Multi-tier ground truth elicitation framework with application to artifact classification for predicting patient instability

Carnegie Mellon Auton Lab

University of Pittsburgh

Donghan Wang, Lujie Chen, Madalina Fiterau, Artur Dubrawski (Carnegie Mellon University), and Marilyn Hravnak, Eliezer Bose, David Wallace, Murat Kaynar, Gilles Clermont, Michael R. Pinsky (University of Pittsburgh)

Context:

- Robust health monitoring systems would **identify artifactual alerts** (due to equipment malfunction or misapplication of sensors) and **distinguish them from true alerts** signifying actual patient's instability
- This capability is important in **reducing alert fatigue** in clinical personnel

Our approach:

- We prototyped a **protocol to collect reliable training data** for development of such robust algorithms
- It uses **Active Machine Learning** (AML) for multi-tier **alert annotation elicitation** at the **minimal effort from a team of expert clinicians**

Methodology

Framework:

- **Machine Learning (ML)** is a powerful methodology which has been successfully applied to handle complexities of bedside data
- It can be used to e.g. discriminate true patient instability from artifact in hemodynamic monitoring data
- ML relies on a **library of expert-annotated examples** to **establish ground truth**
- Collecting expert-annotations can be a **burden on expert clinicians**

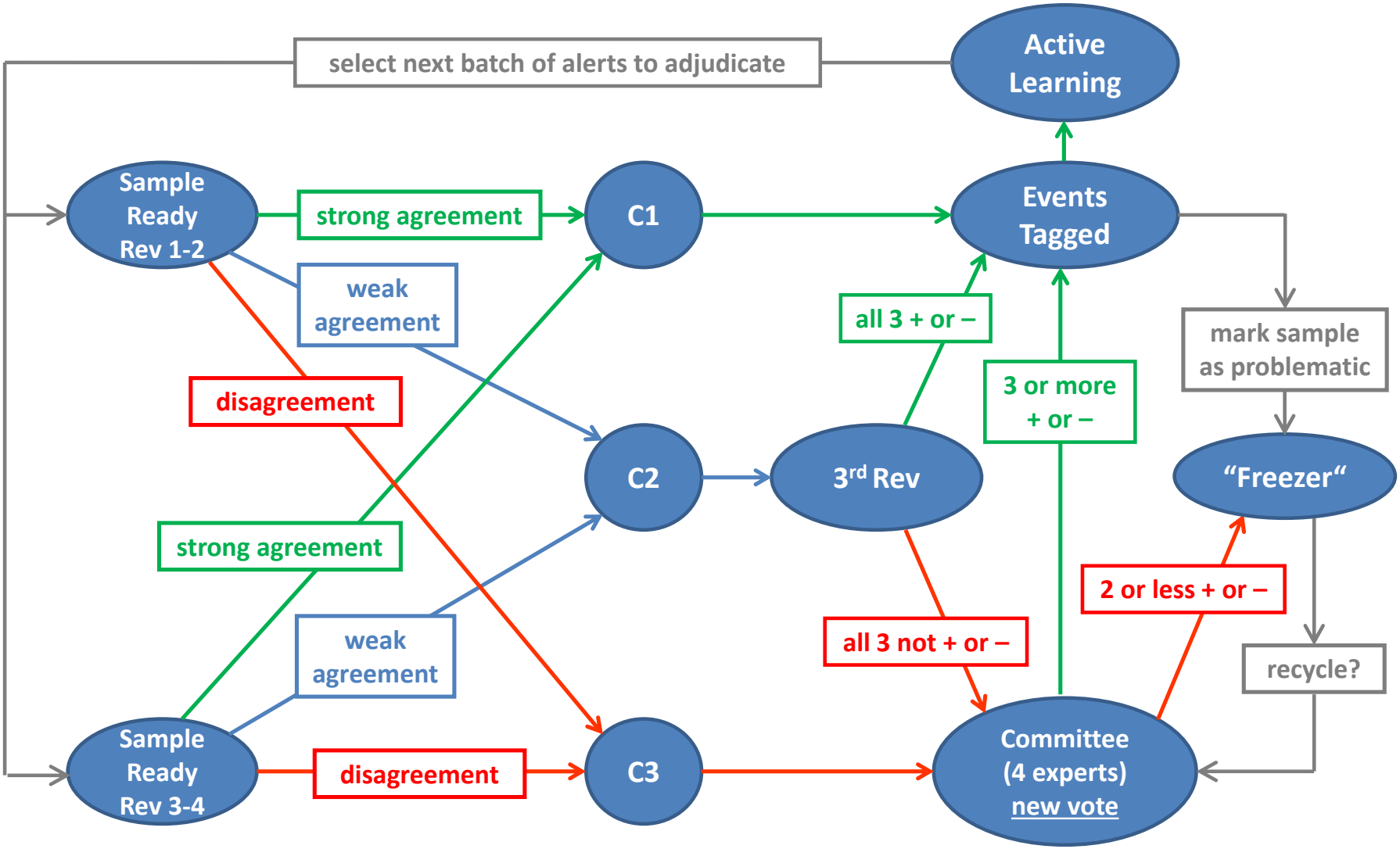
Specific Aim:

