Sunny with a Chance of Hurricane

Decision Analytic Metrics for Forecast Evaluation

Alyssa Bilinski
Brown University

April 2025

Collaborators and Funding

Gabriel Norris
Brown University

Joshua A. Salomon
Stanford University

This work was supported in part by the Council of State and Territorial Epidemiologists (NU38OT000297) and the National Institute of Health through the National Institute of General Medical Sciences (1R35GM155224).

Decision analysis

Prediction models

What might happen?

Decision analysis

Prediction models + **User objectives**

What might happen?

Which goals? Which options?

Decision analysis

Prediction models **+** **User objectives**

What might happen? Which goals? Which options?

↓

Policy recommendations

What is the best course of action, given uncertainty?

Decision analysis

Prediction models + **User objectives**

What might happen?

Which goals? Which options?

Policy recommendations

What is the best course of action, given uncertainty?

Evaluation $\circlearrowleft$

Prior work

Throughout the life cycle of an outbreak, “triggers” for starting and stopping interventions should be:

1. Predictive of outcomes of policy interest

Prior work

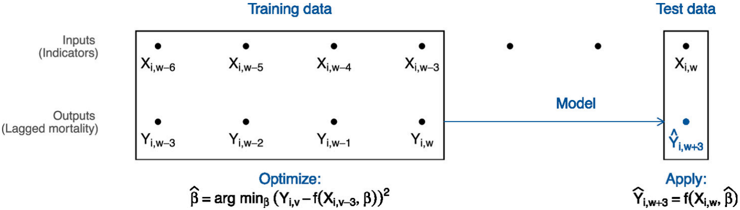
Throughout the life cycle of an outbreak, “triggers” for starting and stopping interventions should be:

1. Predictive of outcomes of policy interest
2. Account for the context-specific risk and costs of acting (or not) on false negative and false positive signals

Prior work

Throughout the life cycle of an outbreak, "triggers" for starting and stopping interventions should be:

1. Predictive of outcomes of policy interest
2. Account for the context-specific risk and costs of acting (or not) on false negative and false positive signals
3. Transparently and regularly updated



PNAS 2023, *Annals of IM* 2022

Prior work

Throughout the life cycle of an outbreak, "triggers" for starting and stopping interventions should be:

1. Predictive of outcomes of policy interest
2. Account for the context-specific risk and costs of acting false negative and false positive signals
3. Transparently and regularly updated

Community Risk Metrics

1. *PNAS* 2023
2. *Annals of IM* 2022
3. *PNAS* 2021

Special Populations

1. Schools: *JAMA NO* 2022, *Annals* 2021, *JAMA Peds* 2022
2. Nursing homes: *JAMA HF* 2024
3. "High" respiratory disease season guidance for HC facilities (w/RIDOH, in progress)

Disclosure

I am not a forecaster.

Disclosure

I am not a forecaster.

Forecasting seems very hard.

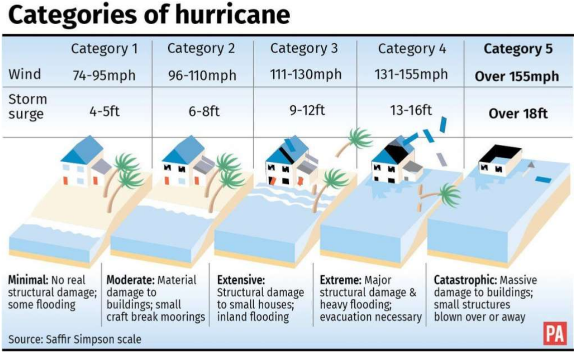
Disclosure

I am not a forecaster.

Forecasting seems very hard.

But we thought it might be valuable to transport a decision-analytic framework to this context.

Back to hurricanes...



Back to hurricanes...

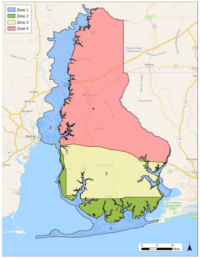
Scenario 1:

Category 1 - Zone 1: All areas of Pleasure Island along with individuals living in manufactured homes, and those living in low lying flood prone areas countywide. (Pleasure Island consists of all areas south of the Intra-coastal Canal to include Fort Morgan, Gulf Shores, Orange Beach and Ono Island.)

