Sequential algorithmic modification with test data reuse: Supplementary material

Jean Feng[1] **Gene Pennello**[2] **Nicholas Petrick**[2] **Berkman Sahiner**[2] **Romain Pirracchio**[3] **Alexej Gossmann**[2]

[1]Department of Epidemiology and Biostatistics, University of California, San Francisco
[2]U.S. Food and Drug Administration
[3]Department of Anesthesiology, University of California, San Francisco

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

A PROOFS

Lemma 1. *The adaptive SRGP in Algorithm 1 with a fixed strategy is equivalent to a prespecified SRGP.*

Proof. We define a filtration over approval histories up to the maximum number of iterations T. That is, define sample space Ω as the set of approval histories over T iterations, i.e. $\Omega = \{0,1\}^{T-1}$, and σ-algebras $\mathcal{F}_t$ for $t = 1, ..., T$ over approval histories up to iteration t. To show that the adaptive SRGP is equivalent to a prespecified SRGP, we need to show that that the adaptive procedure defines a set of hypotheses, node weights, and edge weights for the initial set of hypotheses I_0, the hypotheses and weights are $\mathcal{F}_1$-measurable functions, and the weight constraints are satisfied. First, we note that the edge weights being elicited at iteration t in Algorithm 1 is equivalent to eliciting the edge weights for the initial set of hypotheses I_0, i.e. $g_{a_{t'},a_t} = g_{a_{t'},a_t}(I_0)$ in Algorithm 1. This is because we only elicit the edge weight $g_{a_{\tau_t},a_t}$ if there has been no approval since time τ_t so the edge weights being elicited are never updated via the edge-weight renormalization step in SRGPs. As such, the adaptive SRGP in Algorithm 1 for a model developer with a fixed strategy for selecting hypotheses and weights can be described to have a fixed hypothesis testing tree structure with

- $\mathcal{F}_t$-measurable hypotheses $H_{a_t}(I_0)$ for all a_t
- $\mathcal{F}_1$-measurable node weights $w_{a_t}(I_0)$ for all $a_t \in \{0,1\}^{T-1}$ that satisfy the constraint that they sum to one,
- and $\mathcal{F}_t$-measurable edge weights $g_{a_{t'},a_t}(I_0)$ for all valid edges $(a_{t'}, a_t)$ in the graph that satisfy the constraint that all outgoing edge weights sum to one.

Although the hypotheses and edge weights are $\mathcal{F}_t$-measurable, they can also be viewed as $\mathcal{F}_1$-measurable functions over the input space a_t and $(a_{t'}, a_t)$, respectively. Moreover, the edge weights satisfy the edge weight constraints by design. Thus the adaptive SRGP satisfies the node and edge weights constraints with respect to $\mathcal{F}_1$.

□

Lemma 2. *If the adaptive SRGP in Algorithm 1 controls the FWER for any fixed strategy, then the adaptive SRGP in Algorithm 1 controls the FWER for any stochastic strategy.*

Proof. Let $\mathcal{S}$ be the set of all fixed strategies. The stochastic adaptive strategy is a random distribution over $\mathcal{S}$. Its FWER is

$$\Pr\left(\text{incorrectly reject some } H_t^{\text{adapt}}\right) = \sum_{s\in\mathcal{S}} \Pr(S=s) \Pr\left(\text{incorrectly reject some } H_t^{\text{adapt}} \mid S=s\right)$$

where the latter probability on the right hand side is the FWER for a fixed strategy s. As such, the FWER of the stochastic strategy is properly controlled as long as the FWER of any fixed strategy is properly controlled. □

Corollary 1. *Algorithm 1 with the significance thresholds defined per*

$$c_{a_t}(I_t) = w_{a_t}(I_t)\alpha \tag{S.1}$$

controls the FWER at level α.

