AMDIS Machine Learning Survey and Update

Michael Wang, MD
Xinran (Leo) Liu, MD
University of California, San Francisco

sli.do

Join at
slido.com
#Y182

What barriers have you run into with deploying AI

Disclosures

Leo Liu:

- Clinical Advisor – Advanced Clinical L.L.C
  - Based at Google Brain
- Fiancée
  - Machine Learning Scientist at Amazon

Mike Wang:

- Clinical Consultant - Commure, Inc

Roadmap

Roadmap

Survey Results

Roadmap

Survey Results

PICO for ML Papers

Roadmap

Survey Results

PICO for ML Papers

Discussing Barriers to ML

WHAT IS MACHINE LEARNING

“Machine learning is the science of getting computers to act without being explicitly programmed.”

– Andrew Ng, Stanford University

Taking advantage of computational power

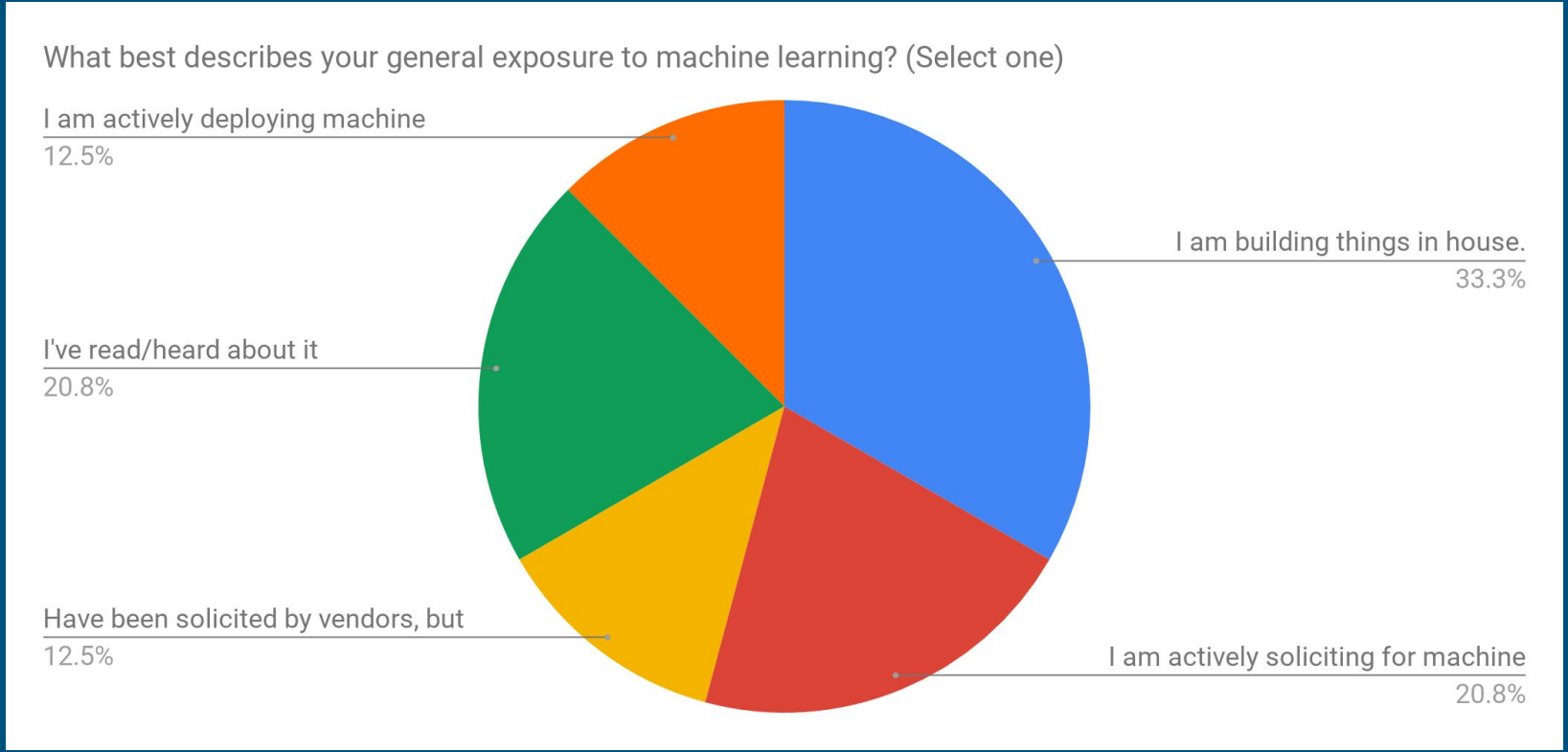
Roadmap

Survey Results

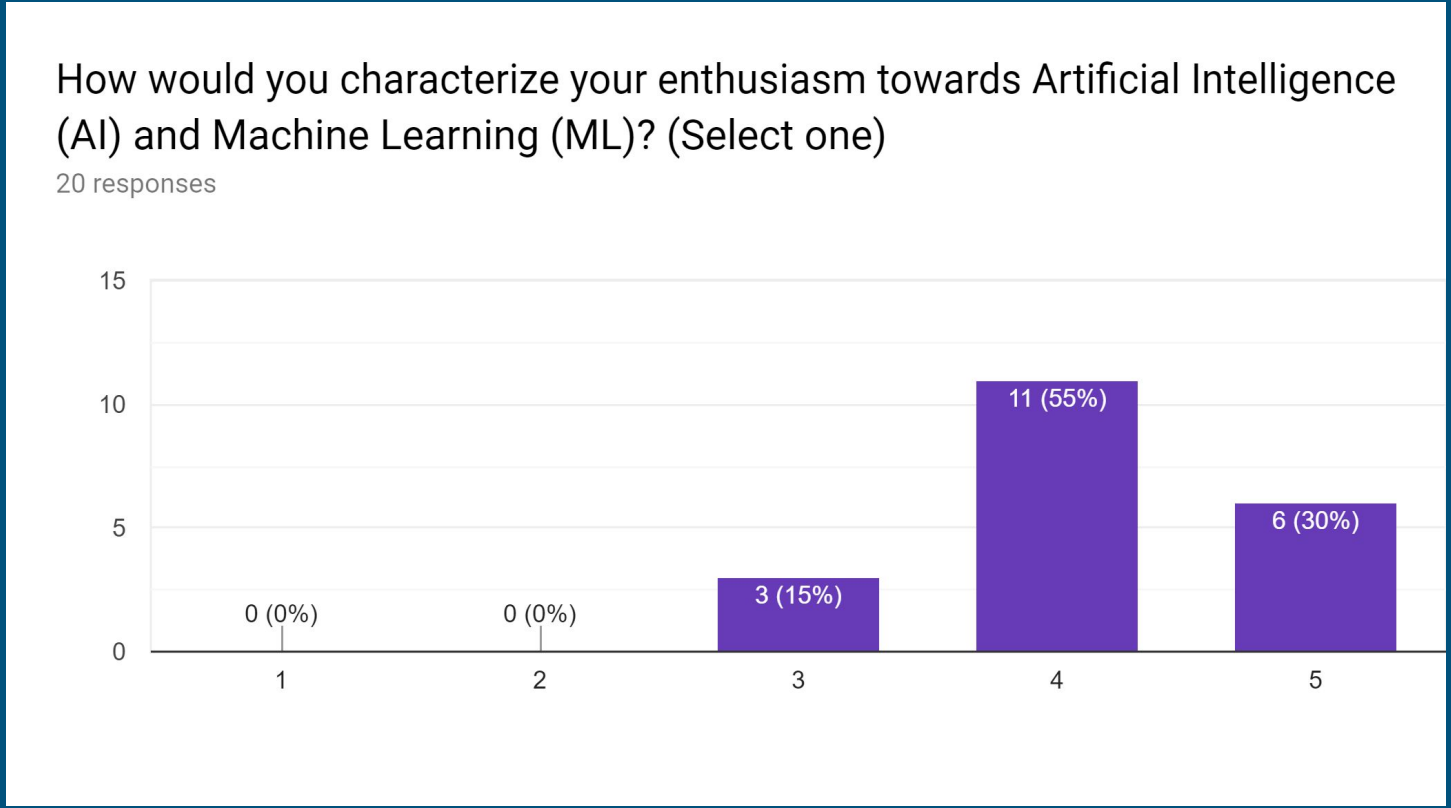
PICO for ML Papers

Discussing Barriers to ML

Survey Results



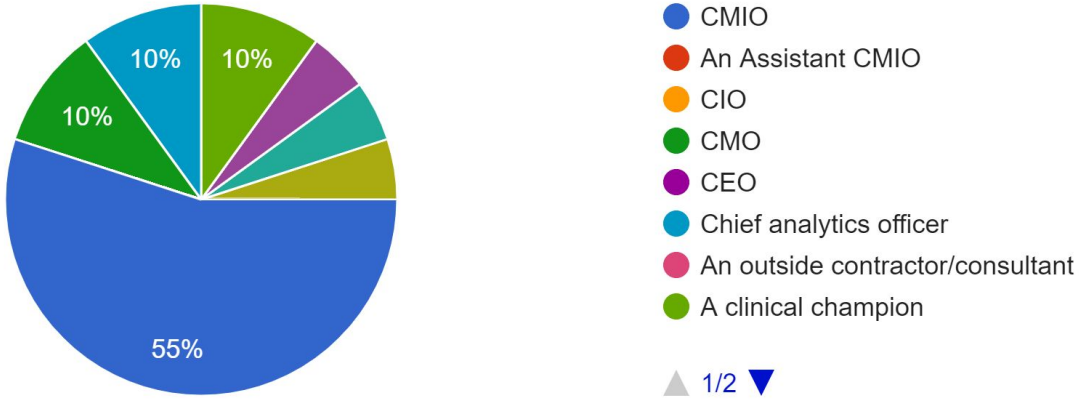
Survey Results



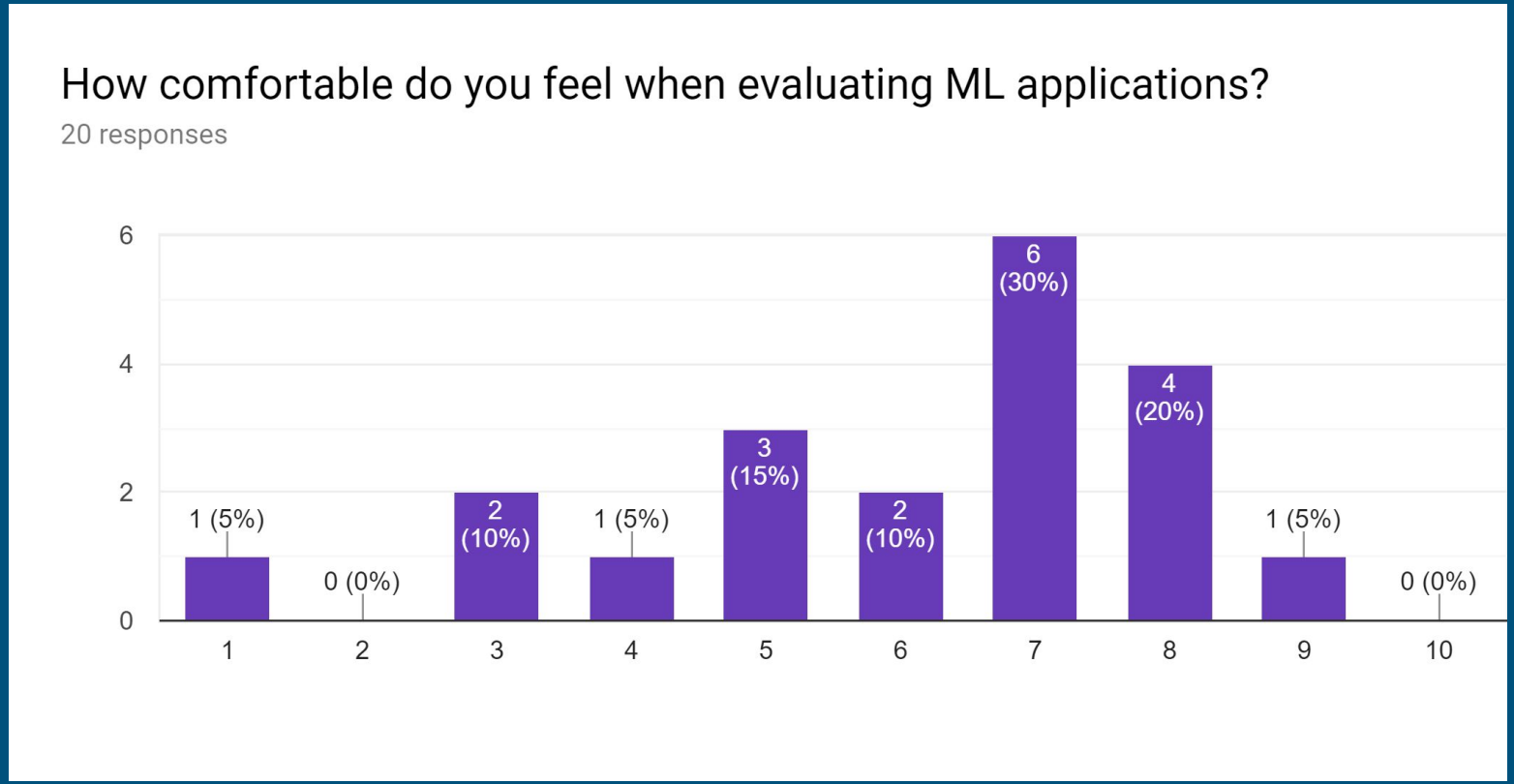
Survey Results

In an ideal world, who in your organization should be most responsible for decisions surrounding ML algorithms in the clinical setting? (Select one)

20 responses



Survey Results



Roadmap

Survey Results

PICO for ML Papers

Discussing Barriers to ML

Roadmap

Survey Results

PICO for ML Papers

Discussing Barriers to ML

sli.do

Join at
slido.com
#Y182

