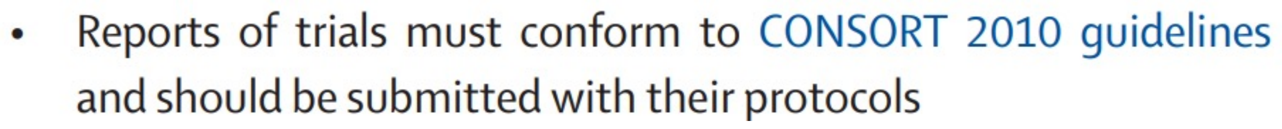
Enhancing the QUAlity and Transparency Of health Research

- **The international ‘standard bearer’ for reporting guidelines**
- Committed to improving ‘the reliability and value of published health research literature by promoting transparent and accurate reporting and wider use of robust reporting guidelines’.

https://www.equator-network.org/about-us/

Reporting Guidelines by Study Type

Study type	Reporting Guidelines
Randomized trials	CONSORT



- Reports of trials must conform to CONSORT 2010 guidelines and should be submitted with their protocols

- **Other reporting guidelines**
  - CLAIM (Checklist for Artificial Intelligence in Medical Imaging)
  - PROBAST-AI (Prediction model Risk Of Bias ASsessment Tool-AI)
  - Etc.

Reading Responses

- What type of data is used and what is the source of the data
- Definition of what the outcome is
- How missing data was handled
- Model architecture choices and explanations for the choices (e.g. if there is a custom optimization loss function)
- Evaluation metrics
- Source code

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD)-AI

1. Title
2. Abstract
3. Introduction
4. Methods
   - Source of data
   - Participants
   - Data preparation
   - Outcome/labelling
   - Predictors
   - Sample size
   - Missing data
   - Analytical methods
   - Risk groups
   - Model development vs. validation
   - Software
5. Results
   - Participants
   - Model development
   - Model specification
   - Model performance
   - Model updating
   - Usability of the model
   - Sensitivity analysis
6. Discussion
7. Other

https://osf.io/nskme/

Today's Outline

- Critical appraisal
- **Case 1: Early prediction of Sepsis**
- Case 2: Diagnosis of COVID-19 with imaging
- Case 3: Detection of diabetic retinopathy

Definition of Sepsis

- **Definition has changed over time.**
  - 1991 consensus definition
    - SIRS + known or suspected infection
      - Definition of Systemic Inflammatory Response Syndrome (SIRS)
        - ≥2 or more of the following:
          1. **Temperature** >38 °C or <36 °C
          2. **Heart rate** >90 beats per minute
          3. **Respiratory rate** > 20 breaths per minute
          4. **Arterial carbon dioxide** < 32 mm Hg
          5. **White blood cell count** (>12,000/µL or <4000/mL or >10%immature band forms

Singer, Mervyn, et al. "The third international consensus definitions for sepsis and septic shock (Sepsis-3)." *Jama* 315.8 (2016): 801-810.

Definition of Sepsis

- **Definition has changed over time.**
  - 2016 consensus definition
    - Life-threatening **organ dysfunction** caused by a dysregulated host **response to infection**
      - Organ dysfunction: ≥2 increase in baseline Sequential Organ Failure Assessment (**SOFA)** score

Singer, Mervyn, et al. "The third international consensus definitions for sepsis and septic shock (Sepsis-3)." *Jama* 315.8 (2016): 801-810.

SOFA score

1. Respiration
2. Platelets
3. Bilirubin
4. Blood pressure
5. Glasgow Coma score
6. Creatinine

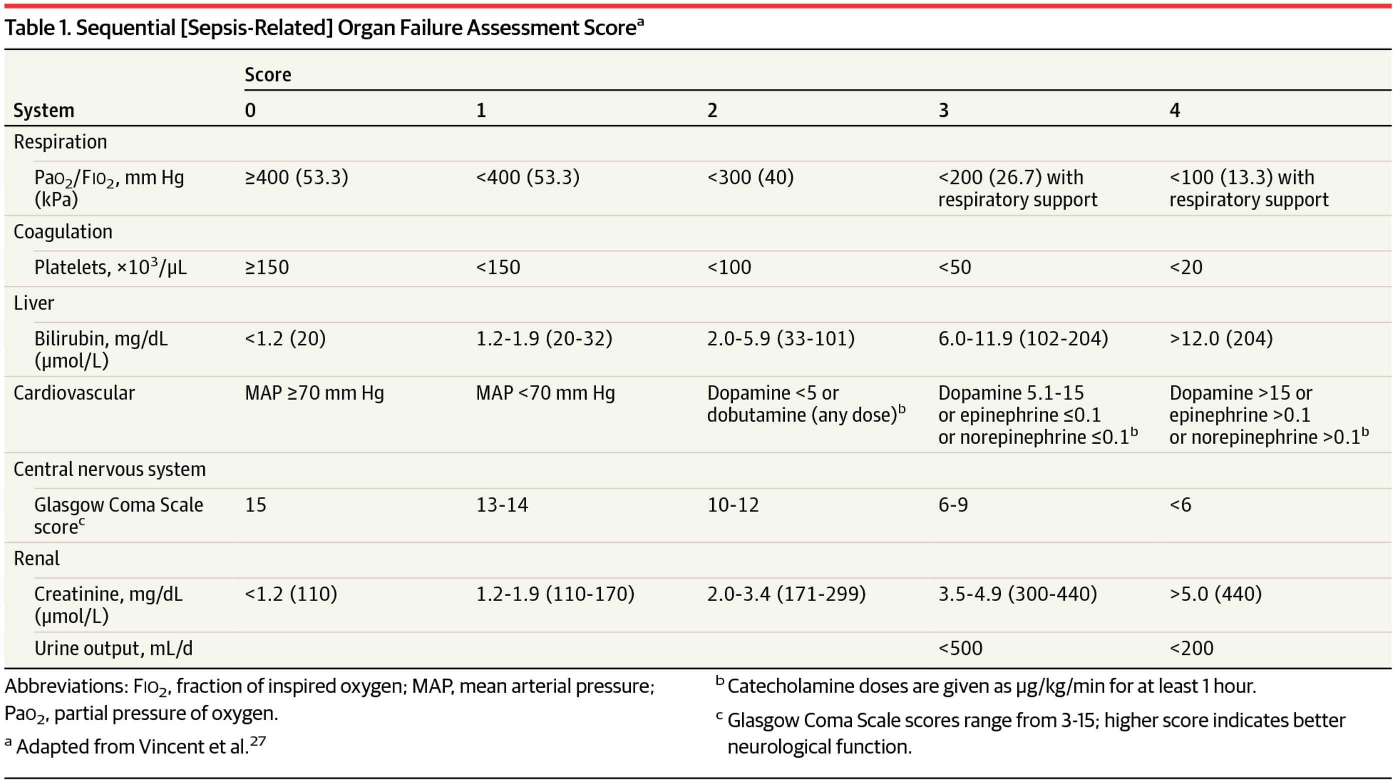


Table 1. Sequential [Sepsis-Related] Organ Failure Assessment Score[a]

	Score				
System	0	1	2	3	4
Respiration					
Pao$_2$/Fio$_2$, mm Hg (kPa)	≥400 (53.3)	<400 (53.3)	<300 (40)	<200 (26.7) with respiratory support	<100 (13.3) with respiratory support
Coagulation					
Platelets, ×10^3/μL	≥150	<150	<100	<50	<20
Liver					
Bilirubin, mg/dL (μmol/L)	<1.2 (20)	1.2-1.9 (20-32)	2.0-5.9 (33-101)	6.0-11.9 (102-204)	>12.0 (204)
Cardiovascular	MAP ≥70 mm Hg	MAP <70 mm Hg	Dopamine <5 or dobutamine (any dose)[b]	Dopamine 5.1-15 or epinephrine ≤0.1 or norepinephrine ≤0.1[b]	Dopamine >15 or epinephrine >0.1 or norepinephrine >0.1[b]
Central nervous system					
Glasgow Coma Scale score[c]	15	13-14	10-12	6-9	<6
Renal					
Creatinine, mg/dL (μmol/L)	<1.2 (110)	1.2-1.9 (110-170)	2.0-3.4 (171-299)	3.5-4.9 (300-440)	>5.0 (440)
Urine output, mL/d				<500	<200

Abbreviations: Fio$_2$, fraction of inspired oxygen; MAP, mean arterial pressure; Pao$_2$, partial pressure of oxygen.

[a] Adapted from Vincent et al.[27]

[b] Catecholamine doses are given as μg/kg/min for at least 1 hour.

[c] Glasgow Coma Scale scores range from 3-15; higher score indicates better neurological function.

Singer, Mervyn, et al. "The third international consensus definitions for sepsis and septic shock (Sepsis-3)." *Jama* 315.8 (2016): 801-810.

Why is Sepsis an Important Problem?

2017 estimated worldwide incidence:

49 million

2017 estimated worldwide incidence:

11 million

Sepsis represents ≈20% of global deaths

2011 estimated total US hospital costs:

20 billion

Most expensive condition treated in US hospitals

Weiss, Audrey J., and Anne Elixhauser. "Overview of hospital stays in the United States, 2012: statistical brief# 180." (2014).
Rudd, Kristina E., et al. "Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study." *The Lancet* 395.10219 (2020): 200-211.

Diagnosing Sepsis

- **Early identification of sepsis risk may result in** earlier treatment, resulting in **improved outcomes.**
    - What outcomes would you consider meaningful?
- **Problem: current sepsis risk detection methods perform modestly**
- **Potential solution:** Electronic health record (EHR) data are becoming generally more widely available, and represent a rich if complex data source that can be applied to the prediction and detection of sepsis

InSight®
by dascena

- "A **clinical decision support (CDS)** software tool that leverages readily-available **data in the Electronic Health Record** system to **help clinicians identify sepsis earlier."**
- "Built with advanced **machine learning** capabilities, InSight can identify patterns to predict the risk of sepsis onset more accurately than rules-based tools."

https://www.dascena.com/insight

InSight

- Calvert et al. 2016
  - Objective: Evaluate the sensitivity and specificity of the InSight algorithm in the prediction of sepsis
  - **TRIPOD-AI**
    - **Methods**
      - **Sources of Data**
        - MIMIC II, a database composed of anonymized clinical documentation from approximately 32,000 patients at the Beth Israel Deaconess Medical Center (BIDMC) collected between 2001 and 2008.

✓

Calvert, Jacob S., et al. "A computational approach to early sepsis detection." *Computers in biology and medicine* 74 (2016): 69-73.

InSight

- Calvert et al. 2016
  - **TRIPOD-AI**
    - **Methods**
      - **Participants**
        - Inclusion criteria (3):
          1. Adult patients admitted to the MICU
          2. Does not meet SIRS criteria at the time of admission to the ICU of within first 4 hours of stay
          3. Measurements available for (i) systolic blood pressure (ii) pulse pressure (iii) heart rate (iv) temperature (v) respiration rate (vi)white blood cell count (vii) pH (viii) blood oxygen saturation (ix) age

✓

Calvert, Jacob S., et al. "A computational approach to early sepsis detection." *Computers in biology and medicine* 74 (2016): 69-73.

InSight

- Calvert et al. 2016
  - TRIPOD-AI
    - Methods
      - ✗ Data preparation: describe any data pre-processing steps, including cleaning, harmoisation, sampling, linkage, de-identiciation methods, and quality checks.
        - Not described

Calvert, Jacob S., et al. "A computational approach to early sepsis detection." *Computers in biology and medicine* 74 (2016): 69-73.

InSight

- Calvert et al. 2016
  - TRIPOD-AI
    - Methods
      - Outcome labeling: clearly define the outcome (e.g. ground truth or reference standard) that is predicted by the prediction model (including the time horizon), including how and when assessed and the rationale for choosing this outcome measurement (if alternatives exist).

Calvert, Jacob S., et al. "A computational approach to early sepsis detection." *Computers in biology and medicine* 74 (2016): 69-73.

InSight

- Calvert et al. 2016
  - TRIPOD-AI
    - Methods
      - Outcome labeling
        - Each of the patients underwent a **binary classification** process to designate them as positive or negative for having acquired in-hospital sepsis.
        - Classification was made based on the patient meeting both of the following criteria:

✓

1. The patient record contains an ICD9 code (995.9) indicating in-hospital contraction of sepsis.
2. The patient meets the 1991 Systemic Inflammatory Response Syndrome (SIRS) criteria for sepsis for a persistent 5-hour period of time. The beginning of the patient's first 5-hour SIRS event is defined as the zero hour.

- InSight was used to predict which patients would develop sepsis 3 hours before the zero hour

Calvert, Jacob S., et al. "A computational approach to early sepsis detection." *Computers in biology and medicine* 74 (2016): 69-73.