Category 2 - Zone 1 & 2: All areas south of State Hwy 98 and the area on the Eastern Shore that is South of Interstate 10 and West of State Hwy 98. Additionally, all individuals living in proximity to the Fish, Styx, Blackwater and Perdido Rivers and all individuals living in manufactured homes, and those living in low lying flood prone areas countywide.

Category 3 - Zones 1 through 3: All areas south of State Hwy 98 and the area on the Eastern Shore west of State Hwy 98, and the area west of State Hwy 225 and west of Hwy 59 North of Stockton to the Baldwin/Monroe County line. Additionally, all individuals living in proximity to the Fish, Styx, Blackwater and Perdido Rivers and all individuals living in manufactured homes, and those living in low lying flood prone areas countywide.

Category 4 or 5 - Zones 1 through 4: All areas south of Interstate 10 and the area on the Eastern Shore west of State Hwy 225 and west of Hwy 59 North of Stockton to the Baldwin/Monroe County line. Additionally, all individuals living in manufactured homes and those living in low lying flood prone areas countywide.



Back to hurricanes...

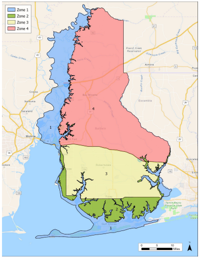
Scenario 1:

Category 1 - Zone 1: All areas of Pleasure Island along with individuals living in manufactured homes, and those living in low lying flood prone areas countywide. (Pleasure Island consists of all areas south of the Intra-coastal Canal to include Fort Morgan, Gulf Shores, Orange Beach and Ono Island.)

Category 2 - Zone 1 & 2: All areas south of State Hwy 98 and the area on the Eastern Shore that is South of Interstate 10 and West of State Hwy 98. Additionally, all individuals living in proximity to the Fish, Styx, Blackwater and Perdido Rivers and all individuals living in manufactured homes, and those living in low lying flood prone areas countywide.

Category 3 - Zones 1 through 3: All areas south of State Hwy 98 and the area on the Eastern Shore west of State Hwy 98, and the area west of State Hwy 225 and west of Hwy 59 North of Stockton to the Baldwin/Monroe County line. Additionally, all individuals living in proximity to the Fish, Styx, Blackwater and Perdido Rivers and all individuals living in manufactured homes, and those living in low lying flood prone areas countywide.

Category 4 or 5 - Zones 1 through 4: All areas south of Interstate 10 and the area on the Eastern Shore west of State Hwy 225 and west of Hwy 59 North of Stockton to the Baldwin/Monroe County line. Additionally, all individuals living in manufactured homes and those living in low lying flood prone areas countywide.



- Notification time varies (24-48 hours for Cat 1-3, 72 hours for Cat 4-5)

Back to hurricanes...

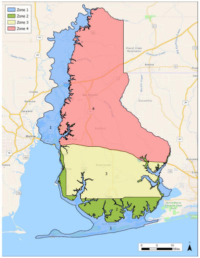
Scenario 1:

Category 1 - Zone 1: All areas of Pleasure Island along with individuals living in manufactured homes, and those living in low lying flood prone areas countywide. (Pleasure Island consists of all areas south of the Intra-coastal Canal to include Fort Morgan, Gulf Shores, Orange Beach and Ono Island.)

Category 2 - Zone 1 & 2: All areas south of State Hwy 98 and the area on the Eastern Shore that is South of Interstate 10 and West of State Hwy 98. Additionally, all individuals living in proximity to the Fish, Styx, Blackwater and Perdido Rivers and all individuals living in manufactured homes, and those living in low lying flood prone areas countywide.

Category 3 - Zones 1 through 3: All areas south of State Hwy 98 and the area on the Eastern Shore west of State Hwy 98, and the area west of State Hwy 225 and west of Hwy 59 North of Stockton to the Baldwin/Monroe County line. Additionally, all individuals living in proximity to the Fish, Styx, Blackwater and Perdido Rivers and all individuals living in manufactured homes, and those living in low lying flood prone areas countywide.

Category 4 or 5 - Zones 1 through 4: All areas south of Interstate 10 and the area on the Eastern Shore west of State Hwy 225 and west of Hwy 59 North of Stockton to the Baldwin/Monroe County line. Additionally, all individuals living in manufactured homes and those living in low lying flood prone areas countywide.



