
LECTURE 13-1

Deployment and Human-AI Teaming

What happens after the model is good enough



Where We Are

The last content week: a model meets a hospital.

Week 12

**Evaluation, robustness,
fairness**

What a score does and does
not tell you

Week 13: today

**Deployment and human-
AI teaming**

What happens after the model
is good enough

Weeks 14-15

Project presentations

No lectures or paper sessions

Four Things Training Does Not Decide

Every one of them changes what the model is worth.

Where it is validated

performance at your hospital, not the vendor's

What threshold it runs at

how many patients get reviewed per true case

Whether anyone is watching

the inputs change after go-live

What the clinician does with it

the combination has to be measured



What External Validation Found

The Model and the Study

A proprietary sepsis model, scored every 15 minutes, validated at one health system.

The model

proprietary, deployed at hundreds of hospitals, scored every 15 minutes

The cohort

27,697 patients, 38,455 hospitalizations, Dec 2018 to Oct 2019

The outcome

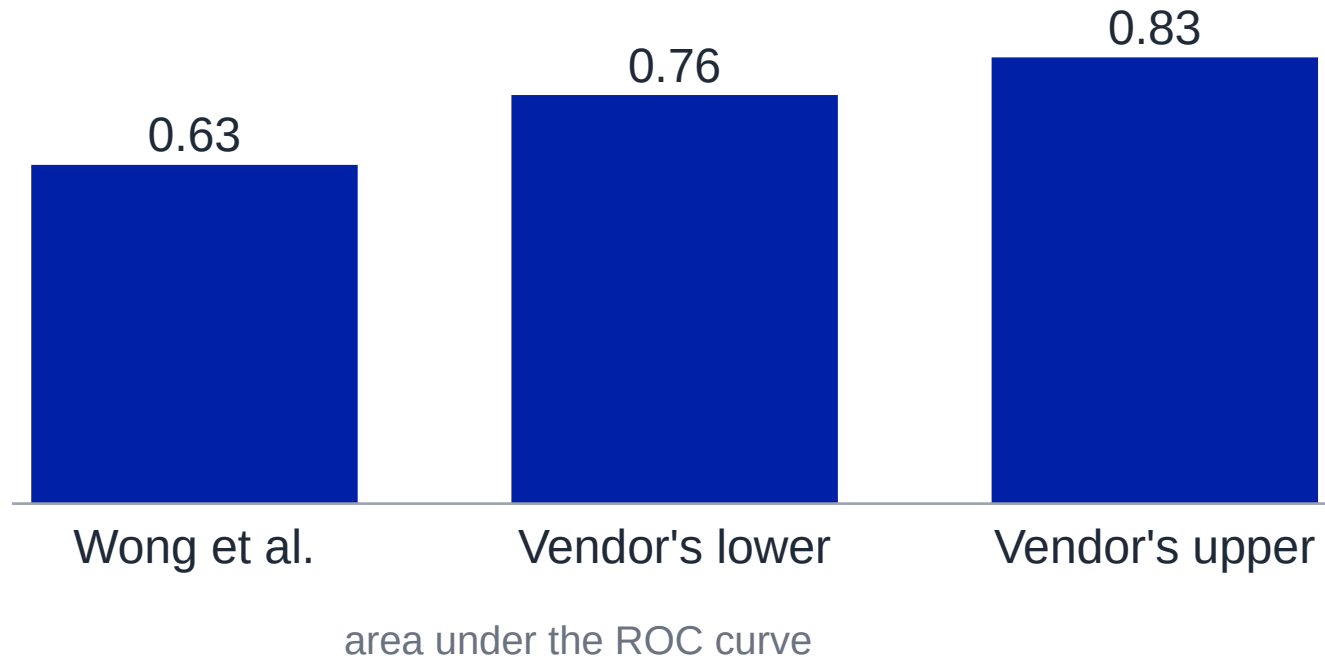
sepsis in 2,552 hospitalizations: 7% prevalence

The comparison

the vendor's own reported performance, from internal documentation

Discrimination

AUC 0.63 at this health system, against the 0.76 to 0.83 the vendor reported.



The objection, answered

per-horizon AUCs are 0.72 to 0.76,
and the authors call them
misleading: the predictions
are not independent

calibration was poor throughout

At the Recommended Threshold

Two thirds of sepsis missed, and alerts on 18% of hospitalizations.