InSight

- Calvert et al. 2016
  - **TRIPOD-AI**
    - **Methods**
      - **Predictors: clearly define all predictors/features used in developing the multivariable prediction model, including how and when they were measured.**
        - Measurements available for (i) systolic blood pressure, (ii) pulse pressure, (iii) heart rate, (iv) temperature, (v) respiration rate, (vi) white blood cell count, (vii) pH, (viii) blood oxygen saturation and (ix) age
          - Selected for their standard availability, medical relevance to sepsis, and the reliable likelihood of their frequent determination in a clinical setting.
        - Beginning with ICU admission, the patient ICU stay was divided into one-hour intervals and measurement timestamps were rounded up to the nearest hour.
        - How was blood oxygen saturation measured? Pulse oximeter vs. direct arterial blood gas
          - Pulse-oximeter may overestimate oxygenation saturation by 2%

Calvert, Jacob S., et al. "A computational approach to early sepsis detection." *Computers in biology and medicine* 74 (2016): 69-73.

Seguin, Philippe, et al. "Evidence for the need of bedside accuracy of pulse oximetry in an intensive care unit." *Critical care medicine* 28.3 (2000): 703-706.

InSight

- Calvert et al. 2016
  - **TRIPOD-AI**
    - **Methods**
      - ✓ **Missing data: Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation or other data augmentation method**

        ⚠
        - Patients without an observation for each measurement were excluded
        - For intervals without observations, missing values were taken to be the most recent available observation.

Calvert, Jacob S., et al. "A computational approach to early sepsis detection." *Computers in biology and medicine* 74 (2016): 69-73.

InSight

- Calvert et al. 2016
  - TRIPOD-AI
    - Methods
      - ❌ Analytical methods: Describe how predictors/features were handled in the analyses (functional form and any standardization)
        - Not described

Calvert, Jacob S., et al. "A computational approach to early sepsis detection." *Computers in biology and medicine* 74 (2016): 69-73.

InSight

- Calvert et al. 2016
  - **TRIPOD-AI**
    - **Methods**
      - ✗ **Analytical methods: Specify the type of model, all model-building procedures (including any predictor selection), and method for internal validation (e.g. bootstrapping, cross-validation)**

- Mi: average value over last 5 hours
- Di: difference between current and value 5 hours prior classified as positive, negligible, or negative
- Dij: trends among pairs of measurements
- Dijk: trends among triplets of measurements

$$Score = a \sum_{i \in A} p(M_i) + b \sum_{i \in B} p(\hat{D}_i) + c \sum_{(i,j) \in C} p(\hat{D}_{ij}) + d \sum_{(i,j,k) \in D} p(\hat{D}_{ijk})$$

- Constants a–d were chosen to maximize the area under the training set receiver operator characteristic (ROC) curve (AUROC), using a standard optimization technique.

Calvert, Jacob S., et al. "A computational approach to early sepsis detection." *Computers in biology and medicine* 74 (2016): 69-73.

InSight

- Calvert et al. 2016
  - TRIPOD-AI
    - Methods
      - Analytical methods

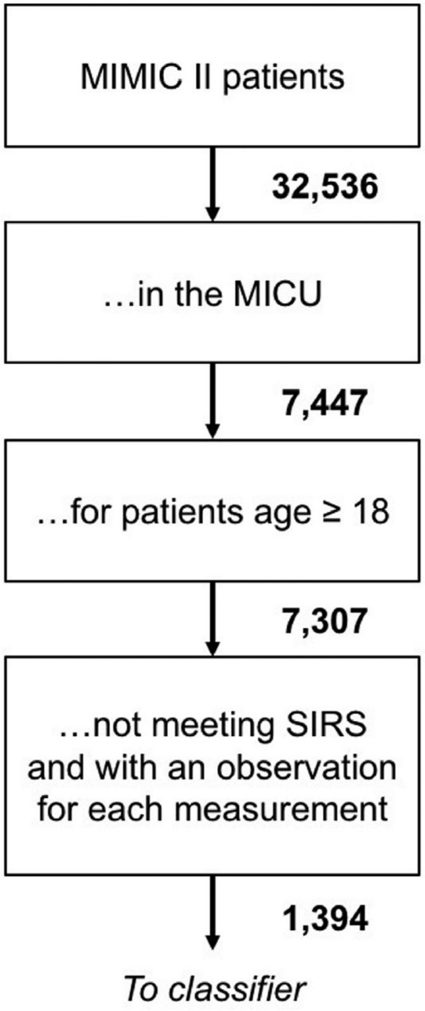
✗

- Details of model training approaches, including hyperparameters, number of models trained, used data sets
  - Not described
- Specify all measures used to assess model performance (e.g. discrimination, calibration) and, if relevant, to compare multiple models
  - Not described
- Describe the method for selecting the final model
  - Not described

Calvert, Jacob S., et al. "A computational approach to early sepsis detection." *Computers in biology and medicine* 74 (2016): 69-73.

InSight

- Calvert et al. 2016
  - **TRIPOD-AI**
    - **Results**
      - ✓ **Participants: describe the flow of the participants through the study**



MIMIC II patients

32,536

…in the MICU

7,447

…for patients age ≥ 18

7,307

…not meeting SIRS and with an observation for each measurement

1,394

To classifier

InSight

- Calvert et al. 2016
  - **TRIPOD-AI**
    - **Results**
      - ✗ **Report the characteristics overall and where applicable for each data source or setting, including the key dates, key predictors/features (including demographics, ethnicity), treatments received, sample size, number of outcome events, follow-up time, and amount of missing data**
        - 1394 patients met inclusion criteria
        - 159 (11%) met outcome criteria
        - **Overall characteristics, predictors not described**
        - **Missing data not described**

Calvert, Jacob S., et al. "A computational approach to early sepsis detection." *Computers in biology and medicine* 74 (2016): 69-73.

InSight

- Calvert et al. 2016
  - **TRIPOD-AI**
    - **Results**
      - **Model specification:**
        - ❌ **Provide details on the full prediction model to allow predictions for individuals to allow third-party evaluation and implementation (e.g. regression coefficients, input parameters, sharing of code/any dependencies). Provide reasons for not sharing code.**
          - **Not described**

Calvert, Jacob S., et al. "A computational approach to early sepsis detection." *Computers in biology and medicine* 74 (2016): 69-73.

InSight

- Calvert et al. 2016
  - **TRIPOD-AI**
    - **Results**
      - **Model performance: report performance measures (with confidence intervals, Cis for the prediction model).**
        - AUROC 0.92 (0.86 – 0.93)
          - Better than published AUROC of procalcitonin 0.85 (0.81 – 0.88)
        - Using score of 0.30 as the cutoff (scores higher than 0.30 indicate prediction of sepsis, sensitivity 90%, specificity 81%
          - Better than published 63% sensitivity and 80% specificity of procalcitonin

Calvert, Jacob S., et al. "A computational approach to early sepsis detection." *Computers in biology and medicine* 74 (2016): 69-73.

InSight

- Calvert et al. 2016
  - **TRIPOD-AI**
    - **Results**
      - **Model performance: report performance measures**
        - Confusion matrix

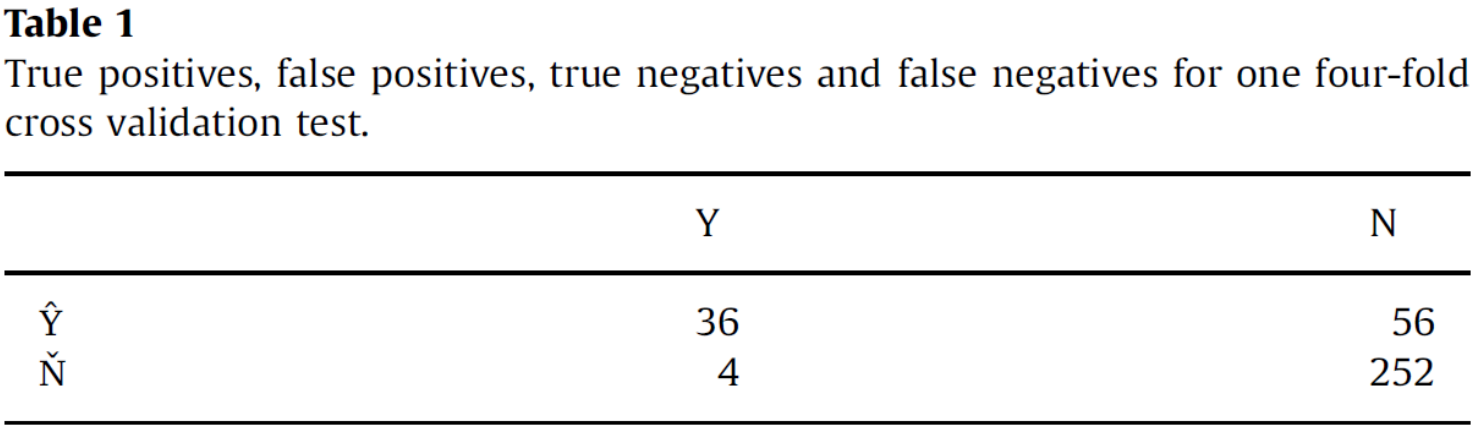


Table 1
True positives, false positives, true negatives and false negatives for one four-fold cross validation test.

	Y	N
Ŷ	36	56
Ň	4	252

^ *Y* indicates the number of patients predicted to become septic, while *Y* denotes the set of patients satisfying the gold standard criteria for sepsis.

Calvert, Jacob S., et al. "A computational approach to early sepsis detection." *Computers in biology and medicine* 74 (2016): 69-73.

Subsequent study

- **Objective:** Validate InSight for **new Sepsis-3 definition** and to investigate the effects of **data sparsity** on its performance
- Methods
  - Data source: MIMIC-III
  - 2016 Sepsis definition: Life-threatening organ dysfunction (≥2 SOFA score) caused by a dysregulated host response to infection

⚠

    - Outcome predicted was suspicion of infection, defined with an order for a culture lab draw, together with a dose of antibiotics, within a specified window

Desautels, Thomas, et al. "Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach." *JMIR medical informatics* 4.3 (2016): e5909.

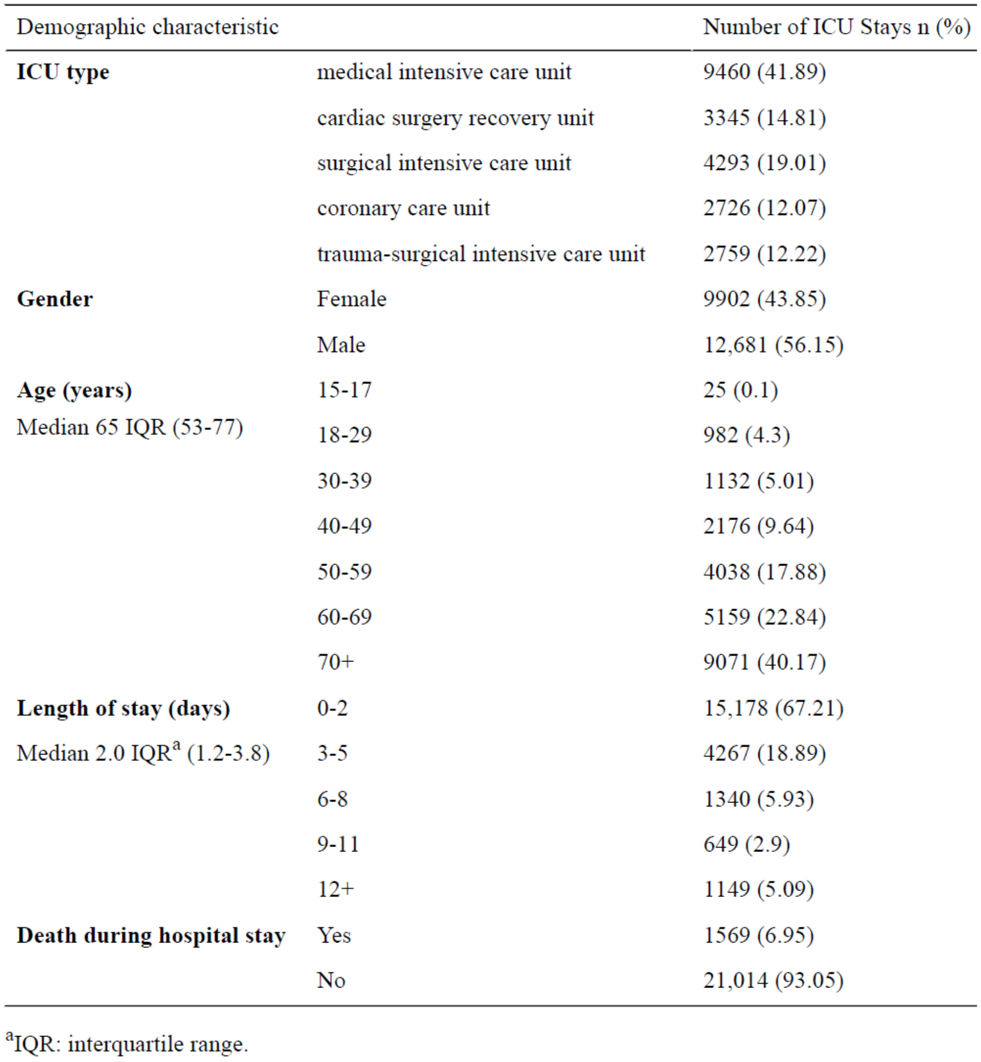
Subsequent study

- Methods
  - **Missing data**
    - Missing data are imputed using a “carry-forward” system, where the most recent bin value is carried forward to fill subsequent empty bins.
    - If the data required to calculate one of the SOFA subscores is not present in the imputed data, that subscore is given the value 0 (ie, “normal”).
    - **Are these assumptions reasonable?**
  - **Limited reporting on data preparation and analytical methods**

Desautels, Thomas, et al. "Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach." *JMIR medical informatics* 4.3 (2016): e5909.