- Notification time varies (24-48 hours for Cat 1-3, 72 hours for Cat 4-5)
- Similar guidance for government, first responders

Key principles

Forecast $\longrightarrow$ Threshold $\longrightarrow$ Recommendation/Action

Key principles

$$\text{Forecast} \longrightarrow \text{Threshold} \longrightarrow \text{Recommendation/Action}$$

- There is some probability the threat materializes.

Key principles

$$\text{Forecast} \longrightarrow \text{Threshold} \longrightarrow \text{Recommendation/Action}$$

- There is some probability the threat materializes.
- There is a cost of evacuating (which may or may not be needed).
- There is a cost of staying put if a storm materializes.

Key principles

Forecast $\longrightarrow$ Threshold $\longrightarrow$ Recommendation/Action

- There is some probability the threat materializes.
- There is a cost of evacuating (which may or may not be needed).
- There is a cost of staying put if a storm materializes.
- Ideally, we choose recommendations accounting for these.

Key principles

Forecast $\longrightarrow$ Threshold $\longrightarrow$ Recommendation/Action

- There is some probability the threat materializes.
- There is a cost of evacuating (which may or may not be needed).
- There is a cost of staying put if a storm materializes.
- Ideally, we choose recommendations accounting for these.
  - Whether to cancel elective surgeries in a respiratory outbreak
  - Whether to start an mpox vaccination strategy

Popular forecast evaluation metrics

The most popular metric for forecast evaluation is **weighted interval score (WIS)**:

Popular forecast evaluation metrics

The most popular metric for forecast evaluation is **weighted interval score (WIS)**:

absolute error for probabilistic forecasts

Popular forecast evaluation metrics

The most popular metric for forecast evaluation is **weighted interval score (WIS)**:

absolute error for probabilistic forecasts

1. considers both point estimate and uncertainty
   $\longrightarrow$ balances accuracy and sharpness

Popular forecast evaluation metrics

The most popular metric for forecast evaluation is **weighted interval score (WIS)**:

absolute error for probabilistic forecasts

1. considers both point estimate and uncertainty
   $\longrightarrow$ balances accuracy and sharpness
2. "strictly proper scoring rule"
   $\longrightarrow$ aligned with reporting best forecast

Popular forecast evaluation metrics

The most popular metric for forecast evaluation is **weighted interval score (WIS)**:

absolute error for probabilistic forecasts

1. considers both point estimate and uncertainty
   $\longrightarrow$ balances accuracy and sharpness
2. "strictly proper scoring rule"
   $\longrightarrow$ aligned with reporting best forecast
3. but...equally weights all points in time and can be difficult to substantively interpret

Recent innovations

1. WIS of log-transformed estimates(Funk et. al., 2023)
2. Predicting shapes (Srivastava et. al., 2022, Srivastava et. al., 2023)
3. Optimizing allocation of a continuous finite resource (Gerding et. al., 2023)

Recent innovations

1. WIS of log-transformed estimates(Funk et. al., 2023)
2. Predicting shapes (Srivastava et. al., 2022, Srivastava et. al., 2023)
3. Optimizing allocation of a continuous finite resource (Gerding et. al., 2023)

...our questions were more basic.

This work

What information would make forecasts most interpretable and actionable to a decision-maker?

1. Propose simple forecast evaluation metrics tied to binary "threshold" outcomes
2. Evaluate performance on COVID-19 case and hospitalization predictions
3. Propose how to operationalize forecast error and uncertainty in decision-making

Methods

Results

Discussion

Metrics

We consider 3 types of metrics:

1. **Trends:** Is the outcome monotonically increasing or decreasing over the horizon?

Metrics

We consider 3 types of metrics:

1. **Trends:** Is the outcome monotonically increasing or decreasing over the horizon?
2. **Combined level/trend thresholds:** (e.g., for cases: $>$20 per 100k & $>$100% of forecast date value, for hospitalizations $<$10 per 100k & $<$100% of forecast date value)

Metrics

We consider 3 types of metrics:

1. **Trends:** Is the outcome monotonically increasing or decreasing over the horizon?
2. **Combined level/trend thresholds:** (e.g., for cases: $>$20 per 100k & $>$100% of forecast date value, for hospitalizations $<$10 per 100k & $<$100% of forecast date value)
3. **Turning points:** Monotonic increase followed by decrease (or the converse)
   - Also considered a “fuzzy” version of this (e.g. predicting peak within 1-2 weeks)

Decision analysis

Objective:
Maximize accuracy, weighting for preferences over different error types. We assume a decision-analytic framework.

