**METHODS**

**D. Expert Annotation Workflow**

Each alert is initially reviewed by two experts who classify it as genuine or artifact with some level of confidence. A strong expert agreement: if the confidence of the reviewers is high, the label is added to the repository, whereas the artifact disagreement: if the reviewers disagree, the decision regarding the alert is made by a committee of 4 experts.

**E. Training Procedure using Active Learning**

We used a method derived from (Fiterau, 2012): the committee of 4 experts discusses the alerts on which a disagreement exists. If the initial reviewers disagree, the alert is given directly to the committee.

**RESULTS**

The proportion of alerts escalated to the 2nd tier review was 26% (22.5% BP, 11% (60%) SpO2. Almost all of HR and RR alerts could be adjudicated in the first tier.

The results show that the consensus for alerts initially conflicted improved significantly as a result of the 2nd tier committee review. Weighed pairwise Kappa statistic increases from -0.10 to 0.28 for BP, and from -0.10 to 0.28 for SpO2 alerts.

**CONCLUSIONS**

We implemented a multi-tier framework to elicit ground truth from multiple reviewers to support development of a prototype of the artifact adjudication system. The initial results show that precious human expertize can be utilized efficiently and without loss of performance of the resulting model for alerting artifacts. 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.