Subsequent study

- Results
  - Table 2. Demographics of the included MIMIC-III intensive care unit stays.



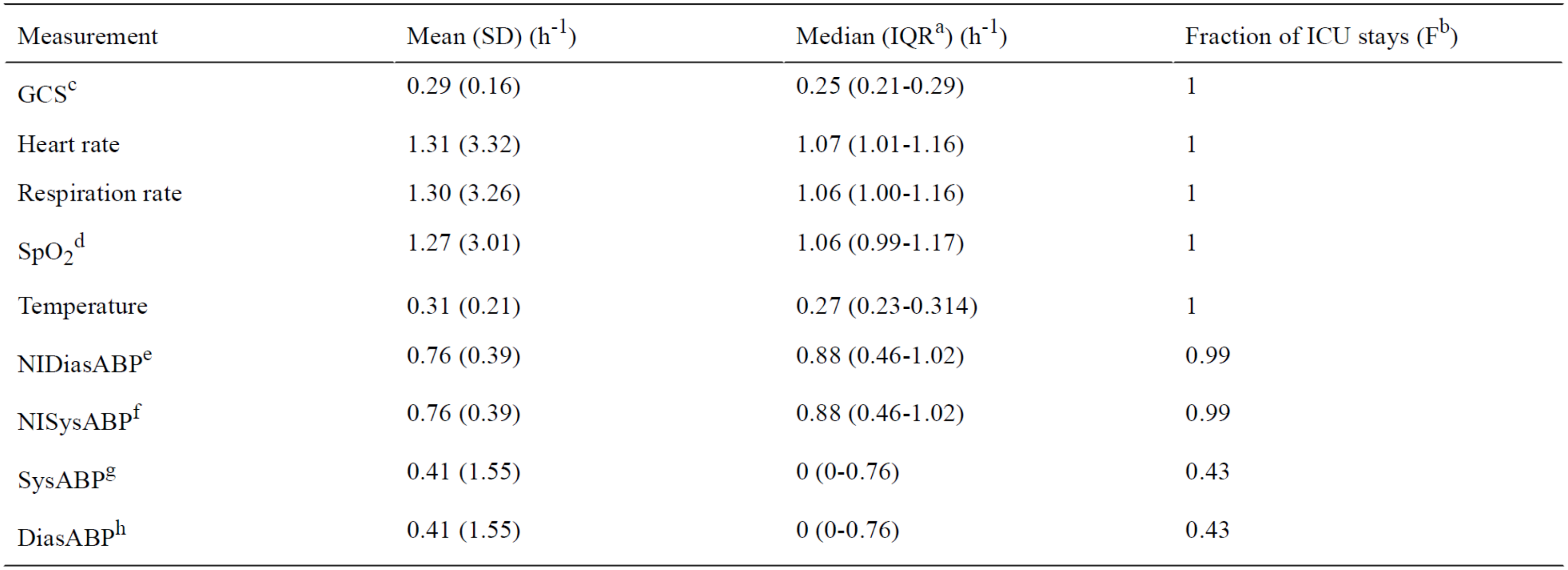
Demographic characteristic		Number of ICU Stays n (%)
ICU type	medical intensive care unit	9460 (41.89)
	cardiac surgery recovery unit	3345 (14.81)
	surgical intensive care unit	4293 (19.01)
	coronary care unit	2726 (12.07)
	trauma-surgical intensive care unit	2759 (12.22)
Gender	Female	9902 (43.85)
	Male	12,681 (56.15)
Age (years)	15-17	25 (0.1)
Median 65 IQR (53-77)	18-29	982 (4.3)
	30-39	1132 (5.01)
	40-49	2176 (9.64)
	50-59	4038 (17.88)
	60-69	5159 (22.84)
	70+	9071 (40.17)
Length of stay (days)	0-2	15,178 (67.21)
Median 2.0 IQR[a] (1.2-3.8)	3-5	4267 (18.89)
	6-8	1340 (5.93)
	9-11	649 (2.9)
	12+	1149 (5.09)
Death during hospital stay	Yes	1569 (6.95)
	No	21,014 (93.05)

[a] IQR: interquartile range.

Desautels, Thomas, et al. "Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach." *JMIR medical informatics* 4.3 (2016): e5909.

Subsequent study

- Results
  - Table 3. Per-hour observation frequencies among included ICU stays (n=22,853).



Measurement	Mean (SD) (h^{-1})	Median (IQR[a]) (h^{-1})	Fraction of ICU stays (F[b])
GCS[c]	0.29 (0.16)	0.25 (0.21-0.29)	1
Heart rate	1.31 (3.32)	1.07 (1.01-1.16)	1
Respiration rate	1.30 (3.26)	1.06 (1.00-1.16)	1
SpO_2[d]	1.27 (3.01)	1.06 (0.99-1.17)	1
Temperature	0.31 (0.21)	0.27 (0.23-0.314)	1
NIDiasABP[e]	0.76 (0.39)	0.88 (0.46-1.02)	0.99
NISysABP[f]	0.76 (0.39)	0.88 (0.46-1.02)	0.99
SysABP[g]	0.41 (1.55)	0 (0-0.76)	0.43
DiasABP[h]	0.41 (1.55)	0 (0-0.76)	0.43

[b] F: the fraction of these ICU stays with **at least one measurement** of the given type.

Subsequent study

- Results
  - **Limited description of model specification**
  - Model performance
    - AUROC at sepsis onset 0.88
      - Better than AUROC of other scores (SIRS, quick SOFA, MEWS, SAPS II, SOFA)
    - Performance measures of InSight when tested and trained with raw data dropouts
      - 10% dropout: 0.87
      - 20% dropout: 0.84
      - 40% dropout: 0.83
      - 60% dropout: 0.78

Desautels, Thomas, et al. "Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach." *JMIR medical informatics* 4.3 (2016): e5909.

Application of Insight

- Objective: evaluate improvements in sepsis-related outcomes with the use of InSight at an acute care hospital
- **Study design: pre-implementation and post-implementation analysis**
- Methods
  - Date source: EHR
  - Population: CRMC emergency and hospital populations
    - Cape Regional Medical Center (CRMC)
      - 242-bed acute care hospital located in Cape May Court House, New Jersey
    - Encounters included if they met 2 or more SIRS criteria at some point during their stay
  - Comparison: pre- vs. post-implementation of InSight
  - Primary outcome: sepsis-related in-hospital mortality rate at CRMC
  - Secondary outcomes: average sepsis-related hospital length of stay and the sepsis-related 30-day readmission rate

McCoy, Andrea, and Ritankar Das. "Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units." *BMJ open quality* 6.2 (2017): e000158.

Application of Insight

- Pre-implementation workflow
    - **Hospital patients**
        - **Manual sepsis scoring system**, tabulated for all non-ED patients twice per day.
            - Nurses checked each patient every 12 hours, or on identification of a potential source of infection, to determine if ≥2 SIRS criteria met.
                - if ≥2 SIRS criteria = true then
                    - Nurse ordered the nursing sepsis bundle
                    - Physician assessed the patient for severe sepsis and accordingly administered all or a portion of the physician sepsis bundle
    - **ED patients**
        - **No formalized sepsis screening process**
        - Similar interventions were made for patients suspected of or diagnosed with severe sepsis or septic shock.

McCoy, Andrea, and Ritankar Das. "Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units." *BMJ open quality* 6.2 (2017): e000158.

Application of Insight

- **Post-implementation workflow**
  - Use of Insight AND
  - Nurses continued tabulation of SIRS criteria every 12 hours for patients in non-ED units

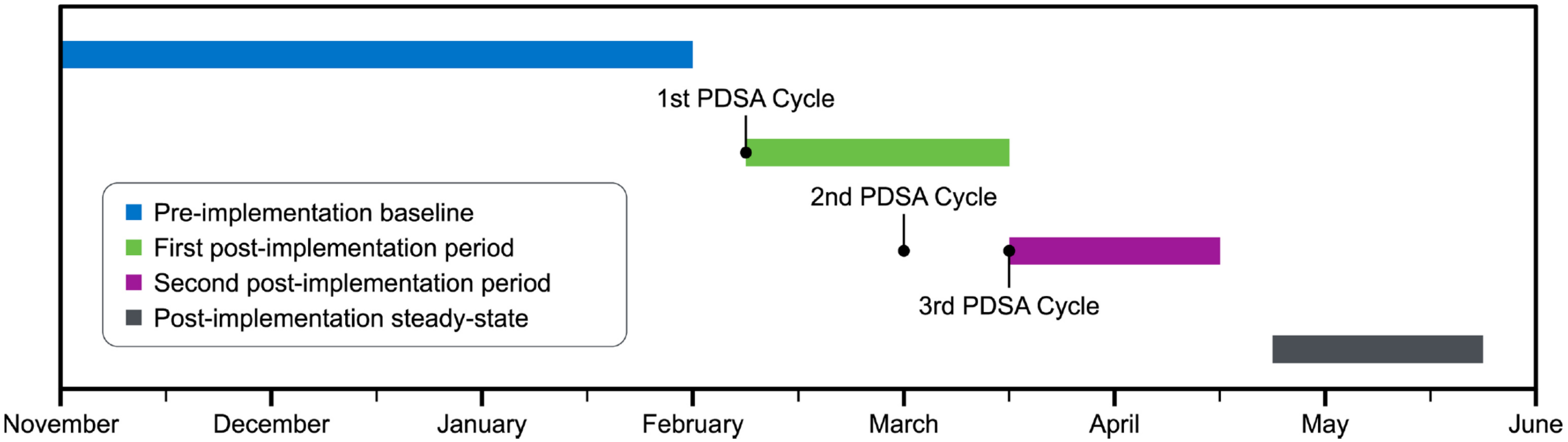


Figure 1 Timeline of patient outcome measurement collection periods and Plan-Do-Study-Act (PDSA) cycles for the study.

Application of Insight

- The quality improvement team regularly incorporated feedback from clinical leadership and end users through the **Plan-Do-Study-Act (PDSA) cycles**.
- **PDSA cycle 1**
  - Focused implementation: few units
  - After implementation, meetings to discuss systemic improvements.
    - Primary areas for improvement concerned the algorithm threshold and the reassessment of patients with sepsis.
    - **Clinicians indicated that due to the use of the algorithm, more patients required bedside assessment than the clinical staff could accommodate.**

McCoy, Andrea, and Ritankar Das. "Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units." *BMJ open quality* 6.2 (2017): e000158.

Application of Insight

- PDSA cycle 2
  - Objective: reduce alert fatigue
    - **Alert threshold adjusted** to reduce the number of flagged patients, increasing specificity of the alert
    - Incorporated a **6-hour 'snooze'** feature to prevent reassessment by the algorithm of any given patient in a 6-hour period
  - Implemented in ED

McCoy, Andrea, and Ritankar Das. "Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units." *BMJ open quality* 6.2 (2017): e000158.

Application of Insight

- **PDSA cycle 3**
  - Objective: adjusting the system's call logic.
  - **Clinicians indicated a lag time** between a prediction score call to a hospitalist and response time to an ED patient.
    - Due to the distance between the ED and other hospital units, it was quicker to direct all ED alerts to a charge nurse or clinical coordinator, rather than to a hospitalist.
    - Accordingly, calls were streamed based on patient location.

McCoy, Andrea, and Ritankar Das. "Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units." *BMJ open quality* 6.2 (2017): e000158.

Application of Insight

- Results

Outcome	Baseline
Mortality rate	7.4%

$3.6 million of cost savings per year

- Improvement in the 3-hour severe sepsis SEP-1 bundle compliance.
  - Pre-implementation 49% vs. post-implementation: 73%

McCoy, Andrea, and Ritankar Das. "Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units." *BMJ open quality* 6.2 (2017): e000158.

InSight Randomized Clinical Trial

- **First time a machine learning-based sepsis prediction system has been investigated in a randomized, interventional design.**
- Population: all patients (age ≥18) admitted in 2 ICUs at a UCSF Medical Center between December 2016 and February 2017
- **Randomized to experimental vs. control group**
  - Healthcare providers, patients, and investigators **blinded** to assignment, although assignments revealed for patients who generated alerts
  - A patient admitted with a sepsis diagnosis was still monitored by the prediction algorithm for potential further septic episodes; thus, these patients were not excluded from the trial.

Shimabukuro, David W., et al. "Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial." *BMJ open respiratory research* 4.1 (2017): e000234.