Proof. Per Lemmas 1 and 2, it suffices to show that the fully prespecified SRGP controls the FWER. Recall that (S.1) is a closed weighted Bonferroni test in Bretz et al. (2011). As such, any fixed or stochastic adaptive strategy would control FWER. □

Proof for Theorem 1. Per Lemmas 1 and 2, it suffices to show that the fully prespecified SRGP controls the FWER.

First, per the proof in Bretz et al. (2009), we note that node weights for any intersection hypothesis I calculated using Algorithm 1 are well-defined, in that it does not depend on ordering in which we remove nodes from the graph.

We begin with proving that for any intersection hypothesis I, the critical values calculated using (4) controls the Type I error. First we show that for any a_t ending with success (i.e. $a_{t,t-1} = 1$) and any I, the calculated critical values for testing the

intersection hypotheses $\bigcap_{a_k \in G_{a_t} \cap I} H_{a_k}$ controls the Type I error at level $\left(\sum_{a_k \in G_{a_t} \cap I} w_{a_k}(I)\right) \alpha$. Per the definition of the critical values in (4), we have that

$$\begin{aligned}
&\Pr\left(\text{we reject for some } a_j \in G_{a_t} \cap I \,\middle|\, H_{G_{a_t} \cap I}\right) \\
&= \sum_{a_j \in (G_{a_t} \cap I)} \Pr\left(p_{a_k} > c_{a_k}(I) \forall a_k \in G_{a_t} \cap I, k < j, p_{a_j} < c_{a_j}(I) \,\middle|\, \bigcap_{a_k \in G_{a_t} \cap I, k \leq j} H_{a_k}\right) \\
&\leq \left(\sum_{a_j \in (G_{a_t} \cap I)} w_{a_j}(I)\right) \alpha.
\end{aligned}$$

Therefore, as long as the total node weight across I is no more than one, we control the Type I error at level α. Because Type I error control holds for all intersection hypotheses I, we have established that this procedure is a valid closed test.

Next, per the proof in Bretz et al. (2009), we must show that the critical values satisfy the monotonicity condition to prove that our procedure is a valid consonant, shortcut procedure. More specifically, we require the following to hold for all $t = 1, ..., T$:

$$c_{a_t}(I) < c_{a_t}(J) \qquad \forall J \subseteq I. \tag{S.2}$$

The proof is by induction. It is easy to see that (S.2) holds for $t = 1$. Suppose (S.2) holds for $1, ..., t-1$. Now consider any history $a_{\tilde{t}}$ that ends with an approval. Consider any a_t and subset $J \subseteq I$ such that $a_t \in G_{a_{\tilde{t}}} \cap J$. We have that

$$\begin{aligned}
&c_{a_t}(J) \\
&= \sup\left\{\tilde{c} : \Pr\left(p_{a_k} > c_{a_k}(J) \forall a_k \in K, p_t < \tilde{c} | H_{K \cup \{a_t\}}\right) \leq \left[\sum_{\substack{a_k \in ((G_{a_{\tilde{t}}} \cap J) \setminus K) \\ k \leq t}} w_{a_k}(J)\right] \alpha \forall K \subseteq \{a_k : a_k \in G_{a_{\tilde{t}}} \cap J, k < t\}\right\} \\
&\geq \sup\left\{\tilde{c} : \Pr\left(p_{a_k} > c_{a_k}(I) \forall a_k \in K, p_t < \tilde{c} | H_{K \cup \{a_t\}}\right) \leq \left[\sum_{\substack{a_k \in ((G_{a_{\tilde{t}}} \cap J) \setminus K) \\ k \leq t}} w_{a_k}(J)\right] \alpha \forall K \subseteq \{a_k : a_k \in G_{a_{\tilde{t}}} \cap J, k < t\}\right\} \\
&\geq \sup\left\{\tilde{c} : \Pr\left(p_{a_k} > c_{a_k}(I) \forall a_k \in K, p_t < \tilde{c} | H_{K \cup \{a_t\}}\right) \leq \left[\sum_{\substack{a_k \in ((G_{a_{\tilde{t}}} \cap I) \setminus K) \\ k \leq t}} w_{a_k}(I)\right] \alpha \forall K \subseteq \{a_k : a_k \in G_{a_{\tilde{t}}} \cap I, k < t\}\right\} \\
&= c_{a_t}(I)
\end{aligned}$$

where the first inequality follows by induction and the second inequality is because the weights are monotonic. □

