

Supporting Information:

**Approval policies for modifications to Machine Learning-Based Software as a
Medical Device: A study of bio-creep**

Jean Feng^{*}, Scott Emerson^{}, and Noah Simon^{***}**

Department of Biostatistics, University of Washington, Seattle, WA, USA

**email:* jeanfeng@uw.edu

***email:* semerson@uw.edu

****email:* nrsimon@uw.edu

Supporting Information

A. Proofs

THEOREM 1: *aACP-BAC achieves uniform control of $\text{BAC}_W(\cdot)$ at level α , i.e.*

$$\text{BAC}_W(T) \leq \alpha \quad T = 1, 2, \dots \quad (1)$$

Proof. At each time point, aACP-BAC launches a set of hypothesis tests comparing \hat{f}_t to models with indices $\hat{M}_t = \{\hat{A}_1, \dots, \hat{A}_t, \hat{A}_t + 1, \dots, t - 1\}$. Let the $\tilde{\mathcal{F}}_t$ -measurable random variable G_t indicate the indices of the true null hypotheses, i.e.

$$\hat{G}_t = \{j \in \hat{M}_t : \hat{f}_j \not\rightarrow_\epsilon \hat{f}_t\}.$$

It is easy to see that the number of bad approvals is upper bounded by the number of incorrect rejections of the launched null hypotheses, i.e.

$$\begin{aligned} & \sum_{1 \vee (T-W)}^T \mathbb{1} \left\{ \exists t' = 1, \dots, t-1 \text{ s.t. } \hat{f}_{\hat{A}_{t'}} \not\rightarrow_\epsilon \hat{f}_{\hat{A}_t} \right\} \\ & \leq \sum_{1 \vee (T-W)}^T \mathbb{1} \left\{ \exists j \in \hat{G}_t, \exists t' = 1, \dots, \Delta_t, \text{ s.t. reject } \hat{f}_j \not\rightarrow_\epsilon \hat{f}_t \text{ at time } t+t' \right\}. \end{aligned} \quad (2)$$

Taking the expectations on both sides, $\text{BAC}_W(T)$ is upper-bounded by

$$\sum_{1 \vee (T-W)}^T \Pr \left(\exists j \in \hat{G}_t, \exists t' = 1, \dots, \Delta_t, \text{ s.t. reject } \hat{f}_j \not\rightarrow_\epsilon \hat{f}_t \text{ at time } t+t' \right). \quad (3)$$

Since the hypothesis tests are tested using a gatekeeping procedure, each probability in (3) is equal to the probability of rejecting the first true null hypothesis in the gatekeeping sequence.

Thus,

$$\Pr \left(\exists j \in \hat{G}_t, \exists t' = 1, \dots, \Delta_t, \text{ s.t. reject } \hat{f}_j \not\rightarrow_\epsilon \hat{f}_t \text{ at time } t+t' \right) \quad (4)$$

$$= \Pr \left(\hat{G}_t \neq \emptyset, \exists t' = 1, \dots, \Delta_t, \text{ s.t. reject } \hat{f}_{\min \hat{G}_t} \not\rightarrow_\epsilon \hat{f}_t \text{ at time } t+t' \right) \quad (5)$$

$$\leq E \left[\Pr \left(\hat{G}_t \neq \emptyset, \exists t' = 1, \dots, \Delta_t, \text{ s.t. reject } \hat{f}_{\min \hat{G}_t} \not\rightarrow_\epsilon \hat{f}_t \text{ at time } t+t' \mid \tilde{\mathcal{F}}_t \right) \right] \quad (6)$$

$$\leq E \left[\hat{\alpha}_t \mathbb{1} \left\{ \hat{G}_t \neq \emptyset \right\} \right]. \quad (7)$$