InSight Randomized Clinical Trial

- Intervention
  - Control group: normal standard of care (nurse evaluation) and monitored by the existing EHR-based severe sepsis detector
  - Experimental group: monitored by InSight and the existing severe sepsis detector
- Outcomes:
  - Primary: average hospital length of stay
  - Secondary outcomes: in-hospital mortality rate and ICU length of stay

Shimabukuro, David W., et al. "Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial." *BMJ open respiratory research* 4.1 (2017): e000234.

InSight Randomized Clinical Trial

- 142 patients randomized (67 experimental vs. 75 control)

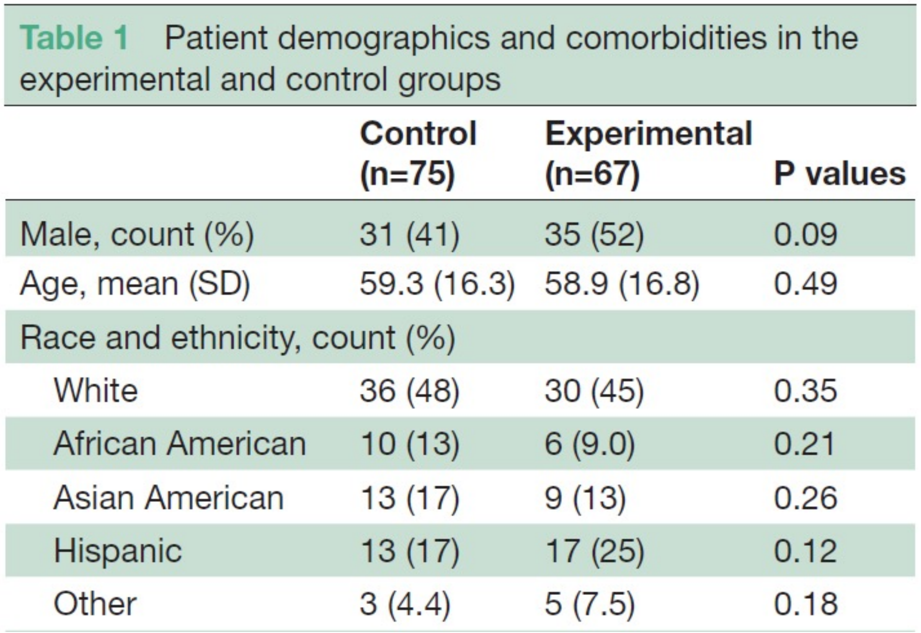
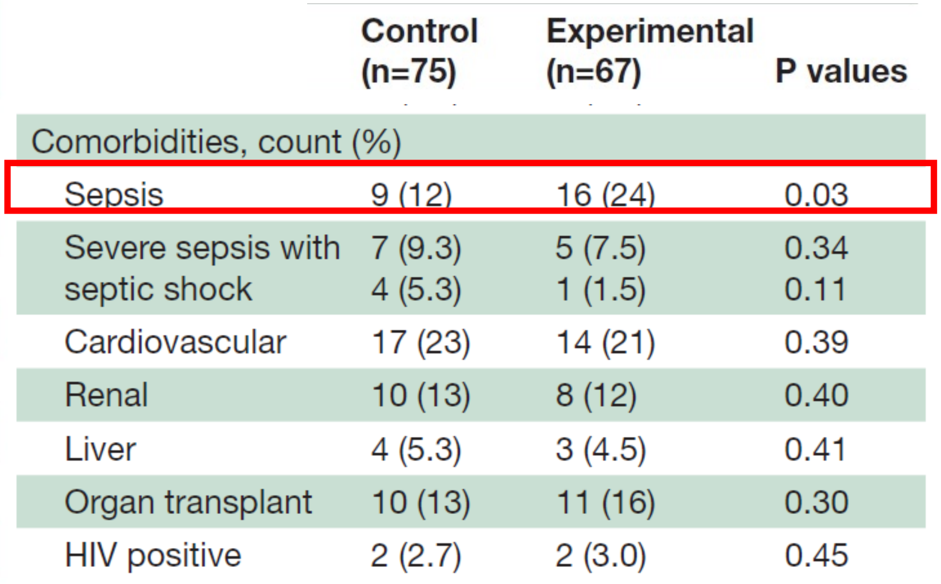


Table 1 Patient demographics and comorbidities in the experimental and control groups

	Control (n=75)	Experimental (n=67)	P values
Male, count (%)	31 (41)	35 (52)	0.09
Age, mean (SD)	59.3 (16.3)	58.9 (16.8)	0.49
Race and ethnicity, count (%)			
White	36 (48)	30 (45)	0.35
African American	10 (13)	6 (9.0)	0.21
Asian American	13 (17)	9 (13)	0.26
Hispanic	13 (17)	17 (25)	0.12
Other	3 (4.4)	5 (7.5)	0.18



	Control (n=75)	Experimental (n=67)	P values
Comorbidities, count (%)			
Sepsis	9 (12)	16 (24)	0.03
Severe sepsis with	7 (9.3)	5 (7.5)	0.34
septic shock	4 (5.3)	1 (1.5)	0.11
Cardiovascular	17 (23)	14 (21)	0.39
Renal	10 (13)	8 (12)	0.40
Liver	4 (5.3)	3 (4.5)	0.41
Organ transplant	10 (13)	11 (16)	0.30
HIV positive	2 (2.7)	2 (3.0)	0.45

Shimabukuro, David W., et al. "Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial." *BMJ open respiratory research* 4.1 (2017): e000234.

InSight Randomized Clinical Trial

- Results

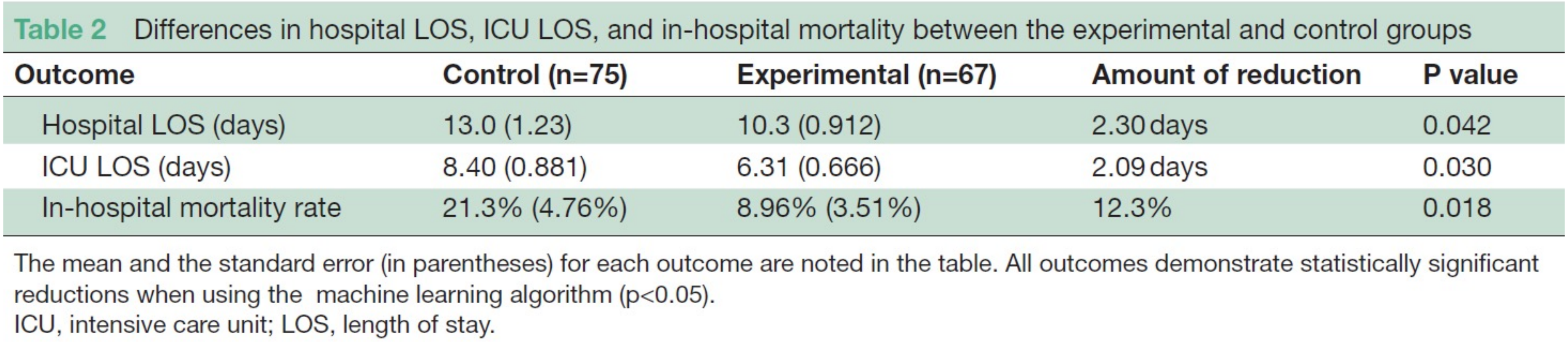


Table 2 Differences in hospital LOS, ICU LOS, and in-hospital mortality between the experimental and control groups

Outcome	Control (n=75)	Experimental (n=67)	Amount of reduction	P value
Hospital LOS (days)	13.0 (1.23)	10.3 (0.912)	2.30 days	0.042
ICU LOS (days)	8.40 (0.881)	6.31 (0.666)	2.09 days	0.030
In-hospital mortality rate	21.3% (4.76%)	8.96% (3.51%)	12.3%	0.018

The mean and the standard error (in parentheses) for each outcome are noted in the table. All outcomes demonstrate statistically significant reductions when using the machine learning algorithm ($p<0.05$).
ICU, intensive care unit; LOS, length of stay.

Shimabukuro, David W., et al. "Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial." *BMJ open respiratory research* 4.1 (2017): e000234.

InSight Multicenter Evaluation

- **Pre-implementation vs. post-implementation analysis at 9 hospitals**
- Results

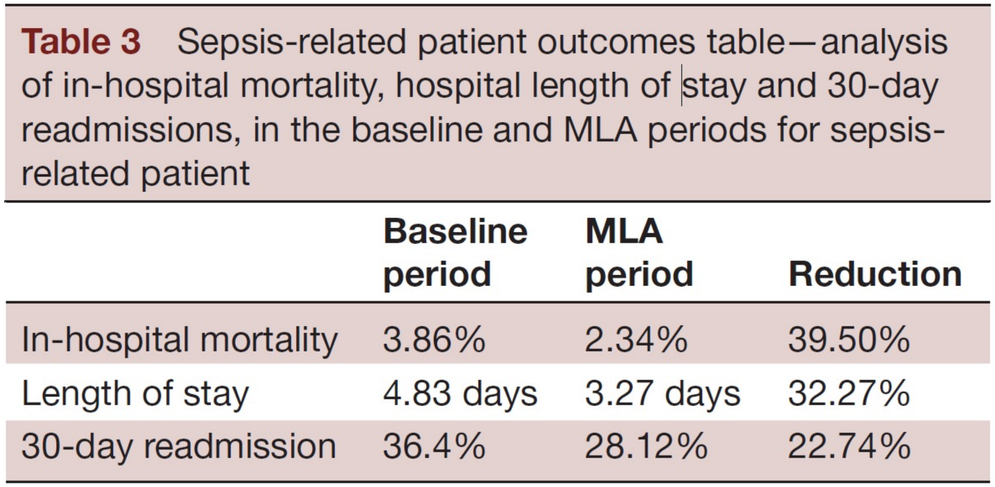


Table 3 Sepsis-related patient outcomes table—analysis of in-hospital mortality, hospital length of stay and 30-day readmissions, in the baseline and MLA periods for sepsis-related patient

	Baseline period	MLA period	Reduction
In-hospital mortality	3.86%	2.34%	39.50%
Length of stay	4.83 days	3.27 days	32.27%
30-day readmission	36.4%	28.12%	22.74%

Reduction of LOS translates to ≈US$14.5 million of annual cost savings across all 9 hospitals included in this analysis

Burdick, Hoyt, et al. "Effect of a sepsis prediction algorithm on patient mortality, length of stay and readmission: a prospective multicentre clinical outcomes evaluation of real-world patient data from US hospitals." *BMJ health & care informatics* 27.1 (2020).

InSight Multicenter Randomized Trial

NIH U.S. National Library of Medicine

ClinicalTrials.gov

Home > Search Results > Study Record Detail

RCT of Sepsis Machine Learning Algorithm

- The focus of this study will be to conduct **a prospective, multi-center randomized controlled trial (RCT)** at Cape Regional Medical Center (CRMC), Oroville Hospital (OH), and UCSF Medical Center (UCSF) in which a machine-learning algorithm will be applied to EHR data for the detection of sepsis.

https://clinicaltrials.gov/ct2/show/NCT03882476

EPIC Sepsis Prediction Model

- Epic Sepsis Model (ESM)
  - Proprietary sepsis prediction model developed by Epic Systems Corporation
    - EPIC is largest EHR vendor in US
  - Uses Demographic, comorbidity, vital sign, laboratory, medication, and procedural variables data.
- **Limited information on performance, with no independent validation**

Bennett, Tellen, et al. "Accuracy of the Epic sepsis prediction model in a regional health system." *arXiv preprint arXiv:1902.07276* (2019).
Wong, Andrew, et al. "External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients." *JAMA Internal Medicine* 181.8 (2021): 1065-1070.