Decision analysis

Objective:
Maximize accuracy, weighting for preferences over different error types. We assume a decision-analytic framework.

	Predicted negative: ($\hat{Y}_{w+3} = 0$)	**Predicted positive:** ($\hat{Y}_{w+3} = 1$)
True negative: ($Y_{w+3} = 0$)		
True positive: ($Y_{w+3} = 1$)		

Decision analysis

Objective:
Maximize accuracy, weighting for preference over different error types. We assume a decision-analytic framework.

	Predicted negative: ($\hat{Y}_{w+3} = 0$)	**Predicted positive:** ($\hat{Y}_{w+3} = 1$)
True negative: ($Y_{w+3} = 0$)	0	
True positive: ($Y_{w+3} = 1$)		

Decision analysis

Objective:
Maximize accuracy, weighting for preference over different error types. We assume a decision-analytic framework.

	Predicted negative: $(\hat{Y}_{w+3} = 0)$	**Predicted positive:** $(\hat{Y}_{w+3} = 1)$
True negative: $(Y_{w+3} = 0)$	0	S_0
True positive: $(Y_{w+3} = 1)$		

Decision analysis

Objective:
Maximize accuracy, weighting for preference over different error types. We assume a decision-analytic framework.

	Predicted negative: ($\hat{Y}_{w+3} = 0$)	**Predicted positive:** ($\hat{Y}_{w+3} = 1$)
True negative: ($Y_{w+3} = 0$)	0	S_0
True positive: ($Y_{w+3} = 1$)	D	

Decision analysis

Objective:
Maximize accuracy, weighting for preference over different error types. We assume a decision-analytic framework.

	Predicted negative: ($\hat{Y}_{w+3} = 0$)	**Predicted positive:** ($\hat{Y}_{w+3} = 1$)
True negative: ($Y_{w+3} = 0$)	0	S_0
True positive: ($Y_{w+3} = 1$)	D	$(1 - \alpha)D + S_1$

Decision analysis

The expected cost of following a metric (M) is:

$$C(M) = \underbrace{Pr(\hat{Y} = 1, Y = 0)S_0}_{\text{expected cost: false positives}} + \underbrace{Pr(\hat{Y} = 0, Y = 1)D}_{\text{expected cost: false negatives}} + \underbrace{Pr(\hat{Y} = 1, Y = 1)\,((1 - \alpha)D + S_1)}_{\text{expected cost: true positives}}$$

Decision analysis

We can rearrange this:

$$C(M) = Pr(\hat{Y} = 1, Y = 0)S_0 + \\ Pr(\hat{Y} = 0, Y = 1)(\alpha D - S_1) + \\ \underbrace{Pr(Y = 1)((1 - \alpha)D + S_1)}_{\text{constant across all metrics}}$$

Decision analysis

We can rearrange this:

$$\begin{aligned} C(M) &\propto p_{FP} S_0 + p_{FN}(\alpha D - S_1) \\ &\propto p_{FP} + p_{FN} w, \end{aligned}$$

where w is the ratio of the net benefit from taking action on a true positive ($\alpha D - S_1$) to costs incurred by unnecessary action in the case of a false positive (S_0)

Return

Decision analysis

If we define relative costs of acting on false positives vs. negatives, we can:

1. **Set decision thresholds:** Pick the optimal cutoff point for a prediction rule from a probabilistic forecast.

Decision analysis

If we define relative costs of acting on false positives vs. negatives, we can:

1. **Set decision thresholds:** Pick the optimal cutoff point for a prediction rule from a probabilistic forecast.
2. **Rank models:**

$$C(M) \propto p_{FP} + p_{FN} w,$$

Decision analysis

If we define relative costs of acting on false positives vs. negatives, we can:

1. **Set decision thresholds:** Pick the optimal cutoff point for a prediction rule from a probabilistic forecast.
2. **Rank models:**

$$C(M) \propto p_{FP} + p_{FN} w,$$

Decision analysis

If we define relative costs of acting on false positives vs. negatives, we can:

