We sincerely thank all the reviewers for their encouraging and constructive comments, especially during a difficult time. We are pleased that they found the paper well written and acknowledged the novelty and potential significance of the proposed learning framework as "... potentially impactful learning framework for multi-organization ML" (R2) and "can be quite significant when multiple agencies ..." (R4). We will incorporate all the comments into the revised paper.

**Reviewer#1** *Assumptions on label/response and identifier.* We will clarify in the revision that in assisted learning, each participant can initiate and contribute to a task regardless of whether it accesses labels or not. For example, in the task of MIMIC3 (Sec. 3, Fig. 1b), the in-hospital lab aims to predict Length Of Stay. The out-side lab doesn't have access to in-hospital's private labels, but it could still initiate and provide assistance to the in-hospital lab by fitting the received residuals instead of public labels. In contrast, a prerequisite for ensemble methods to work in multi-organization learning is that all participants (including the outside laboratory) have access to labels in order to train ensemble elements. As suggested, we will discuss more on the assumption of the identifiers needed to collate data. *Evidence that assisted learning can outperform stacking.* Assuming that labels are public to each participant, we found that Reviewer 1's suggestion on comparing assisted learning and ensemble approaches such as stacking theoretically and practically intriguing. In some cases, we were able to prove that assisted learning significantly outperforms stacking. An example case is where each participant holds a disjoint subset of the features and uses linear regression. The linear coefficients of Alice's features in a stacking model would be determined by her linear space, which may not be proportional to the linear coefficients of Alice's features in an oracle model determined by all the participants' joint linear space. We also did extensive experiments in the last few days to provide empirical evidence (summarized in **Table 1**). The results show that assisted learning often outperforms stacking (under the same settings as in the paper). *Ways for combining predictions.* We used unweighted sum in Procedure 1. We briefly discussed some sophisticated extensions on page 6, e.g., to exchange and summarize information on stratified data (using categorical variables). We plan to publish a website containing open-source software APIs for a list of case studies and extensions for the work.

**Table 1:** Prediction performances of stacking and assisted learning (LR: linear regression, RF: random forest, GB: gradient boosting, NN: neural network, and RG: ridge regression). The settings are the same as in the paper. Column 1 means: in assisted learning, participants use LR; in stacking, participants produce features using LR, and the meta-model combining them is LR. Columns 2-4 were similarly defined. Column 5 means: in assisted learning, each participant at each round performs model selection to choose from three models; in stacking, each participant produces three features using all the models and ensembled using RG. Standard errors are within 0.05 over 50 replications.

| Data | Friedman | | | | | MIMIC3 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Base model(s) | LR | RF | GB | GB | LR+RF+GB | LR | RF | GB | GB | LR+RF+GB |
| Model for stacking | LR | RF | GB | NN | RG | LR | RF | GB | NN | RG |
| Stacking | 2.63 | 1.80 | 1.76 | 1.68 | 1.60 | 121.7 | 118.1 | 119.3 | 119.7 | 115.9 |
| Assisted learning | 2.64 | 1.31 | 1.23 | 1.23 | 1.25 | 120.5 | 109.7 | 111.3 | 111.3 | 110.8 |

**Reviewer#2** *Alignment of data & Evaluation size.* Each participant needs to hold and distribute identifiers for data items, so that the data from different participants can be conceptually combined. Interesting future work includes assisted learning under partially-aligned identifiers (e.g. timestamps), and robustness against a portion of misalignment. We are developing open-source APIs for large-scale deployment on cloud platforms. Upon the acceptance of this work, we will provide assisted learning services to numerous organizations and further evaluate the framework at larger scales. *Relationship with gradient boosting.* The process of sequentially building models and combining their predictions in assisted learning is similar to that in Boosting methods. However, in Boosting, each model is built based on the same dataset with different sample weights. In contrast, assisted learning uses side-information from heterogeneous data sources to improve the performance of a particular learner. The relationship will be discussed in our revision.

**Reviewer#3** *Related work.* We will cite ResNet and more related work such as ensemble methods in the revision. *Evaluation of stop criterion.* We evaluated the proposed stop criterion in all the experimental studies, and we find that the current criterion (based on cross-validation) leads to a near-optimal stopping number. The probability of choosing the optimum could be theoretically derived from large-deviation bounds (under certain assumptions). Future work includes the study of time-dependent data (so backtesting has to be considered). We will discuss these in the revision. *More experiments.* We will include more comparisons with stacking methods (preliminary results in Table 1).

**Reviewer#4** *Feature splitting.* Feature splitting in the paper is categorized into two types. For real data such as the MIMIC3, the data was naturally split according to data-generating modules (namely hospital divisions). For real data without knowing the data-generating process (Superconductor in Appendix) and synthetic data, we randomly split the features. We reported detailed feature splitting for each task in the paper and included scripts in the supplement to reproduce the results. We will further test on data with complex structure and incorporate the results in the revision. *Pathetic scenarios.* From our experimental studies, if Bob's data is not relevant (e.g., purely noisy or purposefully shuffled) to Alice's task, then Alice will observe unimproved or even degraded performance in the learning or/and prediction stages. If Bob is exactly the same as Alice, whether assisted learning will approach oracle depends on the underlying data and model. We will include more experiments and discussions on the pathetic scenarios in the revision.