EPIC Sepsis Prediction Model

- Objective: external validation of the ESM using data from a large academic medical center
- Methods
  - Population: all patients (age ≥18) admitted to Michigan Medicine between December 6, 2018 and October 20, 2019
  - ESM scores calculated for all hospitalizations

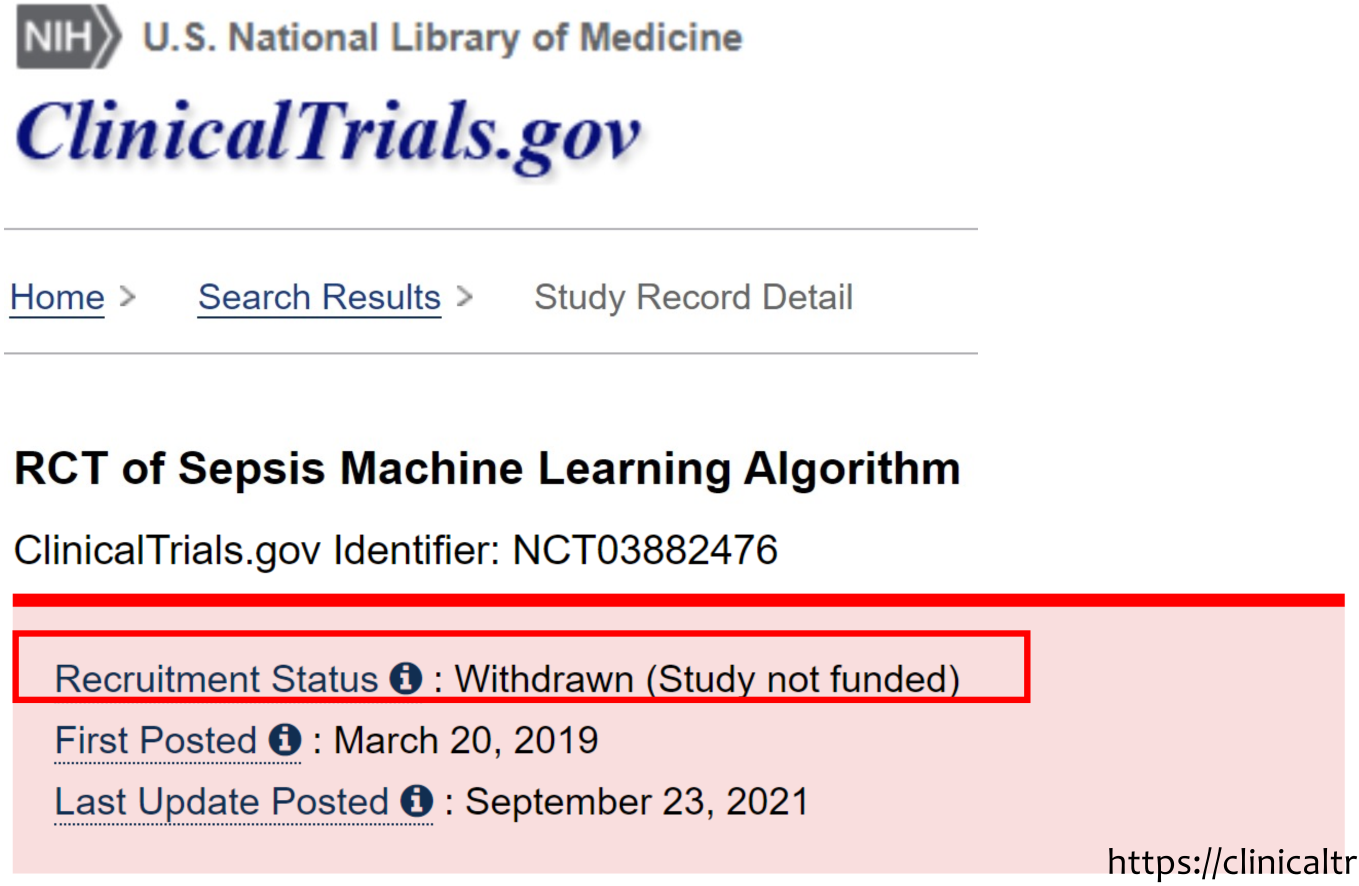
Wong, Andrew, et al. "External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients." *JAMA Internal Medicine* 181.8 (2021): 1065-1070.

EPIC Sepsis Prediction Model

- Results
  - Model performance
    - Hospitalization-level AUC 0.63
      - EPIC internal documentation AUC 0.76-0.83
      - Prior conference proceeding (coauthored with EPIC) AUC 0.73
    - At selected ESM threshold of 6
      - Sensitivity 33%
      - Specificity 83%
      - Positive predictive value 12%
      - Negative predictive value 95%

Wong, Andrew, et al. "External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients." *JAMA Internal Medicine* 181.8 (2021): 1065-1070.

InSight Multicenter Randomized Trial



NIH U.S. National Library of Medicine

ClinicalTrials.gov

Home > Search Results > Study Record Detail

RCT of Sepsis Machine Learning Algorithm

ClinicalTrials.gov Identifier: NCT03882476

Recruitment Status ℹ : Withdrawn (Study not funded)

First Posted ℹ : March 20, 2019

Last Update Posted ℹ : September 23, 2021

https://clinicaltrials.gov/ct2/show/NCT03882476

- **Targeted Real-time Early Warning System (TREWS)**
  - **Developed using MIMIC-II data**
  - TREWScore identified patients before the onset of septic shock with an AUROC 0.83
  - Performance subsequently evaluated in retrospective cohort of unselected patients admitted to a community hospital (ICU and non-ICU patients), AUROC 0.94

Henry, Katharine E., et al. "A targeted real-time early warning score (TREWScore) for septic shock." *Science translational medicine* 7.299 (2015): 299ra122-299ra122.
Henry, Katharine, et al. "Can septic shock be identified early? Evaluating performance of A targeted real-time early warning score (TREWScore) for septic shock in a community hospital: global and subpopulation performance." *D15. Critical Care: Do We Have a Crystal Ball? Predicting Clinical Deterioration and Outcome in Critically Ill Patients*. American Thoracic Society, 2017. A7016-A7016.

BAYESIAN
HEALTH

- **TREWS Application Experience**
  - Objective: identify which patient, provider, and environmental factors influence adoption of TREWS in the real-world setting
  - Population: all adults who presented to the emergency department (ED) or were admitted to a medical or surgical unit at any of five Johns Hopkins Health System hospitals between April 2018 and March 2020

Henry, Katharine E., et al. "Evaluating Adoption, Impact, and Factors Driving Adoption for TREWS, a Machine Learning-Based Sepsis Alerting System." *medRxiv* (2021).

- **TREWS Deployment Experience**
  - Performance
    - 9,805 (2.1%) encounters identified as having sepsis
      - 8,033 of these **(82%) of sepsis encounters were flagged by TREWS**
    - TREWS system screened 469,419 patient encounters
      - System flagged 31,591 (6.7%) patient encounters for sepsis screening
  - **Overall adoption**
    - 89% of all patient encounters with an alert had a provider evaluation entered

Henry, Katharine E., et al. "Evaluating Adoption, Impact, and Factors Driving Adoption for TREWS, a Machine Learning-Based Sepsis Alerting System." *medRxiv* (2021).

BAYESIAN
HEALTH

- **TREWS Deployment Experience**
  - Association between adoption and patient care
    - **A timely evaluation entered by a physician was associated with a 1.12** (95% CI 0.87 - 1.30) **hour reduction in the** adjusted median **time from alert to first antibiotic** order compared with not having a timely evaluation entered in the TREWS tool

Henry, Katharine E., et al. "Evaluating Adoption, Impact, and Factors Driving Adoption for TREWS, a Machine Learning-Based Sepsis Alerting System." *medRxiv* (2021).

BAYESIAN
HEALTH

- **TREWS Deployment Experience**
  - Patient, provider, and environmental factors are associated with **alert adoption**
    - Patient factors:
      - Advanced age (adjusted risk ratio 1.06)
    - Environmental factor:
      - High alert level (aRR 0.94)
      - **Alert occurred 7am-3pm** (aRR 1.03)
    - Provider factors:
      - ED provider (aRR 1.22)
      - Provider experience w/ alert (aRR 1.22)

Henry, Katharine E., et al. "Evaluating Adoption, Impact, and Factors Driving Adoption for TREWS, a Machine Learning-Based Sepsis Alerting System." *medRxiv* (2021).

BAYESIAN
HEALTH

- **TREWS Deployment Experience**
  - Patient, provider, and environmental factors are associated with inappropriate **alert dismissal** on sepsis patients
    - Patient factors:
      - **Absence of key sepsis symptoms** (aRR 1.28)
      - Acute general severity (aRR 1.46)
      - Advanced age (aRR 0.69)
    - Environmental:
      - Alert occurred 3pm-11pm (aRR 1.20)
      - Alert occurred 11pm-7pm (aRR 1.19)
    - Provider factors
      - ED provider (aRR 0.47)
      - Provider experience w/ alert (aRR 0.66)

Henry, Katharine E., et al. "Evaluating Adoption, Impact, and Factors Driving Adoption for TREWS, a Machine Learning-Based Sepsis Alerting System." *medRxiv* (2021).

TREWScore Transportability

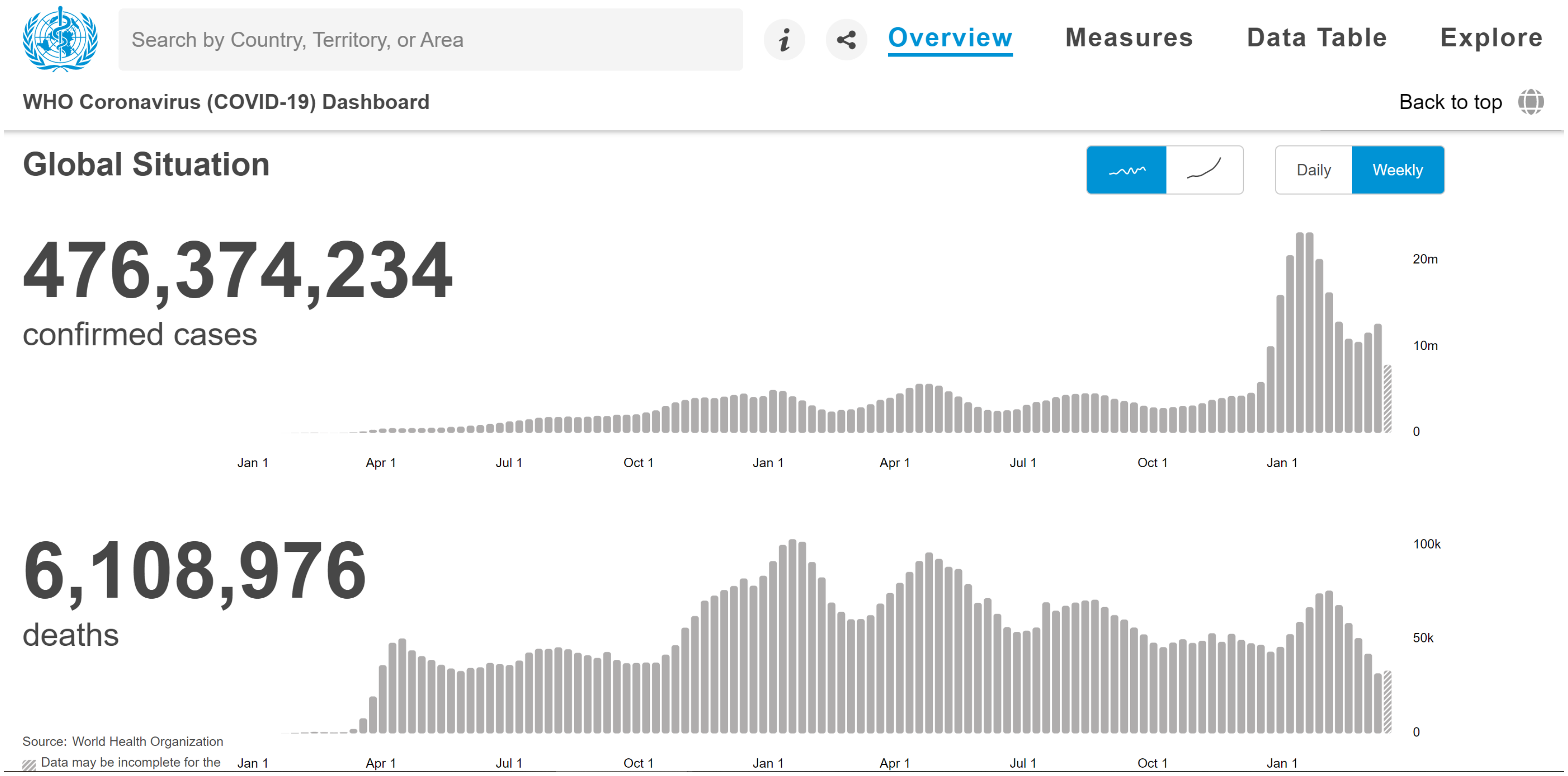
- Objective: evaluate transportability of TREWScore to University Medical Center, Utrecht, Netherlands
- Results
  - **Significant differences in cohort characteristics between MIMIC-III and UMC ICU; UMC ICU more severely ill**
    - UMC ICU cohort was younger with a higher proportion of men
    - Proportions blood pressure monitoring, and mechanical ventilation were all higher in the UMC ICU cohort
    - Total hospital length of stay and hospital mortality longer in the UMC cohort .
  - **Not all 54 TREWScore criteria easily available**
    - 38 available in UMC EHR, 14 require feature engineering, 1 requires text mining, 1 unavailable

Niemantsverdriet, Michael SA, et al. "Transportability and Implementation Challenges of Early Warning Scores for Septic Shock in the ICU: A Perspective on the TREWScore." *Frontiers in medicine* 8 (2021).