Sensitivity	33%
Specificity	83%
Positive predictive value	12%
Hospitalizations alerted	18% (6,971 of 38,455)
Number needed to evaluate	8

Note

the article carries a
2 August 2021 correction
to this statistic

which is the process
working

The Question That Actually Matters

Not what the model scored, but what the alert added.

2,552

sepsis cases

1,709

never flagged (67%)

1,030

of those got timely antibiotics anyway

183

flagged before antibiotics were given in time (7%)

Median lead time where it did fire: 2.5 hours (IQR 0.5 to 15.6).

Animation: What External Validation Found

Animation clip: week13-clip1.mp4

Silent version embedded (click to play). The narrated version is in the course repository.

What to Take From It

Two rules students should carry into their own projects.

Validate locally

your population, your pipeline

case mix, coding, and ordering
practice all differ

Measure what it adds

compare against current practice,
not against nothing

most studies never make
that comparison



Thresholds, Alert Burden, and Drift

One Model, Three Operating Points

At 7% prevalence, the threshold decides the alert load.

Sensitivity	PPV	Alerts per 100 admissions	Patients reviewed per true case
80%	10%	56.0	10
50%	20%	17.5	5
25%	35%	5.0	2.9

The sepsis model at its recommended threshold sits near the first row: 8 reviewed per true case.

Illustrative operating points at 7% prevalence, as in clip 2.

Rarity Governs the Alert Burden

A much better model lands in the same place.

	Deployed model (Wong et al.)	Notebook model
AUC	0.63	0.78
Sensitivity at the chosen point	33%	82%
Positive predictive value	12%	13%
Patients reviewed per true case	8	8.0

A 15-point gain in AUC bought no reduction in the review burden.

Nothing Holds Still

A new assay, a coding change, a different case mix.

Inputs

a new lab assay, a changed unit, a renamed field

Coding

documentation and billing practice change

Population

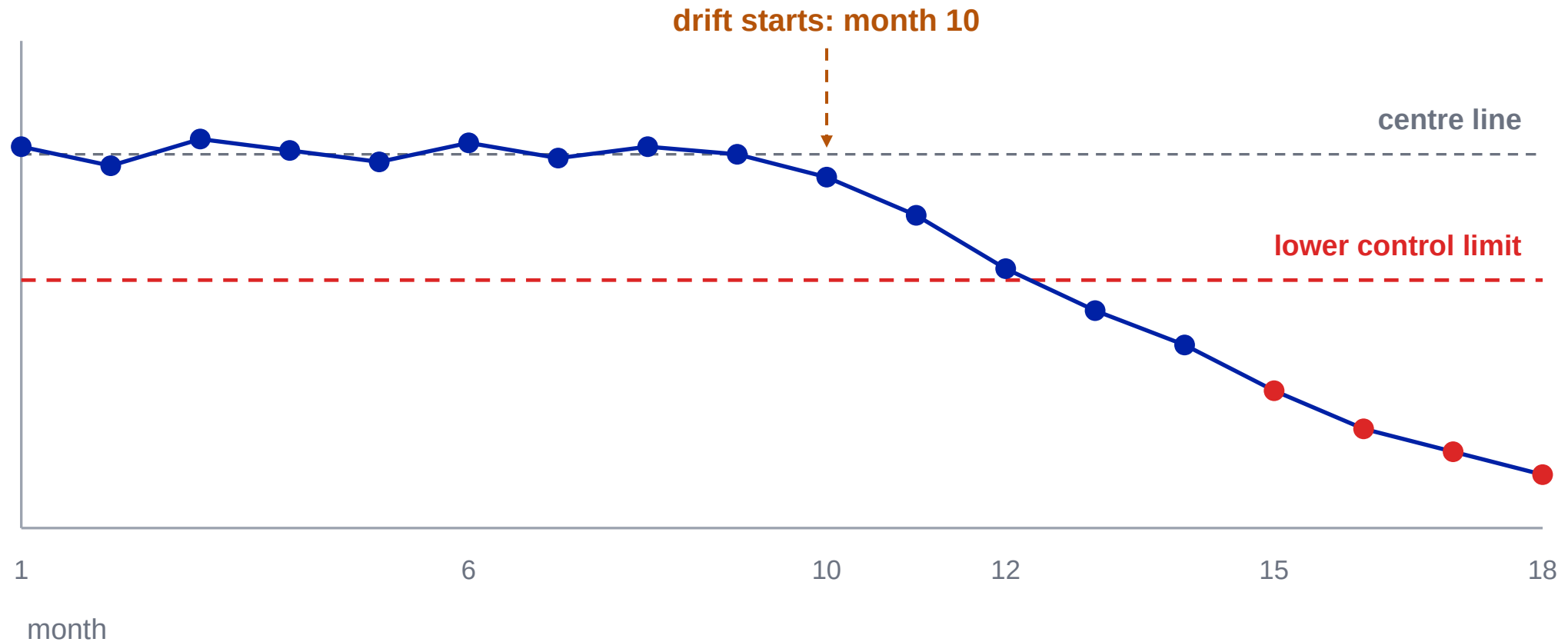
a new service line, a new season, a new referral pattern