Summing together the probabilities within the window, we have

$$\text{BAC}_W(T) \leq E \left[\sum_{1 \vee (T-W)}^T \hat{\alpha}_t \right] \leq \alpha, \quad (8)$$

where the last inequality follows from the fact that $\hat{\alpha}_t$ is always selected such that

$$\sum_{1 \vee (T-W)}^T \hat{\alpha}_t \leq \alpha.$$

THEOREM 2: *aACP-BABR achieves uniform control of $\text{meBAR}_W(\cdot)$ and $\text{meBBR}_W(\cdot)$ at levels α and α' , respectively, i.e.*

$$\text{meBAR}_W(T) \leq \alpha \quad \forall T = 1, 2, \dots \quad (9)$$

$$\text{meBBR}_W(T) \leq \alpha' \quad \forall T = 1, 2, \dots \quad (10)$$

Proof. For all T , $\hat{\alpha}_T$ is selected such that

$$\text{B}\hat{\text{A}}\text{R}_{W'}(T) = \frac{\sum_{t=1}^T \hat{\alpha}_t \mathbb{1}\{t-W \leq t' + \Delta_{t'} \leq t\}}{1 + \sum_{t=1 \vee (T-W)}^{T-1} \mathbb{1}\{\hat{B}_{t-1} \neq \hat{B}_t\}} \leq \alpha \quad \forall W' = 1, \dots, W. \quad (11)$$

Note that we can always set $\hat{\alpha}_T = 0$ to satisfy these constraints, assuming that (11) was satisfied at times $t = 1, \dots, T-1$. Using the result in the proof of Theorem 1, we then bound the numerator of $\text{BAR}_W(T)$ as follows

$$E \left[\sum_{t=1 \vee (T-W)}^T \mathbb{1}\{\exists t' = 1, \dots, t-1 \text{ s.t. } \hat{f}_{\hat{A}_{t'}} \not\rightarrow_{\epsilon, t} \hat{f}_{\hat{A}_t}\} \right] \quad (12)$$

$$\leq E \left[\sum_{t=1}^T \hat{\alpha}_t \mathbb{1}\{t-W \leq t' + \Delta_{t'} \leq t\} \right] \quad (13)$$

$$\leq E \left[\alpha \left(1 + \sum_{t=1 \vee (T-W)}^T \mathbb{1}\{\hat{B}_{t-1} \neq \hat{B}_t\} \right) \right], \quad (14)$$

where the last line follows from (11). Rearranging, we get that $\text{meBAR}_W(T) \leq \alpha$. The proof for uniform control of $\text{meBBR}_W(\cdot)$ is essentially the same, where we replace the alpha-spending function with α'_t and the threshold with α' .

Algorithm 1: aACP-BAC

```

for  $t = 1, 2, \dots$  do
   $\hat{A}_t = \hat{A}_{t-1}$ ;
  /* Determine if there are new approvals */
  for  $j = \hat{A}_{t-1} + 1, \dots, t - 1$  do
    if  $t \leq j + \Delta_j$  ; // If  $\hat{f}_j$  is under consideration for approval
    then
      Run  $\epsilon$ -acceptability tests: Test null hypotheses  $\hat{f}_{j'} \rightarrow_{\epsilon} \hat{f}_j$  for  $j' = \hat{A}_1, \dots, \hat{A}_{j-1}, \hat{A}_{j-1} + 1, \dots, \hat{A}_{t-1}$ 
      with critical value  $c_j(t)$  in gatekeeping style;
      if All  $\epsilon$ -acceptability tests pass then
         $\hat{A}_t = j$ ;
      end
    end
  end
  end
  /* Launch new hypothesis tests for new model proposal */
  Launch family of  $\epsilon$ -acceptability tests with null hypotheses  $\hat{f}_j \rightarrow_{\epsilon} \hat{f}_t$  for  $j = \hat{A}_1, \dots, \hat{A}_t, \hat{A}_t + 1, \dots, t - 1$ ;
  Choose  $\hat{\alpha}_t$  such that (4) is satisfied;
  Select alpha-spending function for  $\hat{\alpha}_t$  and its critical value function  $c_t(\cdot)$  over the next  $\Delta_t$  time points.;
end

```

[Table 1 about here.]