Today’s Outline

- Critical appraisal
- Case 1: Early prediction of Sepsis
- **Case 2: Diagnosis of COVID-19 with imaging**
- Case 3: Detection of diabetic retinopathy

COVID-19



WHO Coronavirus (COVID-19) Dashboard

Global Situation

476,374,234
confirmed cases

6,108,976
deaths

Source: World Health Organization

▨ Data may be incomplete for the

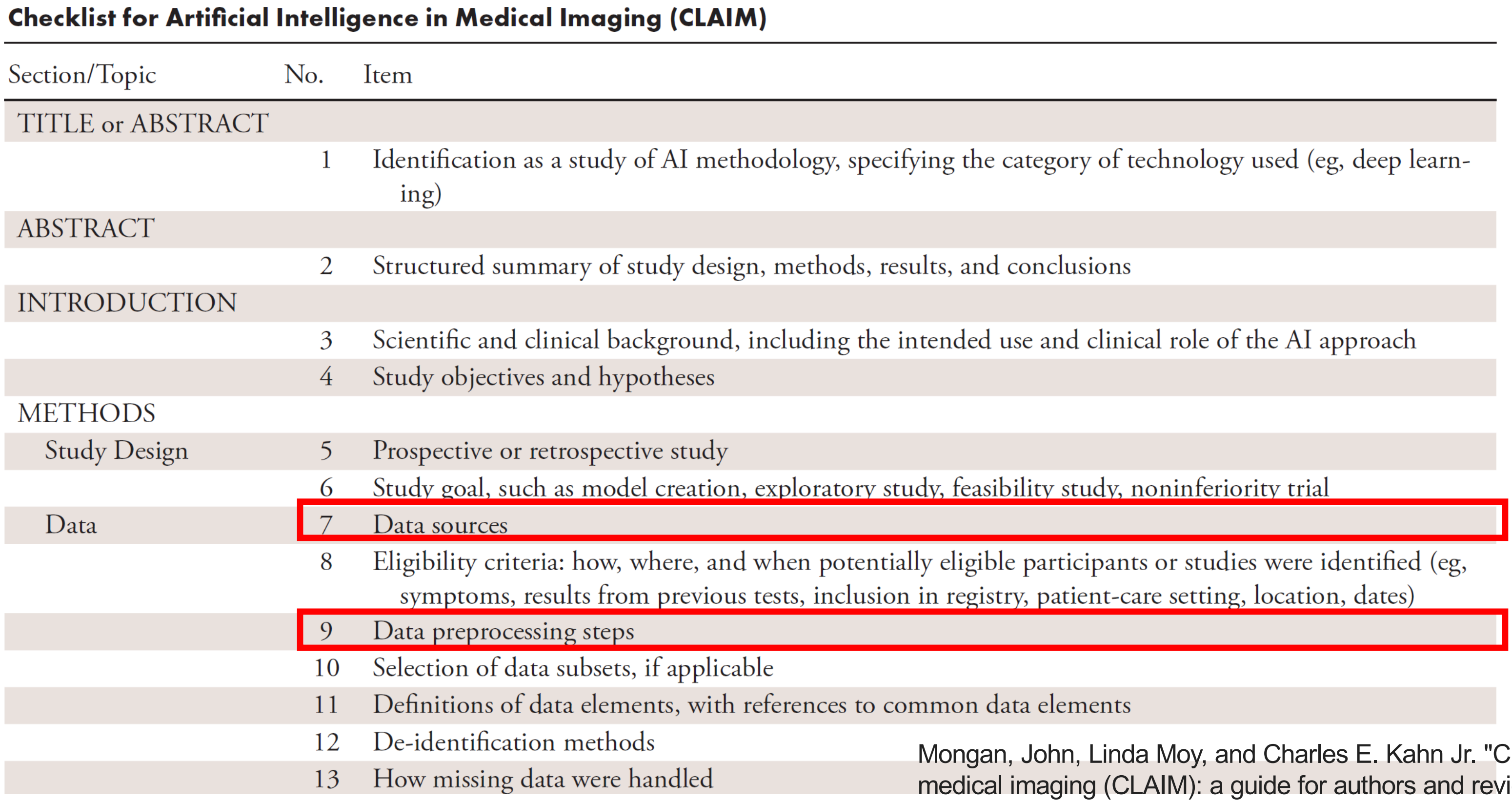
as of **6:15pm CET, 25 March 2022**

https://covid19.who.int/

Diagnosis of COVID-19

- PCR with reverse transcription **(RT-PCR) is the test of choice** for diagnosing COVID-19
- **Potential benefits of image-based diagnosis**
  - Improve speed and accuracy
  - Surrogate in areas with limited access to RT-PCR
  - CXR abnormalities are visible in some patients who initially had a negative RT-PCR
  - CT scan may have higher sensitivity than RT-PCR
- In response to the pandemic, several machine learning models were developed

Checklist for Artificial Intelligence in Medical Imaging (CLAIM)

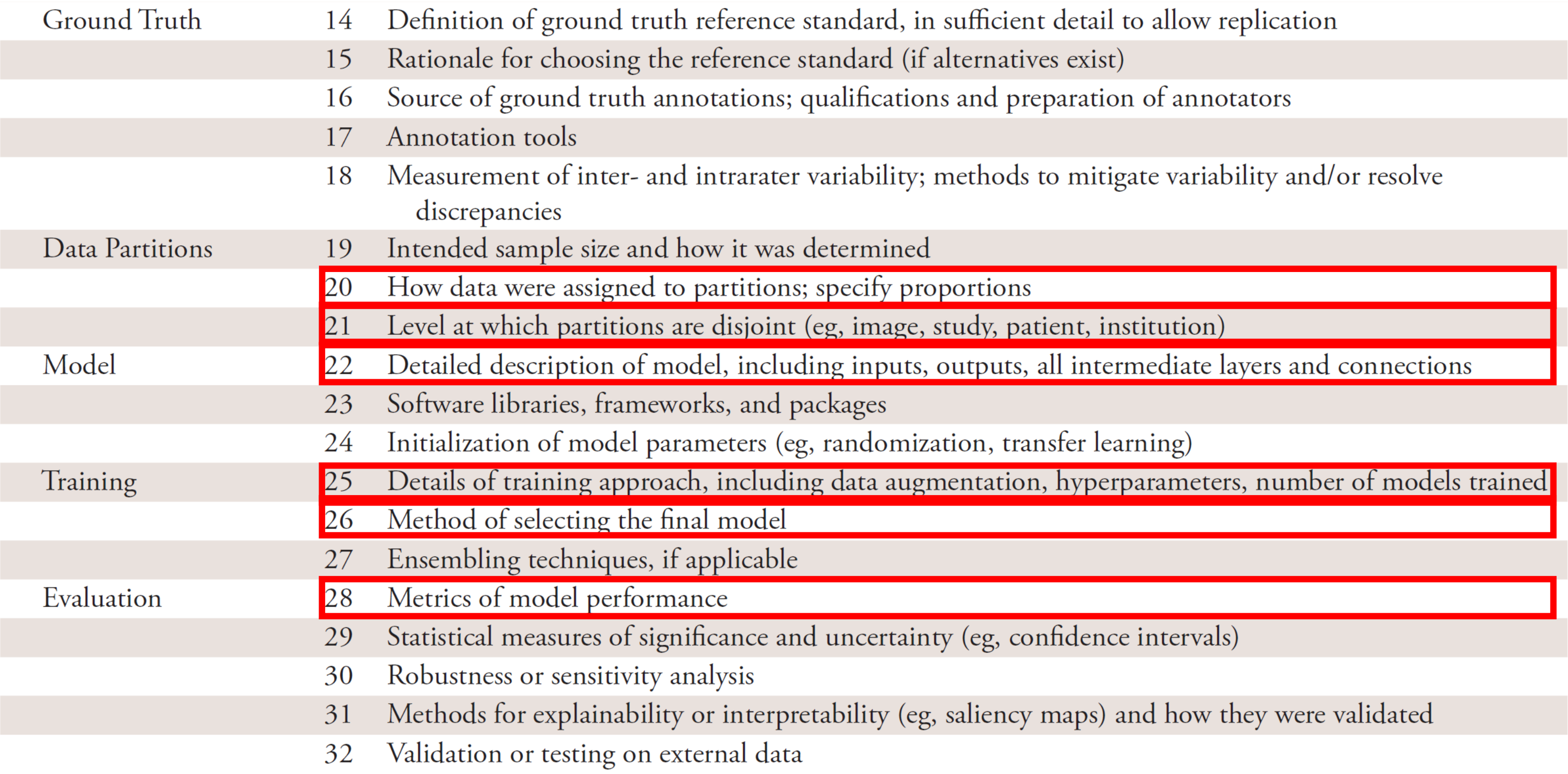


Checklist for Artificial Intelligence in Medical Imaging (CLAIM)

Section/Topic	No.	Item
TITLE or ABSTRACT		
	1	Identification as a study of AI methodology, specifying the category of technology used (eg, deep learning)
ABSTRACT		
	2	Structured summary of study design, methods, results, and conclusions
INTRODUCTION		
	3	Scientific and clinical background, including the intended use and clinical role of the AI approach
	4	Study objectives and hypotheses
METHODS		
Study Design	5	Prospective or retrospective study
	6	Study goal, such as model creation, exploratory study, feasibility study, noninferiority trial
Data	7	Data sources
	8	Eligibility criteria: how, where, and when potentially eligible participants or studies were identified (eg, symptoms, results from previous tests, inclusion in registry, patient-care setting, location, dates)
	9	Data preprocessing steps
	10	Selection of data subsets, if applicable
	11	Definitions of data elements, with references to common data elements
	12	De-identification methods
	13	How missing data were handled

Mongan, John, Linda Moy, and Charles E. Kahn Jr. "Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers." *Radiology: Artificial Intelligence* 2.2 (2020): e200029.

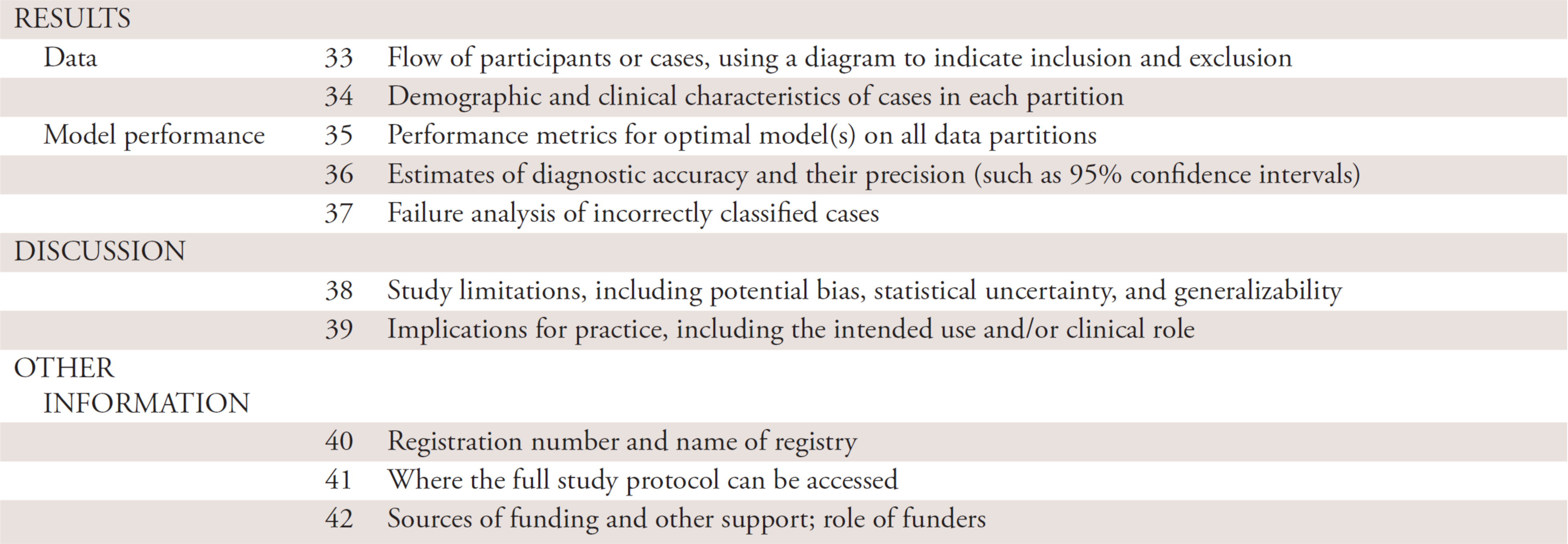
Checklist for Artificial Intelligence in Medical Imaging (CLAIM)



Ground Truth	14	Definition of ground truth reference standard, in sufficient detail to allow replication
	15	Rationale for choosing the reference standard (if alternatives exist)
	16	Source of ground truth annotations; qualifications and preparation of annotators
	17	Annotation tools
	18	Measurement of inter- and intrarater variability; methods to mitigate variability and/or resolve discrepancies
Data Partitions	19	Intended sample size and how it was determined
	20	How data were assigned to partitions; specify proportions
	21	Level at which partitions are disjoint (eg, image, study, patient, institution)
Model	22	Detailed description of model, including inputs, outputs, all intermediate layers and connections
	23	Software libraries, frameworks, and packages
	24	Initialization of model parameters (eg, randomization, transfer learning)
Training	25	Details of training approach, including data augmentation, hyperparameters, number of models trained
	26	Method of selecting the final model
	27	Ensembling techniques, if applicable
Evaluation	28	Metrics of model performance
	29	Statistical measures of significance and uncertainty (eg, confidence intervals)
	30	Robustness or sensitivity analysis
	31	Methods for explainability or interpretability (eg, saliency maps) and how they were validated
	32	Validation or testing on external data

Mongan, John, Linda Moy, and Charles E. Kahn Jr. "Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers." *Radiology: Artificial Intelligence* 2.2 (2020): e200029.