- **Develop a protocol for collecting expert annotations that minimizes clinicians' effort while providing informative training data for ML algorithms**
- Methodology: Active Machine Learning (AMR)

Methodology (contniued)

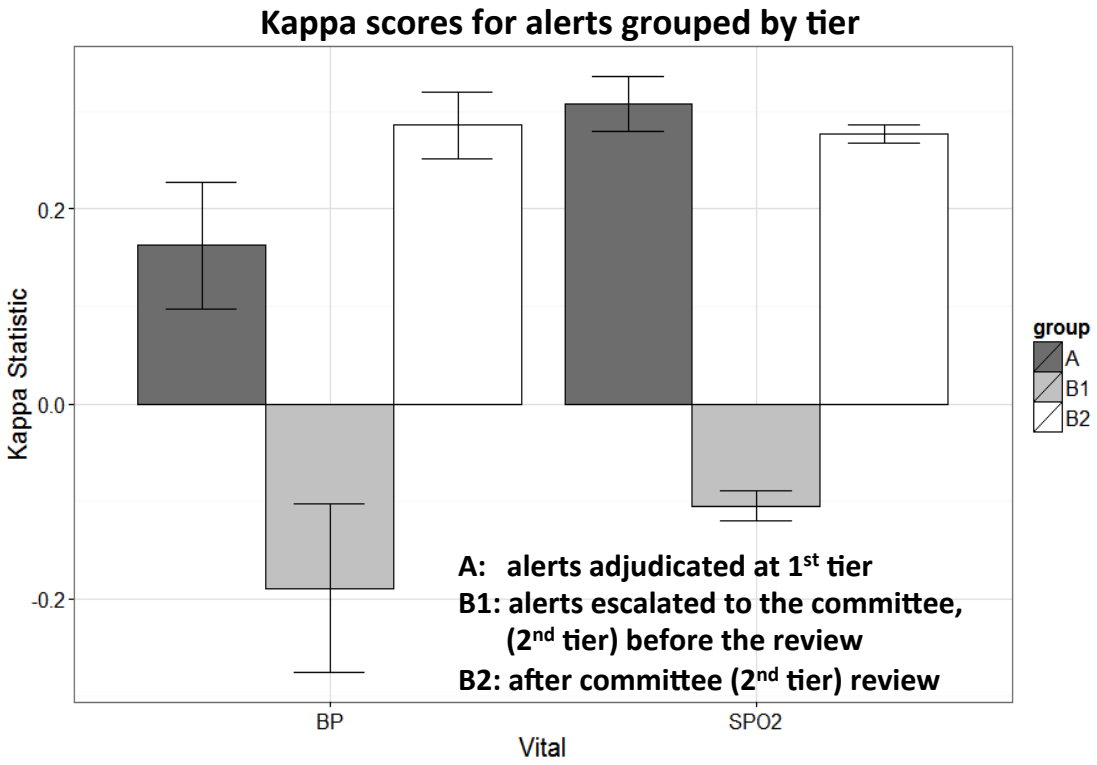
- **The data**, collected from **noninvasive monitoring system** for ECG-derived heart rate (HR) respiratory rate (RR), pulse oximetry (SpO_2), systolic and diastolic blood pressure (BP).
- Data contains **1,582 alert periods** defined as exceedences over pre-set stability thresholds (HR<40 or >140, RR<8 or >36, systolic BP <80 or >200, diastolic BP>110, SpO_2<85%).
- AML **algorithm selects a batch of alerts** which, when adjudicated by experts, would **provide the most useful information** to the alert interpretation algorithm.
- **Reviewers** working individually **score these alerts** on a scale from -3 (artifact) to 3 (true alert) to reflect confidence of assessment.
- Alerts yielding **substantial disagreement or low confidence** are pushed to a **committee review** where consensus can be reached.
- The alerts that cannot be agreed upon are put in a "freezer" and not used in the alert adjudication model building.