Algorithm 2: aACP-BABR

```

for  $t = 1, 2, \dots$  do
   $\hat{A}_t = \hat{A}_{t-1}$ ;
  /* Determine if there are new approvals */
  for  $j = \hat{A}_{t-1} + 1, \dots, t - 1$  do
    if  $t \leq j + \Delta_j$  ; // If  $\hat{f}_j$  is under consideration for approval
    then
      Run  $\epsilon$ -acceptability tests: Test null hypotheses  $\hat{f}_{j'} \rightarrow_{\epsilon} \hat{f}_j$  for  $j' = \hat{A}_j, \dots, \hat{A}_{t-1}$  with critical value
       $c_j(t)$  in gatekeeping style;
      if All  $\epsilon$ -acceptability tests pass then
         $\hat{A}_t = j$ ;
      end
    end
  end
end
 $\hat{B}_t = \hat{B}_{t-1}$ ;
/* Determine if there are new benchmarks */
for  $j = \hat{A}_1, \dots, \hat{A}_{t-1}$  do
  if  $j > \hat{B}_{t-1}$  and  $t \leq j + \Delta_j$  ; // If  $\hat{f}_j$  is under consideration for approval
  then
    Run superiority tests: Test null hypotheses  $\hat{f}_{j'} \rightarrow_{\epsilon} \hat{f}_j$  for  $j' = \hat{B}_j, \dots, \hat{B}_{t-1}$  with critical value  $c'_j(t)$ 
    in gatekeeping style;
    if All superiority tests pass then
       $\hat{A}_t = j$ ;
    end
  end
end
end
/* Launch new hypothesis tests for new model proposal */
Launch family of  $\epsilon$ -acceptability tests with null hypotheses  $\hat{f}_j \rightarrow_{\epsilon} \hat{f}_t$  for  $j = \hat{A}_1, \dots, \hat{A}_t, \hat{A}_t + 1, \dots, t - 1$ ;
Launch family of superiority tests with null hypotheses  $\hat{f}_j \rightarrow_{\epsilon} \hat{f}_t$  for  $j = \hat{B}_t, \dots, t - 1$ ;
Choose  $\hat{\alpha}_t, \hat{\alpha}'_t$  such that (11) and (12) are satisfied;
Select alpha-spending function for  $\hat{\alpha}_t$  and  $\hat{\alpha}'_t$  and their critical value functions  $c_t(\cdot), c'_t(\cdot)$  over the next
 $\Delta_t$  time points.;
end

```

[Table 2 about here.]

B. Simulation settings

We ran 200 replicates for each simulation.

B.1 Hypothesis testing procedure for acceptability

All the aACPs tested for acceptability of a modification from f to f' with null hypothesis $H_0 : f \rightarrow_{\epsilon} f'$ in the following manner. Let the true difference in sensitivities and specificities be denoted (θ_1, θ_2) . At each stage of the group sequential test, we construct rectangular confidence regions for the evaluation metrics using confidence intervals for each metric (Cook, 1994). At any stage, if the rectangular confidence region is completely within the region of acceptable modifications as defined by the NI margin ϵ , then we reject the null hypothesis. For simplicity, we use Pocock's alpha-spending function to determine the confidence levels at each stage (Pocock, 1977). Because our estimates for θ_1 and θ_2 are independent conditional on the number of negative and positive samples, we can control the Type I error of testing $H_0 : f \rightarrow_{\epsilon} f'$ at level α by using the significance thresholds from level $1 - \sqrt{1 - \alpha}$ group sequential tests for the individual metrics. It is easy to see why this works:

$$\Pr(\text{Confidence region fails to cover } (\theta_1, \theta_2) \text{ at some stage}) \quad (15)$$

$$= 1 - \Pr(\text{Confidence region covers } (\theta_1, \theta_2) \text{ at all stages}) \quad (16)$$

$$= 1 - \Pr(\text{CI covers } \theta_1 \text{ at all stages}) \Pr(\text{CI covers } \theta_2 \text{ at all stages}) \quad (17)$$

$$= \alpha \quad (18)$$

We test for superiority by setting $\epsilon = 0$.

