The Value of Clinical Review Labels Over Billing Codes

Evaluating Feature Sets and Training Labels for Retrospective ARDS Identification

Joshua S. Goller[1,2], Paul M. Heider, PhD[2], Charles R. Terry, MD, MSCR[2]
[1]College of Charleston, Charleston, SC, USA
[2]Medical University of South Carolina, Charleston, SC, USA

Introduction

- Acute respiratory distress syndrome (ARDS):
  - strikes about 1 in 10 ICU patients
  - kills about 40%
  - goes unrecognized in about 40% of cases.[1]
- ML can catch missed cases, but models only learn what labels teach.
- Two label sources:
  - Gold standard: provider chart review (accurate but scarce)
  - Silver standard: hospital billing codes (plentiful but noisy)
- Our questions:
  - How badly do billing codes undercount ARDS?
  - Can more billing data close the gap?

Methods

- **Model**: Gradient-boosted trees
- **Features**: physiology + chest X-ray report concepts
- **Label quality Test**:
  - 443 ventilated MUSC patients
  - Chart Labels: MUSC physician review (Berlin Definition)
  - Billing Labels: ARDS binary flag (based on ICD-9/10)
  - 5-fold CV, bootstrap 95% CIs
- **Scaling Test**:
  - Train: MIMIC-III[2], ~23,000 stays, billing labels
  - Test: Morales[3] (same hospital, n=100) and MUSC (n=443)
- **Simulation:** vary label quality and training size to test whether volume closes the gap.

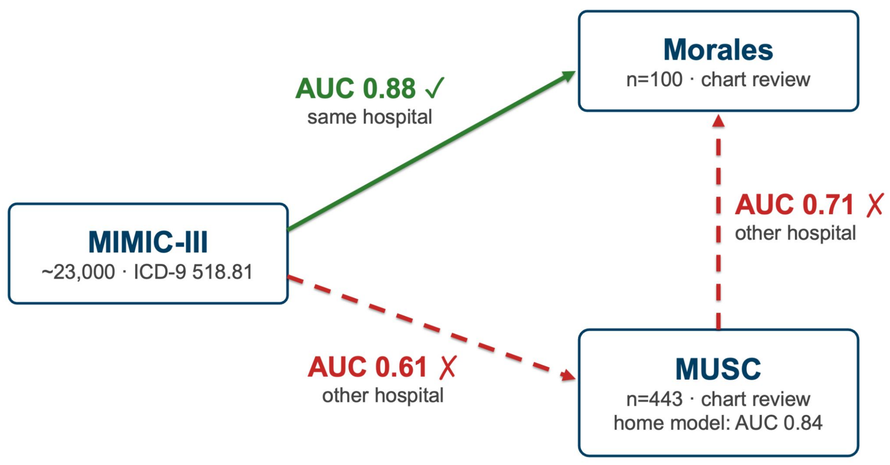
Results

Label Quality Test: Models inherit billing-code ARDS misses

Training Label	Eval Label	AUC	Sens.	Spec.	Prec.
Chart	Chart	0.84	0.76	0.71	0.72
ICD	ICD	0.73	0.41	0.89	0.61
ICD	Chart	0.74	0.29	0.92	0.79

Billing codes flag only 44% of chart-confirmed ARDS (κ = 0.27)

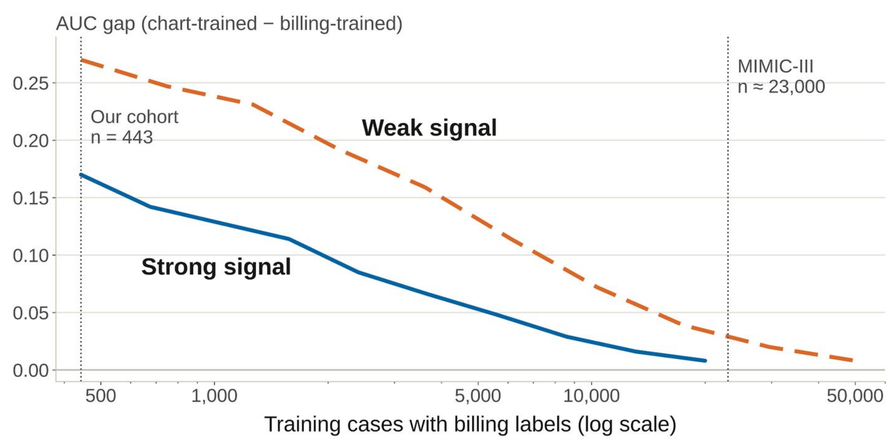
Scaling Test: works within a hospital, fails across hospitals



- Arrow = train → test
- AUC against chart-review labels
- Morales ⊂ MIMIC-III

Bonus Content: Simulation

- Simulated patients, true ARDS status known
- Labels corrupted to billing-code error (κ = 0.27)
- Train on true vs. corrupted labels
- Goal: Measure the AUC gap as training size grows



Takeaway: More data → smaller gap

- Strong signal: gap ≈ 0 by ~20,000 cases
- Weak signal: ~50,000 cases

Conclusions

- In this experiment:
  - Billing codes miss most ARDS. Models trained on them inherit the miss (sensitivity 0.76 → 0.29)
  - Chart-trained model ≈ published detectors (AUC 0.84 vs. 0.79–0.83)[4,5]
  - More billing data closes the gap, within one hospital
  - Across hospitals, it fails → site effects, not just label noise
  - Billing-trained model looked fine on billing labels (AUC 0.73) → must validate against chart review, not billing codes.
- Implication:
  - Billing-defined ARDS cohorts in similar work likely undercount

Future Work

- **Context-aware NLP:** detect negated, uncertain and one-sided findings in chest X-ray reports. “No bilateral opacities” currently counts the same as “bilateral opacities.”
- **Cleaner physiology:** prune noisy measurements and re-test whether vital-sign and ventilator data add signal beyond the notes.
- **Explain the transfer failure:** use feature attributions (SHAP) to find which report phrases reflect site-specific dictation habits.
- **A second chart-labeled site:** to separate label bias from site effects.

Citations

1. Bellani G, et al. JAMA. 2016;315(8):788-800.
2. Johnson AEW, et al. Sci Data. 2016;3:160035.
3. Morales FL, et al. Nat Commun. 2025;16:6787.
4. Afshar M, et al. AMIA Annu Symp Proc. 2018;2018:157-165.
5. Mayampurath A, et al. Crit Care Med. 2020;48(9):e791-e798.

Acknowledgements

We thank D. Weinberg and J. Meng for the chart-reviewed MUSC cohort, and the Amaral lab (Northwestern) for publicly releasing their adjudicated MIMIC-III labels. Funding: Research reported in this publication was supported in part by the National Center for Advancing Translational Sciences of the National Institutes of Health (NIH) under award number **K12TR005297** to the South Carolina Clinical and Translational Research (SCTR) Institute. The authors declare no conflicts of interest.

More Information

Joshua Goller
goller@musc.edu
jsgoller1.github.io/docs/amia_2026_poster.pdf