The model

keeps producing scores, with no error of any kind

Watch It: A Control Chart

Monthly performance against limits set from a stable period.



Monitoring Buys a Dated Signal, Not a Solution

Three-sigma fires in month 15, five months late.

Drift starts

month 10

Run-of-three rule fires

month 12, two months of exposure

Three-sigma rule fires

month 15, five months of exposure

Write the monitoring plan before go-live: metrics, limits, who looks, and what happens when a limit is crossed.

Animation: Thresholds, Alert Burden, and Drift

Animation clip: week13-clip2.mp4

Silent version embedded (click to play). The narrated version is in the course repository.



Humans and Models Together

Two Obvious Rules

Flag when either flags, or flag only when both do.

Flag if either flags

catches more
alarms more

the usual choice when a miss
is worse than a false alarm

Flag only if both flag

alarms less
misses more

the usual choice when alert
fatigue is the problem

Clinician sensitivity 0.80, model 0.75 (illustrative).

Same Two Sensitivities, Opposite Answers

What changes is whether their mistakes coincide.

	Errors independent	Misses coincide
Flag if either flags	0.95	0.80 (the clinician alone)
Flag only if both flag	0.60	0.75 (the model alone)

The whole result lives in a quantity you cannot compute from the two sensitivities.

The notebook sweeps the correlation: either-flags falls from 0.951 to 0.800, both-flag rises from 0.603 to 0.750.

And People Are Not Rules

Automation bias, alert fatigue, and a trial that found no benefit.

Automation bias

a confident output is trusted more than it has earned

Alert fatigue

alerts that are usually wrong stop being read at all

Measured, not assumed

Goh et al. (Week 12) found no benefit from the combination

Ambient Intelligence

Sensing the care space instead of waiting to be told.

Hand hygiene (Singh et al., 2020)

16 depth sensors, acute care unit

algorithm: 92.1% sensitive

human observers: 85.2%

test set: 88 dispenser-use events

ICU mobility (Yeung et al., 2019)

7 depth sensors, one hospital

563 mobility events

activity detected: 87.2% sensitivity

staff counted correctly: 68.8%

The weak numbers are the teaching: 88 events, and a 69% count.

88 Events Is Not Many

The bootstrap from Week 12, applied to the hand-hygiene claim.

Algorithm, 95% CI 0.864 to 0.977



Human observers, 95% CI 0.773 to 0.920



0.70

1.00

Difference: 0.000 to 0.136. Ahead in 96.7% of resamples.

Animation: Humans and Models Together

Animation clip: week13-clip3.mp4

Silent version embedded (click to play). The narrated version is in the course repository.

Governance Before Technique

Sensing a care space is a consent question first.

Who is observed

patients and staff, and staff are rarely asked

What is kept

depth rather than images helps, and is not the whole answer

What it is used for

a record collected for care can be read for other purposes

What the review names

clinical validation, data privacy, model transparency

Recap: After the Model Is Good Enough

Validate

Locally, and against current practice. AUC 0.63, and 7% of cases helped.

Threshold

Rarity governs the burden: AUC 0.78 still gave 8 reviewed per case.

Monitor

A drift from month 10 was caught in month 15. Write the plan first.

Team

0.95 or 0.80 from the same two sensitivities. Measure the pair.

Before Thursday, and Then the Projects

Read

Wong et al., External validation of a sepsis model, and Haque et al., Ambient intelligence. The last paper sessions.

Write

Reading responses on Canvas before 1:00 PM Thursday.

Explore

Week 13 reference notebook (ungraded): the threshold sweep, a drift nobody catches in time, teaming, and 88 events.

Then

Presentations Apr 13, 15 and 20: every team's slides are due 1:00 PM Apr 13. Paper and repository due Apr 21, 11:59 PM.