a) Your ambitious first year informatics fellow with a CS degree wants to predict readmissions in patients getting discharged from the hospital. The model he builds will most likely be a _____ model.

Supervised vs Unsupervised Learning

Supervised vs Unsupervised Learning

Supervised - You supply the “right answers” to the model and it learns how to come up these answers.

Supervised vs Unsupervised Learning

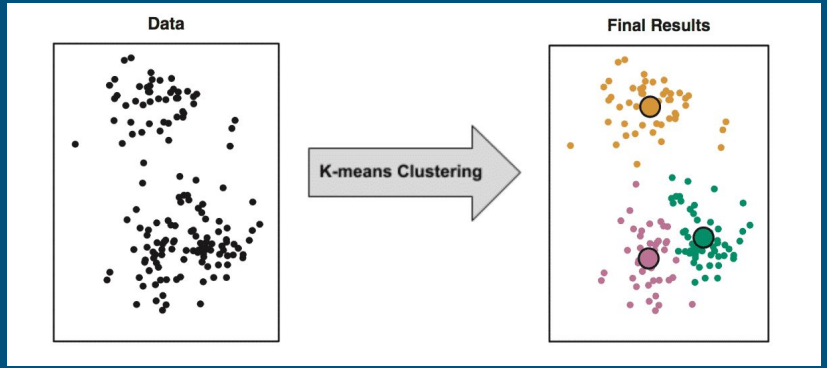
Supervised - You supply the “right answers” to the model and it learns how to come up these answers.

Unsupervised - There are no right answers and the models are trying find “patterns” in the data.

Supervised vs Unsupervised Learning

Supervised - You supply the “right answers” to the model and it learns how to come up these answers.

Unsupervised - There are no right answers and the models are trying find “patterns” in the data.



Supervised vs Unsupervised Learning

Supervised - You supply the “right answers” to the model and it learns how to come up these answers.

Unsupervised - There are no right answers and the models are trying find “patterns” in the data.

Reinforcement - Train the model by “rewarding” it for correct actions

Example: training a robot to grab a ball, training a computer to play chess

A first question

Your ambitious first year informatics fellow with a CS degree wants to predict readmissions in patients getting discharged from the hospital.