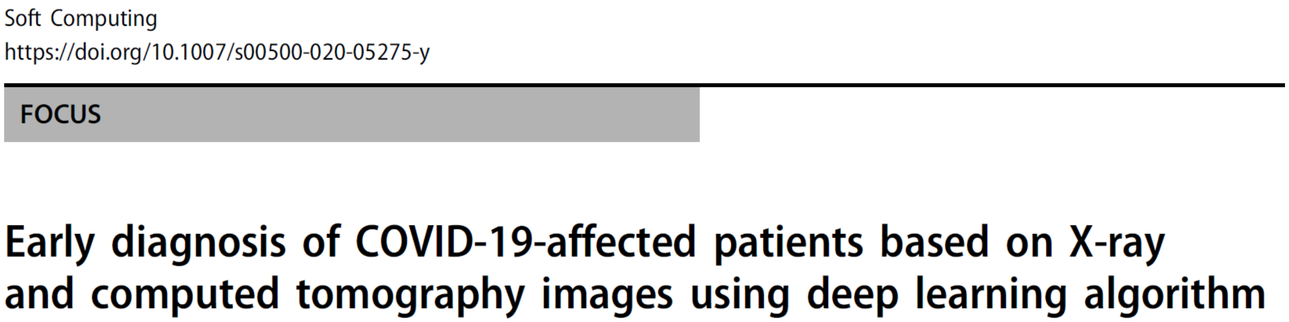
Checklist for Artificial Intelligence in Medical Imaging (CLAIM)



RESULTS		
Data	33	Flow of participants or cases, using a diagram to indicate inclusion and exclusion
	34	Demographic and clinical characteristics of cases in each partition
Model performance	35	Performance metrics for optimal model(s) on all data partitions
	36	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)
	37	Failure analysis of incorrectly classified cases
DISCUSSION		
	38	Study limitations, including potential bias, statistical uncertainty, and generalizability
	39	Implications for practice, including the intended use and/or clinical role
OTHER INFORMATION		
	40	Registration number and name of registry
	41	Where the full study protocol can be accessed
	42	Sources of funding and other support; role of funders

Mongan, John, Linda Moy, and Charles E. Kahn Jr. "Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers." *Radiology: Artificial Intelligence* 2.2 (2020): e200029.

ML for Diagnosis of COVID19 from Images



Soft Computing
https://doi.org/10.1007/s00500-020-05275-y

FOCUS

Early diagnosis of COVID-19-affected patients based on X-ray and computed tomography images using deep learning algorithm

- CLAIM, Methods, Data source
  - The data sources must be clearly identified to allow reproducible collection of the same datasets.
  - In this paper, chest X-ray and CT scan images of 360 patients have been acquired from the open-source database (https://github.com/ieee8023/covid-chestxray-dataset), out of which are 360 images of COVID-19 patients, 16 images of SARS and 18 images of streptococcus. This repository is comprising chest X-ray/CT images for the most part of patients with acute respiratory distress syndrome (ARDS), COVID-19, E-Coli, streptococcus, pneumocystis, pneumonia and severe acute respiratory syndrome (SARS).

Dansana, Debabrata, et al. "Early diagnosis of COVID-19-affected patients based on X-ray and computed tomography images using deep learning algorithm." *Soft Computing* (2020): 1-9.

ML for Diagnosis of COVID19 from Images

- CLAIM, Methods, Data pre-processing
  - Item 9 Describe preprocessing steps fully and in sufficient detail so that other investigators could reproduce them. Specify the use of normalization, resampling of image size, change in bit depth, and/or adjustment of window/level settings. State whether or not the data have been rescaled, threshold-limited ("binarized"), and/or standardized. Specify how the following issues were handled: regional format, manual input, inconsistent data, missing data, wrong data types, file manipulations, and missing anonymization. Define any criteria to remove outliers. Specify the libraries, software (including manufacturer name and location), and version numbers, and all option and configuration settings employed.
  - All images were resized to 224 9 224 pixels.
  - We perform different cleaning steps with the data like preprocessing, splitting and data augmentation.

Dansana, Debabrata, et al. "Early diagnosis of COVID-19-affected patients based on X-ray and computed tomography images using deep learning algorithm." *Soft Computing* (2020): 1-9.

ML for Diagnosis of COVID19 from Images

- CLAIM, Methods, Data partitions
  - Item 20 Specify how the data were assigned into training, validation ("tuning"), and testing partitions; indicate the proportion of data in each partition and justify that selection. Indicate if there are any systematic differences between the data in each partition, and if so, why.
    - Not described

Dansana, Debabrata, et al. "Early diagnosis of COVID-19-affected patients based on X-ray and computed tomography images using deep learning algorithm." *Soft Computing* (2020): 1-9.

ML for Diagnosis of COVID19 from Images

- CLAIM, Methods, Model
  - Item 25 Completely describe all of the training procedures and hyperparameters in sufficient detail that another investigator could exactly duplicate the training process.
  - For neural networks, descriptions of hyperparameters should include at least learning rate schedule, optimization algorithm, minibatch size, dropout rates (if any), and regularization parameters (if any).
  - Discuss what objective function was employed, why it was selected, and to what extent it matches the performance required for the clinical or scientific use case.
  - Define criteria used to select the best-performing model.
  - Not described

Mongan, John, Linda Moy, and Charles E. Kahn Jr. "Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers." *Radiology: Artificial Intelligence* 2.2 (2020): e200029.

ML for Diagnosis of COVID19 from Images

- **CLAIM, Methods, Method of selecting the final model**
  - **Item 26 Describe the method and performance parameters used to select the best-performing model among all the models trained for evaluation against the held-out test set. If more than one model is selected, justify why this is appropriate.**
  - **Not described**
- **CLAIM, Methods, Metrics of model performance**
  - **Item 28 Describe the metric(s) used to measure the model's performance and indicate how they address the performance characteristics most important to the clinical or scientific problem. Compare the presented model to previously published models.**
  - Seven unique metrics were utilized to assess the proposed method. These metrics are precision, recall, F1 score, support, accuracy, micro average and weighted average.

Mongan, John, Linda Moy, and Charles E. Kahn Jr. "Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers." *Radiology: Artificial Intelligence* 2.2 (2020): e200029.

Systematic Review of ML for Diagnosis of COVID19 from Images

- Objective: review the literature of ML methods as applied to Chest CT and CXR for the diagnosis and prognosis of COVID-19
- All studies underwent an initial quality screening stage using “8 mandatory” CLAIM criteria
- 254 deep learning-based studies identified
  - **215 (85%) excluded due to missing ≥1 CLAIM criteria**
    - 110 (51%) fail ≥3 CLAIM criteria
    - 3 most common reasons for a paper failing the quality check was due to insufficient documentation on
      1. How the final model was selected in 61%
      2. The method of pre-processing of the images in 58%
      3. The details of the training approach (for example, the optimizer,the loss function, the learning rate) in 49%

Roberts, Michael, et al. "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans." *Nature Machine Intelligence* 3.3 (2021): 199-217.

Systematic Review of ML for Diagnosis of COVID19 from Images

- 37 deep learning-based studies identified meeting mandatory CLAIM criteria
  - 29 did not complete any external validation
  - 30 did not perform any robustness or sensitivity analysis of their model
  - 26 did not report the demographics of their data partitions
  - 25 did not report the statistical tests used to assess significance of results or determine confidence intervals
  - 23 did not report confidence intervals for the performance
  - 22 did not sufficiently report their limitations, biases or issues around generalizability

Roberts, Michael, et al. "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans." *Nature Machine Intelligence* 3.3 (2021): 199-217.

Today's Outline

- Critical appraisal
- Case 1: Early prediction of Sepsis
- Case 2: Diagnosis of COVID-19 with imaging
- **Case 3: Detection of diabetic retinopathy**

Diabetic Retinopathy

- 2014 estimated worldwide prevalence: **422 million**
  - Prevalence is increasing; 1980 estimated worldwide prevalence: 108 million
- Among US patients with diabetes, **≈1/3 have diabetic retinopathy**
- **Diabetic Retinopathy (DR)**
  - **One of the leading causes of vision impairment in the world**
  - Condition **caused by chronically high blood sugar that damages blood vessels in the retina**, the thin layer at the back of the eye responsible for sensing light and sending signals to the brain.
    - These blood vessels can leak or hemorrhage, causing vision distortion or loss.
  - In early stages of DR, a patient often has no symptoms.
    - **Early detection is key to initiate timely treatment and mitigate the risk of blindness.**

https://www.who.int/news-room/fact-sheets/detail/diabetes

Beede, Emma, et al. "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy." *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020.

Diagnosis of Diabetic Retinopathy

- **Traditionally, retinal photography with manual interpretation** has been used as a screening tool for diabetic retinopathy
- **Potential benefits of automated grading** of diabetic retinopathy
  - Near **instantaneous reporting of results**; improving patient outcomes by providing early detection and treatment.
  - **Reducing barriers to access**
  - **Reproducibility**; consistency of interpretation (because a machine will make the same prediction on a specific image every time)

Gulshan, Varun, et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." *Jama* 316.22 (2016): 2402-2410.

ML for Automated Grading of Diabetic Retinopathy

- Objective: train a deep learning algorithm to detect referable diabetic retinopathy and assess the performance of the algorithm in 2 clinical validation sets.
- CLAIM
  - Data sources
    - Training: 128,175 macula-centered images of which 33,894 were from India (Aravind Eye Hospital, Sankara Nethralaya, and Narayana Nethralaya) and the rest from EyePACS sites.
    - The datasets from India were obtained from both eye hospital clinics and screening camps.
    - The EyePACS data consists of patients that were screened using the EyePACS tele-ophthalmology platform from January 2013 to April 2015. EyePACS clinics serve higher percentages of the latino population in the U.S., therefore, the EyePACS dataset was enriched for Hispanic patients (~55%), with Caucasian, Black, and Asian patients each comprising approximately 5-10% of the population. Cameras were used to acquire the images include Centervue DRS, Optovue iCam, Canon CR1/DGi/CR2, Topcon NW8 using 45-degree fields of view.

Mongan, John, Linda Moy, and Charles E. Kahn Jr. "Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers." *Radiology: Artificial Intelligence* 2.2 (2020): e200029.
Gulshan, Varun, et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." *Jama* 316.22 (2016): 2402-2410.

ML for Automated Grading of Diabetic Retinopathy

Google

- CLAIM
  - Data pre-processing
    - For algorithm training, input images were scale normalized by detecting the circular mask of the fundus image and resizing the diameter of the fundus to be 299 pixels wide.
    - Images for which the circular mask could not be detected were excluded from the development and the clinical validation sets. This corresponded to 117 out of 128,175 on the development set, 17 out of 9,963 in EyePACS-1, and none in Messidor-2.

Mongan, John, Linda Moy, and Charles E. Kahn Jr. "Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers." *Radiology: Artificial Intelligence* 2.2 (2020): e200029.
Gulshan, Varun, et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." *Jama* 316.22 (2016): 2402-2410.

ML for Automated Grading of Diabetic Retinopathy

- **CLAIM**
  - **Methods, Model**
    - CNN architecture: Inception-v3
    - Optimization algorithm: stochastic gradient descent
    - Preinititiallization using weights from ImageNet
    - Early stopping criteria used to terminate training before convergence
    - Ensemble of 10 networks trained on the same data was used, and the final prediction was computed by a linear average over the predictions of the ensemble.