B.2 Incremental deterioration

We set total time $T = 200$ and the maximum wait time $\Delta = 5$ for all models. The number of new monitoring observations at each time point increments by ten to estimate the true performance difference with increasing precision over time, starting with 200 observations.

B.3 *Periodic model deterioration and improvement*

We set total time $T = 100$ and maximum wait time $\Delta = 5$ for all models. We accumulate 200 new observations at each time point.

B.4 *Accumulating data*

Each patient is represented by 30 covariates and the true outcome is generated using a logistic model. The developer performs logistic regression with a lasso penalty and tunes the penalty parameter using cross-validation. To increase the margin of model improvement at later time points and the ability to detect small improvements, we increase the number of training observations at each time point by five, starting with size 20, and use a larger wait time of $\Delta = 10$. The total time is $T = 40$ since the model performance plateaus quickly.

B.5 *Significant model improvements*

In order to make the model improvements significant with high probability, we accumulate 650 observations at each time point, which is more than the other simulation settings. Since large improvements are relatively rare, we used a short total time of $T = 20$. Since a company is likely more confident in these improvements, the maximum wait time is set to $\Delta = 3$.

B.6 *Time trends*

The total time is $T = 100$ and the wait time is $\Delta = 5$. We accumulate 300 new observations at each time point.

C. **Sensitivity to choice of hyper-parameters**

The definition of the error rates depends on two hyper-parameters: the window size W and the non-inferiority margin ϵ . To study how sensitive the aACPs are to these hyper-parameter, we run the same set of simulations from Section 6 but vary either W or ϵ .

First, we vary ϵ over the values 0, 0.01, 0.05, and 0.1 while keeping the window $W = 15$

fixed (Figure 1). (For $\epsilon = 0$, note that approval requires demonstrating superiority.) As the NI margin increases, all the aACPs approve more modifications, both bad and good ones. Compared to aACP-BAC and -BABR, the error rates of aACP-Reset and -Baseline increase more quickly with respect to ϵ . As such, the relative ordering between the aACPs is similar across different values of ϵ .

In Figure 2, we vary the window size W over the values 1, 15, 25, and 50 while keeping the NI margin $\epsilon = 0.05$ fixed. Only aACP-BAC and -BABR depend on W ; The other aACPs are agnostic to the choice of W . As W increases, aACP-BAC and -BABR become more conservative and approve fewer modifications. Since their error rates are already quite low, the error rates are not very sensitive to changes in W . On the other hand, choosing an excessively large value of W , such as our example with $W = 50$, leads to significantly slower approval rates for proposed model improvements. As such, we suggest selecting a value for W that corresponds to the minimal time period one would like to control error rates for.

[Figure 1 about here.]

[Figure 2 about here.]

References

- Cook, R. J. (1994). Interim monitoring of bivariate responses using repeated confidence intervals. *Control. Clin. Trials* **15**, 187–200.
- Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191–199.