Proof for Theorem 2. Per Lemmas 1 and 2, it suffices to show that the fully prespecified SRGP controls the FWER.

We first prove that the critical values per (6) control the Type I error for any intersection hypothesis I. For any I, define $\tilde{I}$ as the union of I and all prespecified nodes. Then the Type I error can be bounded using a sequence of union bounds:

$$\begin{aligned}
&\Pr\left(\exists (t, a_t) \in I \text{ s.t. } p_{a_t} < c_{a_t}(I) \mid H_I\right) \\
&\leq \Pr\left(\exists t \text{ s.t. } \xi^{\text{pres}}_{t,n} \leq z^{\text{pres}}_t(I) \text{ OR } \exists (t, a_t) \in I \text{ s.t. } p_{a_t} < c_{a_t}(I) \mid H_I\right) \\
&\leq \sum_{t=1}^{\infty} \left[\Pr\left(\xi^{\text{pres}}_{t',n} > z^{\text{pres}}_{t'}(I) \forall t' \leq t-1, \xi^{\text{pres}}_{t,n} \leq z^{\text{pres}}_t(I) \,\middle|\, H_I\right) + \sum_{a_t \in I} \Pr\left(\xi^{\text{pres}}_{t',n} > z^{\text{pres}}_{t'}(I) \forall t' \leq t, p_{a_t} < c_{a_t}(I) \,\middle|\, H_I\right)\right] \\
&\leq \left(\sum_{t=1}^{\infty} \left(w^{\text{pres}}_t\left(\tilde{I}\right) + \sum_{a_t \in I} w_{a_t}\left(\tilde{I}\right)\right)\right) \alpha \\
&= \alpha.
\end{aligned}$$

Because the weights are nondecreasing in Algorithm 1, the critical values defined in (6) satisfy the monotonicity condition. As such, Algorithm 1 is a consonant, short-cut procedure for the above closed test.

□

B HYPOTHESIS TEST DETAILS

B.1 TESTING FOR AN IMPROVEMENT IN AUC

In Section 3, we decide whether or not to approve a modification by testing the adaptively-defined null hypothesis (7) at each iteration j, which compares the AUC between the jth adaptively proposed model and the initial model. Per Algorithm 1, we test the adaptive hypotheses by treating them as pre-specified hypotheses from a bifurcating tree, i.e.

$$H_{0,a_j} : \psi\left(\hat{f}_{a_j}, P_0\right) \leq \psi\left(\hat{f}_0; P_0\right) + \delta_{a_j} \tag{S.3}$$

for approval histories a_j. We now describe how the test statistics and significance thresholds are constructed.

Recall that the AUC is equal to the Mann-Whitney U-statistic for comparing ranks across two populations, i.e.

$$\psi(f, P_0) = P_0\left(f(X_1) > f(X_2) \mid Y_1 = 1, Y_2 = 0\right), \tag{S.4}$$

where (X_1, Y_1) and (X_2, Y_2) represent independent draws from P_0. The empirical AUC is defined as

$$\psi(f, P_n) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \mathbb{1}\left\{f(X_j) > f(X_i)\right\} \mathbb{1}\left\{Y_j = 1, Y_i = 0\right\}, \tag{S.5}$$

where n_0 is the number of observations with $Y = 0$ and $n_1 = n - n_0$. To test (S.3), we characterize the asymptotic distribution of (S.5) by analyzing its influence function. Given IID observations from P_0, (S.5) is an asymptotically linear estimator of the model's AUC (LeDell et al., 2015), in that