The model he builds will most likely be a _____ model.

a) **Supervised -** Your fellow will be labeling which patients count as readmissions.
b) Unsupervised
c) Perfect
d) Imaginary

PICO for Machine Learning

PICO for Machine Learning

P - Patient/Problem

PICO for Machine Learning

P - Patient/Problem

I - Intervention

PICO for Machine Learning

P - Patient/Problem

I - Intervention

C - Comparison

PICO for Machine Learning

P - Patient/Problem

I - Intervention

C - Comparison

O - Outcomes

PICO for Machine Learning

P - Patient/Problem

I - Intervention

C - Comparison

O - Outcomes

PICO for Machine Learning - Patient/Problem

PICO for Machine Learning - Patient/Problem

- Patient Population
  - Does this model apply to my patient population?

PICO for Machine Learning - Patient/Problem

- Patient Population
  - Does this model apply to my patient population?
    - Example: MANY models are build off of the publicly available MIMIC dataset (ICU data from a large hospital in Boston)

PICO for Machine Learning - Patient/Problem

- Patient Population
  - Does this model apply to my patient population?
    - Example: MANY models are build off of the publicly available MIMIC dataset (ICU data from a large hospital in Boston)
- Problem
  - For supervised models: How much do I trust the labelling used to build the model?

PICO for Machine Learning - Patient/Problem

- Patient Population
  - Does this model apply to my patient population?
    - Example: MANY models are build off of the publicly available MIMIC dataset (ICU data from a large hospital in Boston)
- Problem
  - For supervised models: How much do I trust the labelling used to build the model?
    - Example: Some predictive models for sepsis use ICD10 sepsis codes as their “truth” label (this is only 50% sensitive)

A Real Paper!

Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks

Pranav Rajpurkar* PRANAVSR@CS.STANFORD.EDU
Awni Y. Hannun* AWNI@CS.STANFORD.EDU
Masoumeh Haghpanahi MHAGHPANAHI@IRHYTHMTECH.COM
Codie Bourn CBOURN@IRHYTHMTECH.COM
Andrew Y. Ng ANG@CS.STANFORD.EDU

A Real Paper!

Testing

We collect a test set of 336 records from 328 unique patients. For the test set, ground truth annotations for each record were obtained by a committee of three board-certified cardiologists; there are three committees responsible for different splits of the test set. The cardiologists discussed each individual record as a group and came to a consensus labeling. For each record in the test set we also collect 6 individual annotations from cardiologists not participating in the group. This is used to assess performance of the model compared to an individual cardiologist.

How is the “true arrhythmia” being determined in the training set:

a. A single cardiologist read in the EHR
b. Majority from 6 different boarded cardiologists
c. ICD10 codes
d. Consensus committee of 3 boarded cardiologists
e. Medical Students
f. Magic 8 ball

boardcertified cardiologists;

there are three committees responsible for different splits of the test set. The cardiologists discussed each individual record as a group and came to a consensus labeling. For each record in the test set we also collect 6 individual annotations

from cardiologists not

A Real Paper!

Testing

We collect a test set of 336 records from 328 unique patients. For the test set, ground truth annotations for each record were obtained by a committee of three board-certified cardiologists; there are three committees responsible for different splits of the test set. The cardiologists discussed each individual record as a group and came to a consensus labeling. For each record in the test set we also collect 6 individual annotations from cardiologists not participating in the group. This is used to assess performance of the model compared to an individual cardiologist.

How is the “true arrhythmia” being determined in the training set:

a. A single cardiologist read in the EHR
b. Majority from 6 different boarded cardiologists
c. ICD10 codes
d. Consensus committee of 3 boarded cardiologists
e. Medical Students
f. Magic 8 ball

PICO for Machine Learning

P - Patient/Problem

I - Intervention

C - Comparison

O - Outcomes

PICO for Machine Learning

P - Patient/Problem

I - Intervention

C - Comparison

O - Outcomes

PICO for Machine Learning - Intervention

PICO for Machine Learning - Intervention

- Actionable Insight
  - Will knowing a model's predictions change provider/system behavior

PICO for Machine Learning - Intervention

- Actionable Insight
  - Will knowing a model's predictions change provider/system behavior
    - Example: Showing mortality predictions to your discharging doctors.

PICO for Machine Learning - Intervention

- Actionable Insight
  - Will knowing a model's predictions change provider/system behavior
    - Example: Showing mortality predictions to your discharging doctors.
- Integration with workflow
  - Applications need defined ROI and incorporation into workflow.