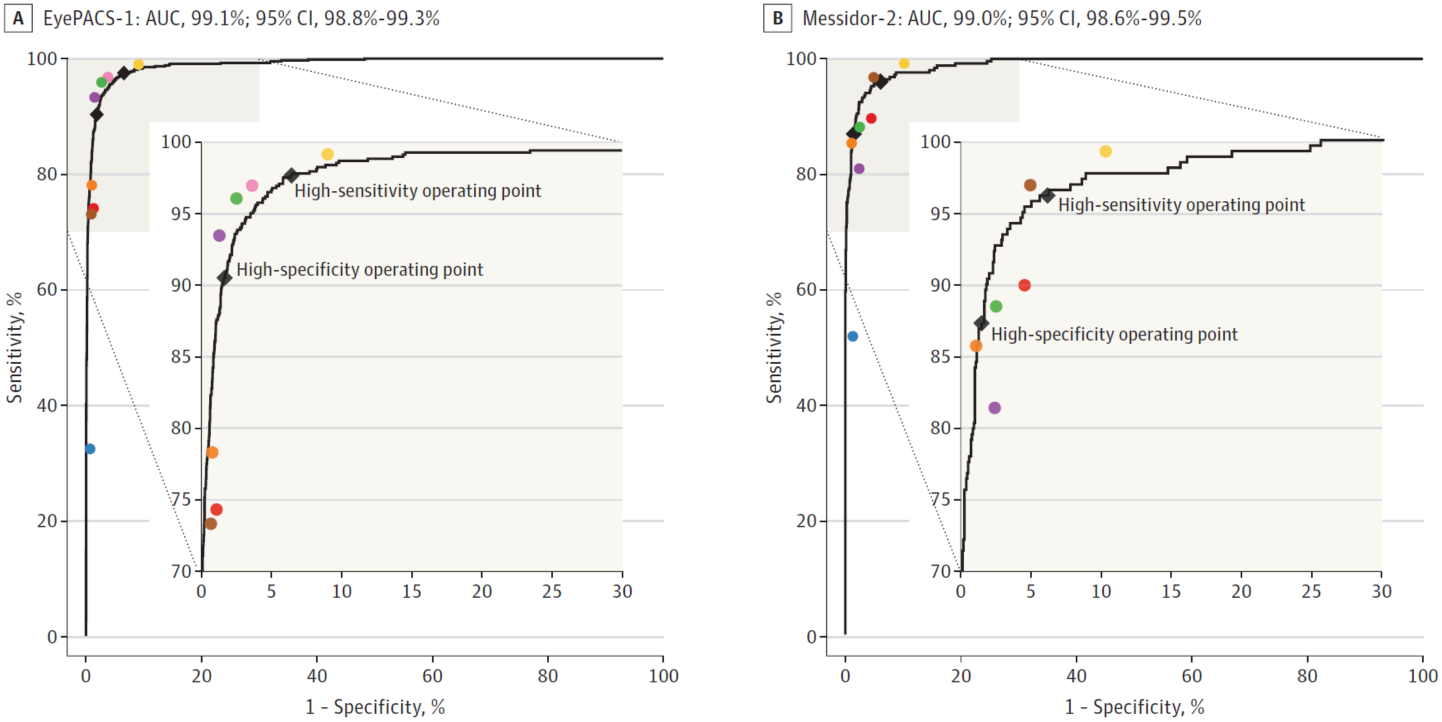
  - For neural networks, descriptions of hyperparameters should include at least learning rate schedule, optimization algorithm, minibatch size, dropout rates (if any), and regularization parameters (if any).

Mongan, John, Linda Moy, and Charles E. Kahn Jr. "Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers." *Radiology: Artificial Intelligence* 2.2 (2020): e200029.
Gulshan, Varun, et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." *Jama* 316.22 (2016): 2402-2410.

ML for Automated Grading of Diabetic Retinopathy

- Results

Figure 2. Validation Set Performance for Referable Diabetic Retinopathy



A EyePACS-1: AUC, 99.1%; 95% CI, 98.8%-99.3%

B Messidor-2: AUC, 99.0%; 95% CI, 98.6%-99.5%

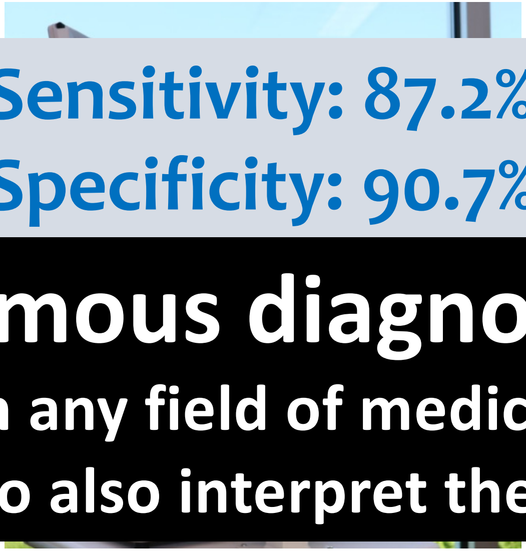
Gulshan, Varun, et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." *Jama* 316.22 (2016): 2402-2410.

Diabetic retinopathy

- Evaluating the performance of autonomous AI algorithm to diagnose diabetic retinopathy

819 patients with diabetes

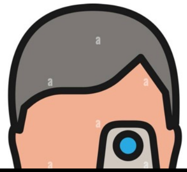
AI system



Sensitivity: 87.2%

Specificity: 90.7%

Human expert



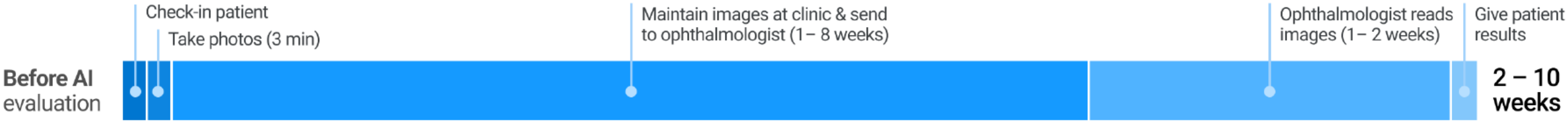
First autonomous diagnostic AI system

authorized by FDA in any field of medicine - without the need for a clinician to also interpret the image or results

Abràmoff, Michael D., et al. "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices." NPJ digital medicine 1.1 (2018): 1-8.

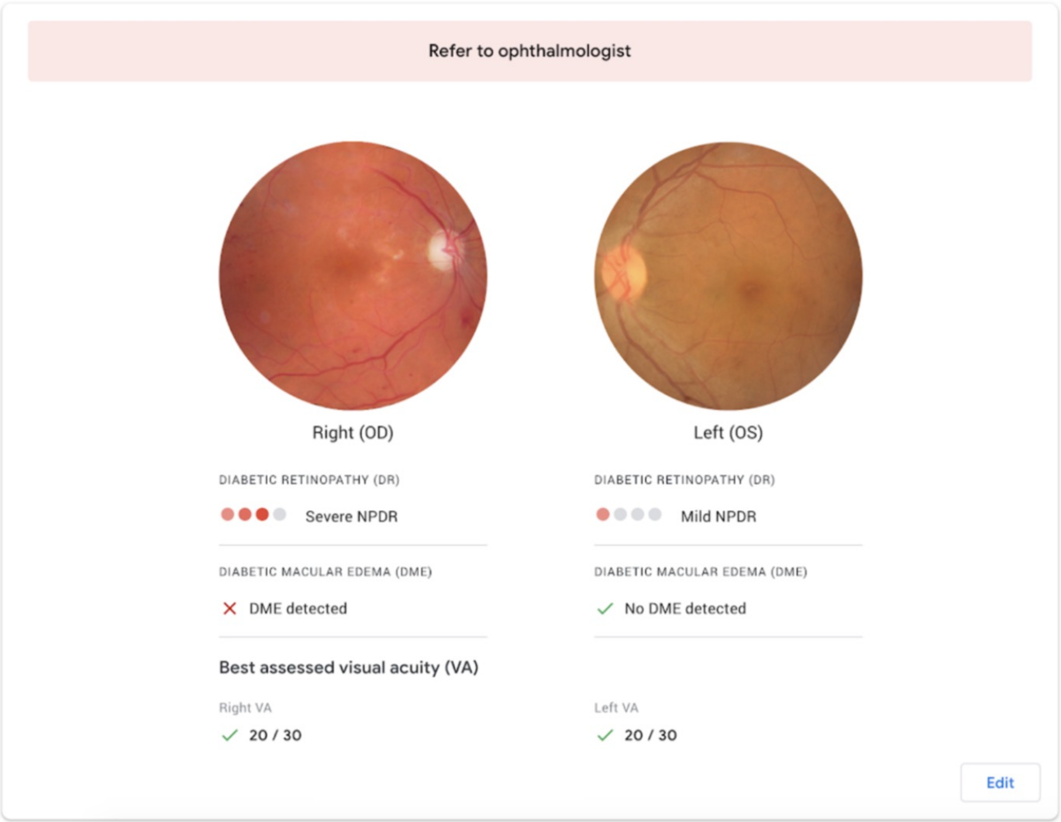
Application of ML for Automated Grading of Diabetic Retinopathy

- In many countries, shortage of clinical specialists limits the ability to screen diabetic patients for retinopathy
- Objective: explore the expectations and realities that nurses encounter in bringing a deep learning model into their clinical practices at 11 clinics across Thailand


Check-in patient
Take photos (3 min)
Maintain images at clinic & send to ophthalmologist (1 – 8 weeks)
Ophthalmologist reads images (1 – 2 weeks)
Give patient results
Before AI evaluation
2 – 10 weeks

Application of ML for Automated Grading of Diabetic Retinopathy

- Once a patient was called for their eye exam, the camera operator verified consent, took photos of each eye using the clinic’s current fundus camera, and uploaded them to the deep learning system.
- The images were sent to the algorithm in the cloud, and an assessment of the presence and severity of DR, was returned in real-time, including a recommendation for whether or not the patient should be referred to an ophthalmologist



Refer to ophthalmologist

Right (OD)

DIABETIC RETINOPATHY (DR)

●●●○ Severe NPDR

DIABETIC MACULAR EDEMA (DME)

✗ DME detected

Left (OS)

DIABETIC RETINOPATHY (DR)

●○○○ Mild NPDR

DIABETIC MACULAR EDEMA (DME)

✓ No DME detected

Best assessed visual acuity (VA)

Right VA

✓ 20 / 30

Left VA

✓ 20 / 30

Figure 3. Web application displaying the deep learning model’s predictions for diabetic retinopathy and diabetic macular edema, along with the fundus photos.

Application of ML for Automated Grading of Diabetic Retinopathy

- **Pre-deployment findings**
  - Lighting often suboptimal for fundus photos
    - We were interested to see how these real-world conditions would affect our model performance
  - Expectations for AI-assisted screening
    - Images need to be prominently displayed alongside the DR prediction
      - Provide confidence to the nurse that the correct image was being used for the assessment
      - Provide nurses with information they could use to convince patients to seek treatment
    - Potential benefits
      - Learning opportunity, to improve their own ability to make accurate DR assessments themselves
      - Use the system’s results to prove their own readings to on-site doctors
        - Several nurses expressed frustration with their assessments being undervalued or dismissed by physicians, and they were excited about the potential to demonstrate their own expertise to more senior clinicians.

Beede, Emma, et al. "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy." *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020.

Application of ML for Automated Grading of Diabetic Retinopathy

- **Pre-deployment findings**
  - Expectations for AI-assisted screening
    - Potential challenges
      - **Concern about increased workload** (following the study protocol (including uploading images)) and reduce ability to screen all patients arriving each day.
      - **Concern about false positives**, including the additional travel burden to follow up on a referral, the cost of missing work associated with travel, and the emotional strain a positive result

Beede, Emma, et al. "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy." *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020.

Application of ML for Automated Grading of Diabetic Retinopathy

- **Post-deployment findings**
  - Consenting patients
    - **Informed consent process was made more complicated** by the need to explain the deep learning system.
    - With deep learning system, **referral recommendations would need to be made immediately**, compared to previous workflow, where results may not be available for up to 10 weeks
      - Some nurses observed to dissuade patients from participations in the prospective study, for fear that it would cause unnecessary hardship.

Beede, Emma, et al. "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy." *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020.

Application of ML for Automated Grading of Diabetic Retinopathy

- **Post-deployment findings**
  - Clinical factors Influence System Performance
    - Gradeability
      - System rejects images not deemed high-quality, since it cannot guarantee that it hasn't missed something
      - **≈20% of images deemed ungradeable**
        - Low-quality images related to
          - Non-darkened environments
          - Dysfunctional camera
          - Lack of using dilating drops on patients
      - In the case of an ungradable image, the system notifies the nurse and recommends the patient be referred to an ophthalmologist, as part of the prospective study protocol.
        - Turned out to be frustrating as images they felt were human-readable were rejected by the system
        - This in-the-moment feedback caused the nurses to take more photos, in an attempt to achieve an image the system will grade.

Application of ML for Automated Grading of Diabetic Retinopathy

- **Post-deployment findings**
  - Clinical factors Influence System Performance
    - **Internet speed and connectivity**
      - One key difference in the eye screening workflow before and after the implementation of the deep learning system is that images are now uploaded to the cloud to get an assessment while the patient waits for results
      - With a strong internet connection, these results appear within a few seconds. However, the **clinics in our study often experienced slower and less reliable connections.** This causes some images to take 60-90 seconds to upload, **slowing down the screening queue and limiting the number of patients that can be screened in a day.**
        - **In one clinic, the internet went out for a period of two hours during eye screening**, reducing the number of patients screened from 200 to only 100.

Application of ML for Automated Grading of Diabetic Retinopathy

- Our research highlights that end-users and their environment determine how a new system will be implemented; that **implementation is of equal importance to the accuracy of the algorithm itself**, and cannot always be controlled through careful planning.
- By incorporating human-centered evaluations into deep learning model evaluations, and studying model performance on live data generated at the clinical site, we can reduce the risk that deep learning systems will fail in the wild, and increase the likelihood for meaningful improvements to patients and clinicians.

Beede, Emma, et al. "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy." *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020.

Summary

- A critical appraisal framework is useful to evaluate the rigor and utility of ML in healthcare studies
- Reproducibility is a key component of the scientific process
- Despite ML models with high accuracy, clinical application remains a challenge

Questions

6.871/HST.956: Machine Learning for Healthcare