$$\psi(f, P_n) - \psi(f, P_0) = \frac{1}{n} \sum_{i=1}^{n} \phi(f, P_0)(X_i, Y_i) + o_p(1/\sqrt{n}) \tag{S.6}$$

with influence function

$$\begin{aligned}
\phi(f, P_0)(X_i, Y_i) =& \frac{\mathbb{1}\{Y_i = 1\}}{P_0(Y = 1)} P_0\left(f(X) < c \mid Y = 0; c = f(X_i)\right) \\
&+ \frac{\mathbb{1}\{Y_i = 0\}}{P_0(Y = 0)} P_0\left(f(X) > c \mid Y = 1; c = f(X_i)\right) \\
&- \left\{\frac{\mathbb{1}\{Y_i = 0\}}{P_0(Y = 0)} + \frac{\mathbb{1}\{Y_i = 0\}}{P_0(Y = 0)}\right\} \psi\left(f, P_0\right).
\end{aligned}$$

Per the Central Limit Theorem, we have that

$$\sqrt{n}\left(\psi(f, P_n) - \psi(f, P_0)\right) \to_d N\left(0, \sigma(f, P_0)^2\right) \tag{S.7}$$

where $\sigma(f, P_0)^2 = \mathrm{Var}(\phi(f, P_0)(X, Y))$. We can then test the null hypothesis $H_0 : \psi(\hat{f}_0, P_0) \leq c$ for some constant c based on the asymptotic normality of (S.5). In addition, we can test (S.3) by deriving the asymptotic distribution of $\psi\left(\hat{f}_{a_j}, P_0\right) - \psi\left(\hat{f}_0; P_0\right)$ based on the difference of the influence functions $\phi(\hat{f}_{a_j}, P_0)(X, Y) - \phi(\hat{f}_0, P_0)(X, Y)$. To run `fsSRGP`, we can extend the above derivations to construct a flexible fixed sequence test for testing a family of null hypotheses (S.3) across multiple iterations j by analyzing the joint asymptotic distribution of the test statistics $\psi\left(\hat{f}_{a_j}, P_n\right) - \psi\left(\hat{f}_0; P_n\right)$ and compute the significance thresholds defined in (4). Similar logic can be used to derive the critical values (5) and significance thresholds (6) in `fsSRGP`.

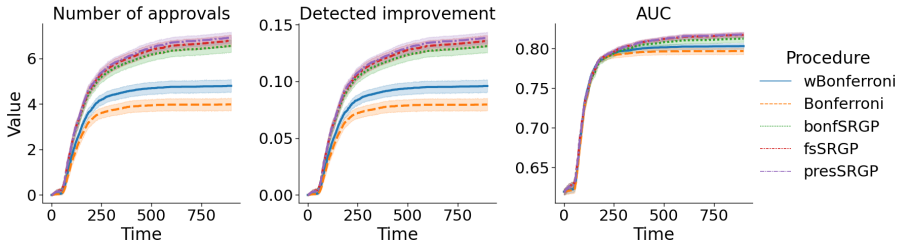

Figure 1: Comparison of MTPs for approving modifications when the model developer is highly risk averse. The simulation is the same as that for Figure 4, except the modifications are submitted only if the calculated power exceeds 80%.

B.2 TESTING MODEL DISCRIMINATION AND CALIBRATION

Section 4 considers the more complex hypothesis test (8), which checks for an improvement in AUC and calibration-in-the-large. We implement this by testing three individual hypothesis tests using sequential gatekeeping. First, we test that the difference between the average risk prediction and the observed event rate is no smaller than $-\epsilon$. Next, we test that this difference is no larger than ϵ. Finally, we test for an improvement in AUC using the procedure described in Section B.1. To control the Type I error for rejecting the overall null hypothesis, we perform alpha spending across the individual hypotheses.