PICO for Machine Learning - Intervention

- Actionable Insight
  - Will knowing a model's predictions change provider/system behavior
    - Example: Showing mortality predictions to your discharging doctors.
- Integration with workflow
  - Applications need defined ROI and incorporation into workflow.
    - Example: Showing mortality predictions in the discharge activity that then auto-populate a care transitions referral to reduce readmissions.

PICO for Machine Learning

P - Patient/Problem

I - Intervention

C - Comparison

O - Outcomes

PICO for Machine Learning - Comparison

PICO for Machine Learning - Comparison

- Static vs Dynamic
  - Will the model change over time or not?

PICO for Machine Learning - Comparison

- Static vs Dynamic
  - Will the model change over time or not?
    - Static - Easier to implement, but often degrades over time.

PICO for Machine Learning - Comparison

- Static vs Dynamic
  - Will the model change over time or not?
    - Static - Easier to implement, but often degrades over time.
    - Dynamic - Much harder to implement because this needs large data feeds.

PICO for Machine Learning - Comparison

- Static vs Dynamic
  - Will the model change over time or not?
    - Static - Easier to implement, but often degrades over time.
    - Dynamic - Much harder to implement because this needs large data feeds.
    - Consider review periods: Ie every 2-3 years (no consensus)

PICO for Machine Learning - Outcomes

PICO for Machine Learning

P - Patient/Problem

I - Intervention

C - Comparison

O - Outcomes

PICO for Machine Learning

P - Patient/Problem

I - Intervention

C - Comparison

O - Outcomes

PICO for Machine Learning - Outcomes

- AUC

PICO for Machine Learning - Outcomes

- AUC
- Accuracy

PICO for Machine Learning - Outcomes

- AUC
- Accuracy
- Sensitivity/Specificity

PICO for Machine Learning - Outcomes

- Binary Classifiers (Ie sepsis vs no sepsis)
  - AUC
  - Sensitivity/Specificity
  - F1 statistic
  - Accuracy

PICO for Machine Learning - Outcomes

- Binary Classifiers (Ie sepsis vs no sepsis)
  - AUC
  - Sensitivity/Specificity
  - F1 statistic
  - Accuracy
- Non-binary Classifiers (Ie low, mid, high risk for PE)
  - Accuracy
  - Kappa

PICO for Machine Learning - Outcomes

- Binary Classifiers (Ie sepsis vs no sepsis)
  - AUC
  - Sensitivity/Specificity
  - F1 statistic
  - Accuracy
- Non-binary Classifiers (Ie low, intermediate, high risk for PE)
  - Accuracy
  - Kappa
- Regressions (Ie what is the % chance of readmission)
  - R-squared
  - MSE

PICO for Machine Learning - Outcomes

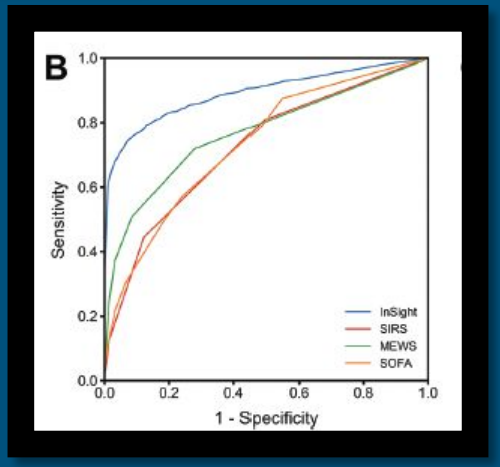
- Binary Classifiers (Ie sepsis vs no sepsis)
  - AUC
  - Sensitivity/Specificity
  - F1 statistic
  - Accuracy
- Non-binary Classifiers (Ie low, mid, high risk for PE)
  - Accuracy
  - Kappa
- Regressions (Ie what is the % chance of readmission)
  - R-squared
  - MSE
- Which metric(s) do you choose?

WHAT ABOUT SEPSIS?

BMJ Open **Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU**

Qingqing Mao,[1] Melissa Jay,[1] Jana L Hoffman,[1] Jacob Calvert,[1]

Receiver Operator Curve for severe sepsis (AUC: 0.87)



WHAT ABOUT SEPSIS?