Schematic diagram of the multi-tier protocol



Results

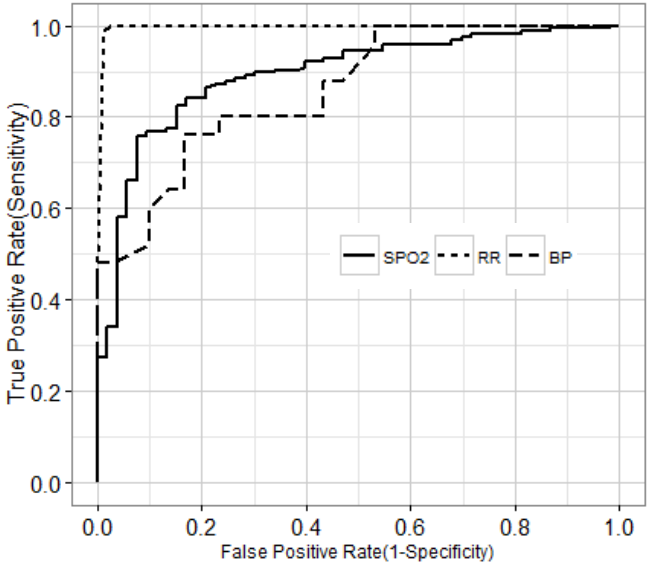
- We **collected 1,941 annotations from 7 expert clinician reviewers on 450 unique alerts** (HR: 60, BP: 80, RR: 80, SpO_2: 230).
- Of those, 26 (32.5%) BP, 115 (50%) SpO_2, and almost none of HR and RR alerts, were **escalated to the 2nd tier review**.
- The results show that the **consensus for alerts initially conflicted improved significantly** as a result of the 2nd tier committee review.
- Weighted pairwise Kappa statistic increases from -0.19 to 0.29 for BP, and from -0.10 to 0.28 for SpO_2 alerts.

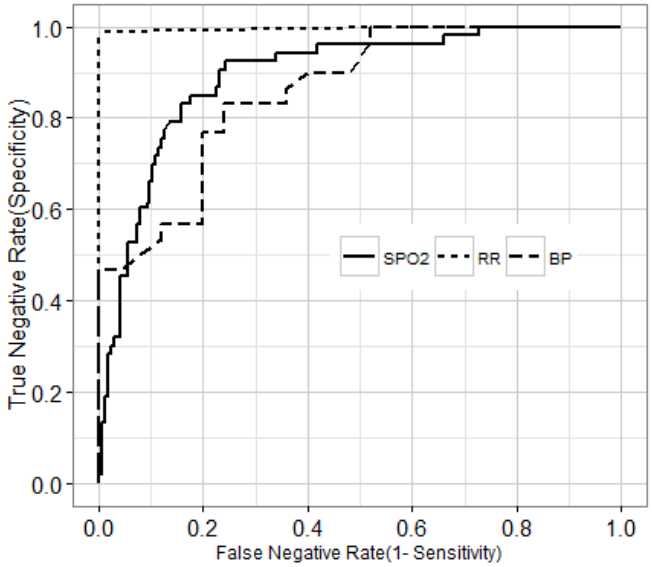


Results (continued)

Preliminary **artifact adjudication model** built using these annotations shown:

- **Very strong ability to identify RR alerts and artifacts** (dense dash line in the ROC diagrams)
- **Ability to very confidently isolate** more than 47% of **true BP alerts** and more than 45% of **BP artifacts**
- **Very good performance in isolating SpO_2** artifact, equivalent to **what can be attained with 50% more annotated training data** if the Active Machine Learning protocol has not been used.





Conclusion

- We implemented a **multi-tier framework** to **elicit ground truth from multiple reviewers** to support development of a prototype of the automated artifact adjudication system.
- The initial results show that **precious human expertize can be utilized efficiently and without loss of performance** of the resulting models of instability.
- The proposed annotation framework can **yield accurate alert adjudication systems while minimizing effort of human experts** required to produce ground truth evidence, **even if very large libraries of reference data are available**.

This work has been partially supported by the NSF Awards 0911032 and 1320347, and by the NIH Grant R01NR013912.