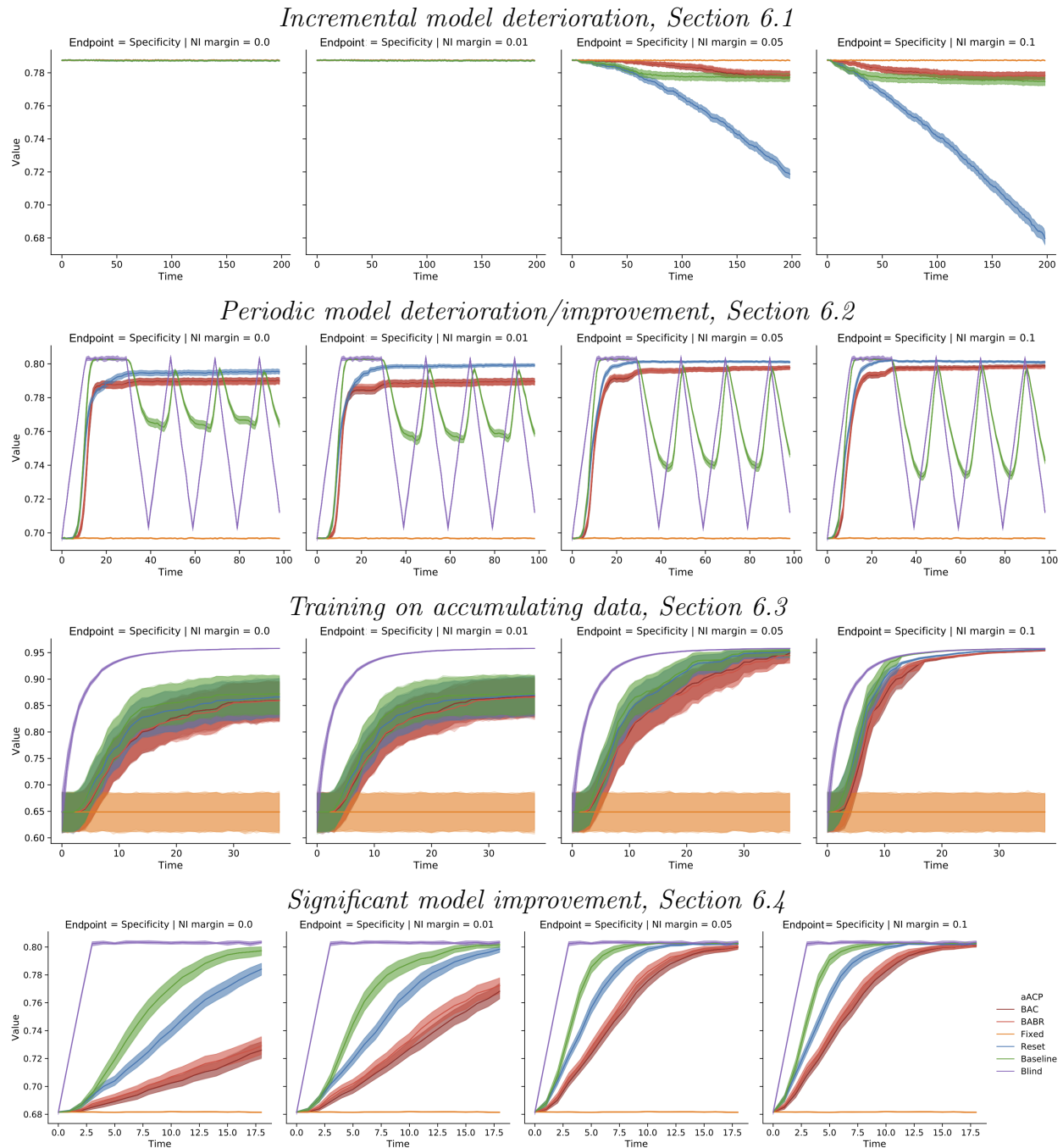


Figure 1. Specificity of the currently approved model over time for different simulation settings (rows) and non-inferiority margin values ϵ (columns). Sensitivity plots are very similar and, thus, have been omitted.