C ADDITIONAL EXPERIMENTS

C.1 SENSITIVITY ANALYSIS TO RISK TOLERANCE OF MODEL DEVELOPER

In Section 3.2, we simulated a model developer who submits a refitted model for testing only if the power exceeds a threshold of 50%. This threshold is a reflection of the model developer's risk tolerance. A model developer who selects a higher threshold is more likely to have their modifications approved, but the time between each model submission is also longer. To understand how a more conservative model developer would affect the results in Section 3.2, we rerun the same simulation except with a threshold of 80%. As seen in Figure 1, the overall rate of model improvement is slower. For example, `presSRGP` previously required 180 observations to reach an AUC of 0.80 when the power threshold was 50%. In comparison, it requires nearly 250 observations when the power threshold is set to 80%. Also, the performance of the different MTPs are now more similar, particularly between the different SRGPs. This is also unsurprising, as the power to approve these modifications is much higher in this simulation; the additional power gain from employing `fsSRGP` and/or `presSRGP` as compared to `bonfSRGP` is now much smaller. Finally, a more conservative model development strategy decreases the variability of the approval histories, as evidenced by the narrower error bars.

C.2 SENSITIVITY ANALYSIS OF SRGP WITH HYPOTHETICAL PRESPECIFIED MODEL UPDATES

The power of SRGP with hypothetical prespecified model updates (`presSRGP`) depends on the similarity between the prespecified and adaptive model updates. As their correlation increases, the power of `presSRGP` will increase, all other things being equal. Here we present a simulation study where we investigate the sensitivity of `presSRGP` to the similarity of the model updates, using the same data stream as that in Section 3.2. We have carefully designed three model developers such that their adaptively generated model updates have different correlations with the prespecified model updates but the rate of improvement in AUC is the same. To do so, the prespecified updating procedure trains on only observations with *even* indices. The first model developer (`Even`) generates updates as close as possible to the prespecified rule: they refit the model using only even indices and adaptively submit modifications if the calculated power exceeds 50%. The second model developer (`Odd`) generates updates in a very different manner: they refit using only observations with odd indices. Finally, the third model developer (`Even/Odd`) deviates moderately from the prespecified model updates: they train modifications on observations with indices that are 0, 3, and 5 mod 6. Figure 2 shows that the power for approving modifications depends on how much the model developer deviates from their prespecified updating procedure. The moderate deviations in `Even/Odd` lead to a small drop in the approval rate and very slight drop in AUC. The drop in performance is more obvious in `Odd`, where the adaptive strategy does not align with the prespecified updating procedure at all.

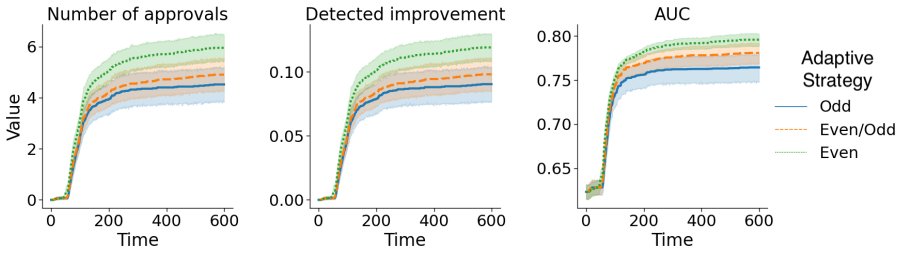

Figure 2: We analyze the sensitivity of `presSRGP` to the similarity between the adaptive and prespecified model updates. We simulate three model developers who increasingly deviate from the prespecified updates: `Even`, `Even/Odd`, and `Odd`, ordered from lowest to highest deviation from the prespecified updating procedure.