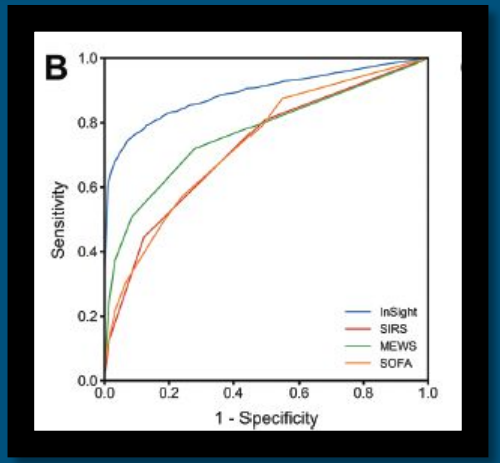
BMJ Open

Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU

Qingqing Mao,[1] Melissa Jay,[1] Jana L Hoffman,[1] Jacob Calvert,[1]

Receiver Operator Curve for severe sepsis (AUC: 0.87)

Sensitivity ~ 80%, Specificity ~80%



WHAT ABOUT SEPSIS?

BMJ Open

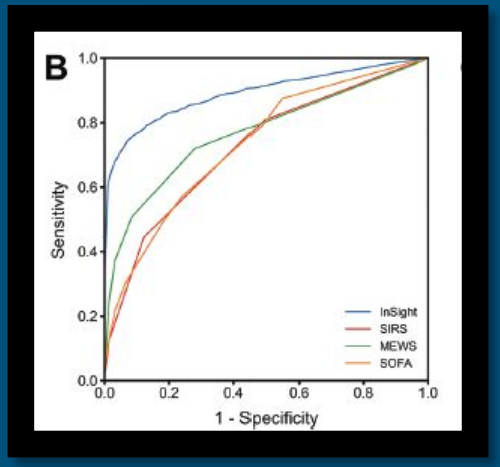
Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU

Qingqing Mao,[1] Melissa Jay,[1] Jana L Hoffman,[1] Jacob Calvert,[1]

Receiver Operator Curve for severe sepsis (AUC: 0.87)

Sensitivity ~ 80%, Specificity ~80%

Assuming prevalence ~10%



WHAT ABOUT SEPSIS?

BMJ Open **Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU**

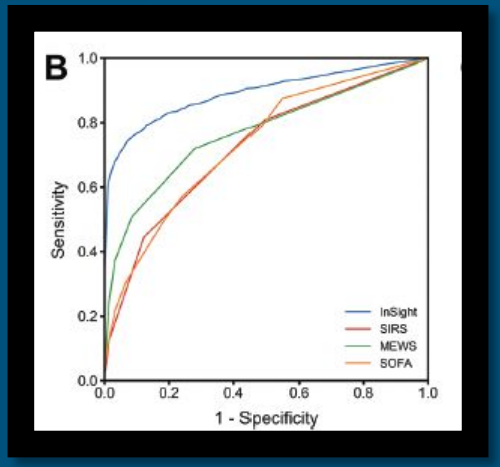
Qingqing Mao,[1] Melissa Jay,[1] Jana L Hoffman,[1] Jacob Calvert,[1]

Receiver Operator Curve for severe sepsis (AUC: 0.87)

Sensitivity ~ 80%, Specificity ~80%

Assuming prevalence ~10%

PPV: 31%



APPLE WATCH

96% sensitive, 98% specific – sounds great!

Prevalence of afib for population under 55 = 0.1%

What do you think is positive predictive value?

a) 1%

b) 5%

c) 20%

d) 50%

e) 90%

sli.do

Join at
slido.com
#Y182

The Apple Watch is 96% sensitive and 98% specific. What do you think its positive predictive value is for the <55 population which has an Afib prevalence of 0.1%.

APPLE WATCH

96% sensitive, 98% specific – sounds great!

Prevalence of afib for population under 55 = 0.1%

What do you think is positive predictive value?