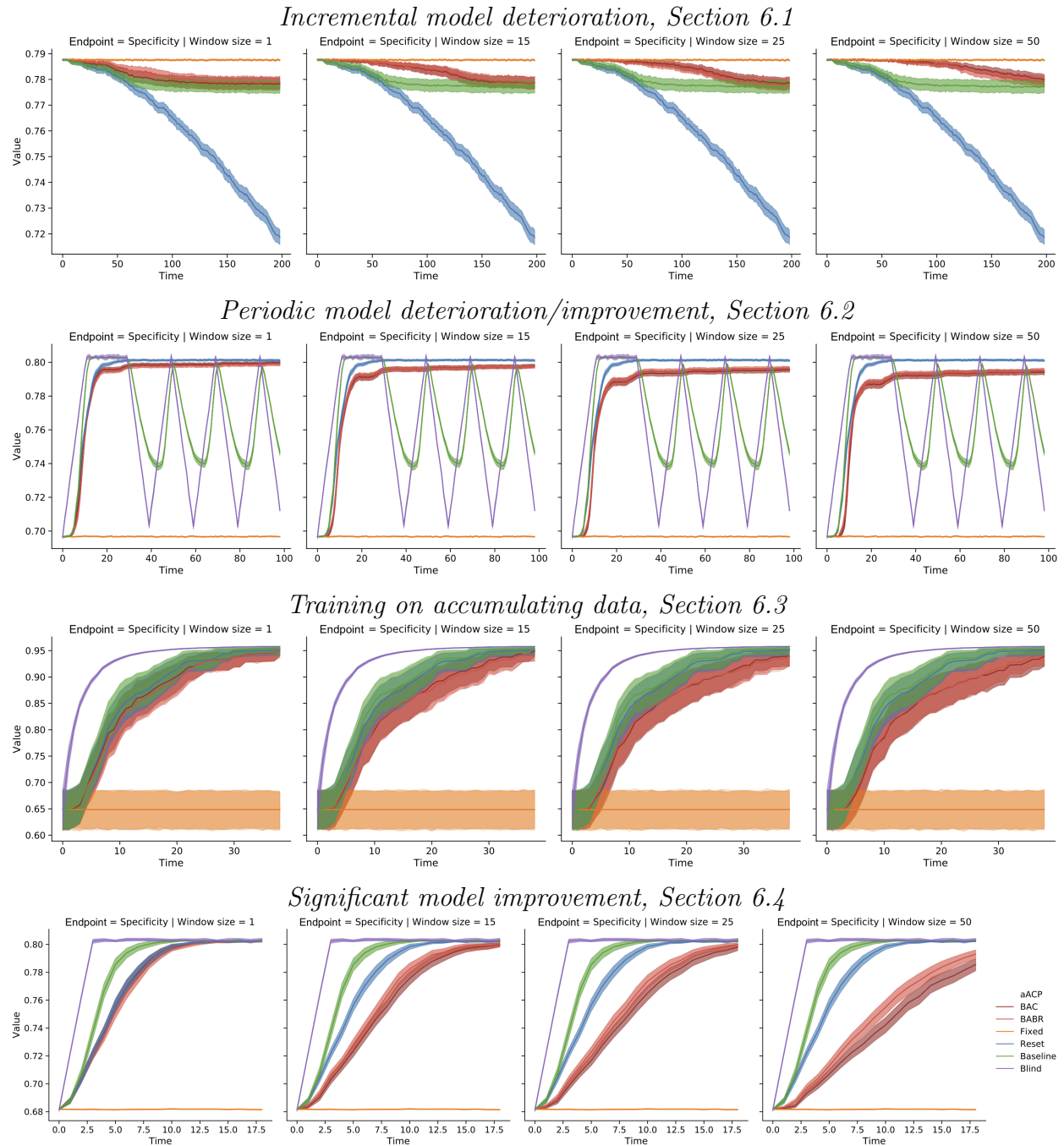


Figure 2. Specificity of the currently approved model over time for different simulation settings (rows) and window sizes W (columns). Sensitivity plots are very similar and, thus, have been omitted.