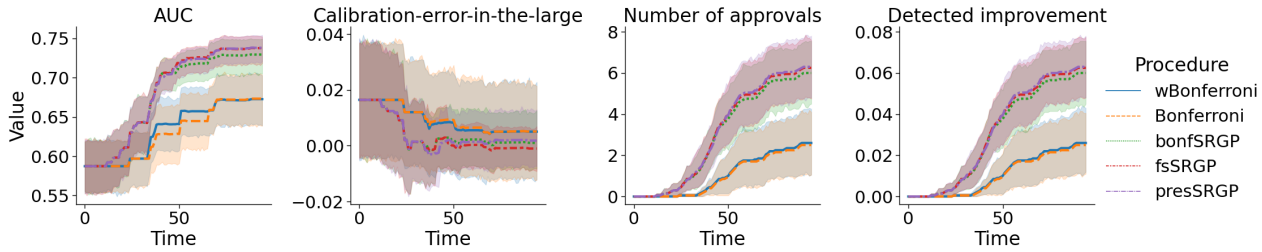

Figure 3: Approving refitted gradient boosted trees for detecting intracranial hemorrhages in head CT scans. Models are approved if the calibration-error-in-the-large is close to the ideal value of zero and the AUC is improving.

C.3 REVISING DETECTION ALGORITHM FOR INTRACRANIAL HEMORRHAGES

Here we present a second data analysis. We analyze data from the RSNA 2019 Brain CT Hemorrhage Challenge (Flanders et al., 2020), where the prediction task was to detect and classify intracranial hemorrhages (ICH) based on head CT scans. We follow nearly the same pre-processing procedure outlined in (Gossmann et al., 2021): we extract two axial slices from each subject's head CT scan and then apply a pre-trained ResNet50 model (without any training or fine-tuning on the medical images) to extract 2048 features from each image. We will consider the binary classification task of detecting the presence of any ICH subtype. The adaptive testing setup is similar to that outlined in Section 4. For each simulation replicate, we randomly select 100 subjects to train the initial GBT model, 900 subjects to generate model updates, and 500 subjects for adaptive test data reuse. The 900 subjects are randomly ordered to construct a data stream, in which data from 10 patients arrive at each time point. At each time point, we refit the GBT on all previously collected data. The model developer is allowed $T = 10$ adaptive tests. Figure 3 shows that the result for 20 replicates. `presSRGP` and `fsSRGP` performed the best; `bonfSRGP` performed very similarly.

References

Frank Bretz, Willi Maurer, Werner Brannath, and Martin Posch. A graphical approach to sequentially rejective multiple test procedures. *Stat. Med.*, 28(4):586–604, February 2009.

Frank Bretz, Willi Maurer, and Gerhard Hommel. Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures. *Stat. Med.*, 30(13):1489–1501, June 2011.

Adam E Flanders, Luciano M Prevedello, George Shih, Safwan S Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T Mongan, Anouk Stein, Felipe C Kitamura, Matthew P Lungren, Gagandeep Choudhary, Lesley Cala, Luiz Coelho, Monique Mogensen, Fanny Morón, Elka Miller, Ichiro Ikuta, Vahe Zohrabian, Olivia McDonnell, Christie Lincoln, Lubdha Shah, David Joyner, Amit Agarwal, Ryan K Lee, Jaya Nath, and RSNA-ASNR 2019 Brain Hemorrhage CT Annotators. Construction of a machine learning dataset through collaboration: The RSNA 2019 brain CT hemorrhage challenge. *Radiol Artif Intell*, 2(3):e190211, May 2020.

Alexej Gossmann, Aria Pezeshk, Yu-Ping Wang, and Berkman Sahiner. Test data reuse for the evaluation of continuously

evolving classification algorithms using the area under the receiver operating characteristic curve. *SIAM Journal on Mathematics of Data Science*, pages 692–714, January 2021.

Erin LeDell, Maya Petersen, and Mark van der Laan. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electron. J. Stat.*, 9(1):1583–1607, 2015.