1%

5%

20%

50%

90%

Overall afib prevalence (21+ yrs old): ~1%

Overall PPV: 33%

PICO for Machine Learning - Outcomes

- Binary Classifiers
  - AUC
  - Sensitivity/Specificity
  - F1 statistic
  - Accuracy
- Non-binary Classifiers
  - Accuracy
  - Kappa
- Regressions
  - R-squared
  - MSE
- Which do you choose? **It depends**

PICO for Machine Learning - Outcomes

The right statistic for you depends on your population, use case, and preferences

- In general:
  - Asking for more sensitivity will hurt positive predictive value and specificity.
  - Consider optimizing something like an F1 score (which penalizes false positives).

Roadmap

Survey Results

PICO for ML Papers

Discussing Barriers to ML

Roadmap

Survey Results

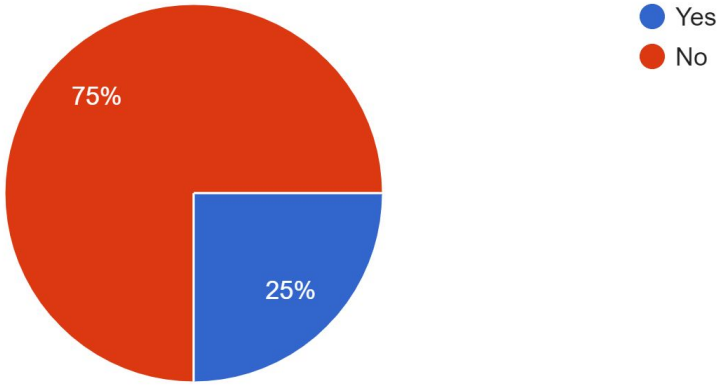
PICO for ML Papers

Discussing Barriers to ML

Survey Results

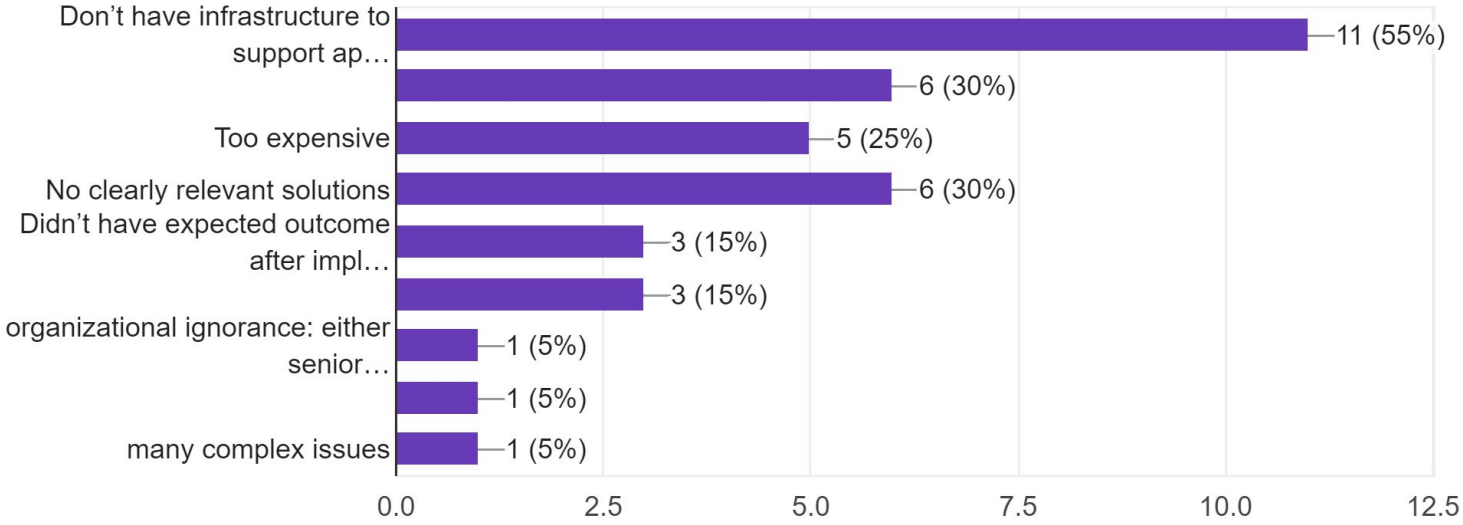
Have you attempted or successfully deployed any machine learning projects? (Select one)

20 responses



What barriers have you faced in implementing ML solutions (Select all that apply)

20 responses



Thank you!