aACP	BAC_W	# approved	Final		Cumulative Utility		meBAR _W	meBBR _W
			Specificity	Sensitivity	Specificity	Sensitivity		
<i>Incremental model deterioration, Section 6.1</i>								
BABR	0.000	1.940	0.782	0.777	0.784	0.784	0.000	0.000
BAC	0.000	1.960	0.781	0.778	0.783	0.784	0.000	0.000
Baseline	0.060	2.220	0.778	0.779	0.781	0.781	0.060	0.000
Fixed	0.000	1.000	0.787	0.788	0.787	0.788	0.000	0.000
Reset	1.180	9.860	0.719	0.715	0.763	0.761	0.541	0.541
<i>Periodic model deterioration/improvement, Section 6.2</i>								
BABR	0.000	2.920	0.799	0.799	0.786	0.787	0.000	0.000
BAC	0.000	2.940	0.799	0.799	0.786	0.787	0.000	0.000
Baseline	6.300	43.360	0.749	0.749	0.765	0.766	6.300	0.000
Blind	15.000	99.000	0.712	0.711	0.762	0.762	15.000	0.000
Fixed	0.000	1.000	0.697	0.697	0.697	0.697	0.000	0.000
Reset	0.020	3.740	0.801	0.801	0.790	0.791	0.020	0.020
<i>Training on accumulating data, Section 6.3</i>								
BABR	0.015	3.495	0.945	0.950	0.842	0.854	0.010	0.000
BAC	0.010	3.430	0.949	0.949	0.844	0.854	0.010	0.000
Baseline	0.105	24.015	0.953	0.958	0.872	0.876	0.105	0.000
Blind	3.070	39.000	0.958	0.958	0.926	0.925	3.070	0.000
Fixed	0.000	1.000	0.649	0.636	0.649	0.636	0.000	0.000
Reset	0.055	4.810	0.951	0.956	0.866	0.871	0.018	0.139
<i>Significant model improvement, Section 6.4</i>								
BABR	0.000	4.020	0.802	0.802	0.757	0.757	0.000	0.000
BAC	0.000	3.980	0.801	0.801	0.753	0.753	0.000	0.000
Baseline	0.000	17.163	0.803	0.803	0.778	0.779	0.000	0.000
Blind	0.000	19.000	0.803	0.803	0.790	0.790	0.000	0.000
Fixed	0.000	1.000	0.681	0.682	0.682	0.682	0.000	0.000
Reset	0.000	4.429	0.802	0.802	0.770	0.771	0.000	0.007

Table 1

Comparison of automatic Algorithm Change Protocols in different simulation settings for IID data. Columns BAC_W , $meBAR_W$, $meBBR_W$ display the maximum error rate over all time points. In the incremental model deterioration, we omit aACP-Blind since it always converges to completely uninformative classifier.

aACP	BAC _w	# approved	Specificity	Final Sensitivity	Cumulative Specificity	Utility Sensitivity	meBAR _w	meBBR _w
<i>No time trend</i>								
BABR	0.010	1.025	0.712	0.712	0.712	0.712	0.010	0.000
BAC	0.010	1.025	0.712	0.712	0.712	0.712	0.010	0.000
Fixed	0.000	1.000	0.712	0.712	0.712	0.712	0.000	0.000
<i>Acceptability graph constant</i>								
BABR	0.010	1.035	0.651	0.651	0.712	0.712	0.010	0.000
BAC	0.015	1.035	0.651	0.651	0.712	0.712	0.015	0.000
Fixed	0.000	1.000	0.651	0.652	0.712	0.712	0.000	0.000
<i>Acceptability graph changing</i>								
BABR	—	1.704	0.652	0.652	0.689	0.689	—	—
BAC	—	1.720	0.649	0.650	0.689	0.689	—	—
Fixed	—	1.000	0.772	0.773	0.712	0.712	—	—

Table 2

Comparison of automatic Algorithm Change Protocols for different time trend scenarios (Section 6.5). Columns BAC_w, meBAR_w, meBBR_w display the maximum error rate over all time points. Their values are omitted for the case where the acceptability graph is changing over time since the error rates are not well-defined.