1. **Set decision thresholds:** Pick the optimal cutoff point for a prediction rule from a probabilistic forecast.
2. **Rank models:**

$$C(M) \propto p_{FP} + p_{FN} w,$$

Decision analysis

If we define relative costs of acting on false positives vs. negatives, we can:

1. **Set decision thresholds:** Pick the optimal cutoff point for a prediction rule from a probabilistic forecast.
2. **Rank models:**

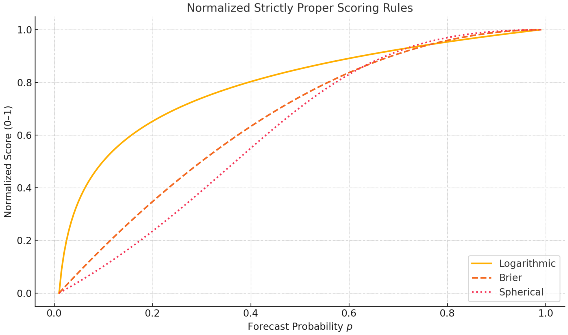
$$C(M) \propto p_{FP} + p_{FN} w,$$

Quick note

Accuracy (and weighted accuracy) do not induce "strictly proper scoring rules."

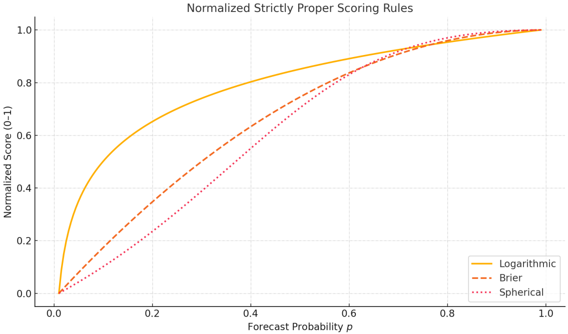
Quick note

Accuracy (and weighted accuracy) do not induce "strictly proper scoring rules."



Quick note

Accuracy (and weighted accuracy) do not induce "strictly proper scoring rules."



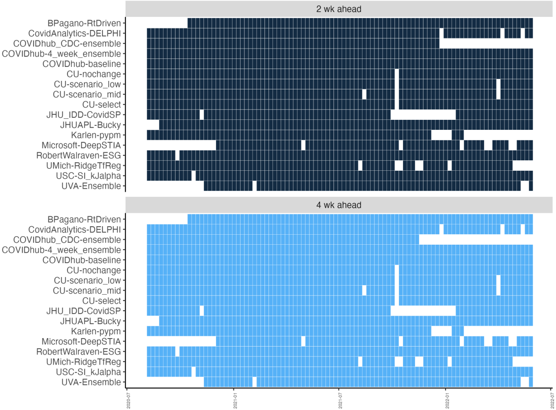
Considerations for both fitting and scoring, but we focus on the latter.

Data

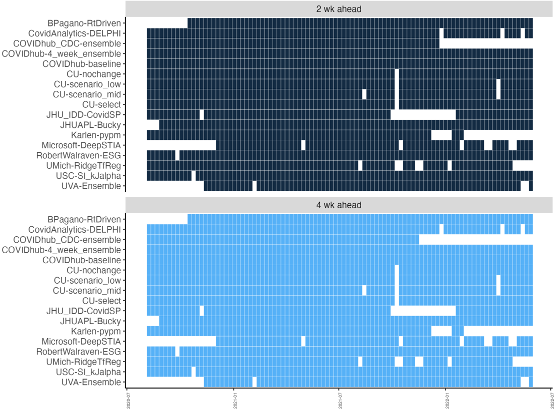
We analyzed COVID-19 Forecast Hub case and hospitalization projections from August 2020 through June 2022:

- National and state predictions
- Top 20 most frequently-reported models, ensemble models, baseline models (day-of prediction)
- quantile predictions $\longrightarrow$ mean $\longrightarrow$ binary outcomes (not sensitive to using median, preferred value)

Imputation


2 wk ahead
BPagano-RtDriven
CovidAnalytics-DELPHI
COVIDhub_CDC-ensemble
COVIDhub-4_week_ensemble
COVIDhub-baseline
CU-nochange
CU-scenario_low
CU-scenario_mid
CU-select
JHU_IDD-CovidSP
JHUAPL-Bucky
Karlen-pypm
Microsoft-DeepSTIA
RobertWalraven-ESG
UMich-RidgeTfReg
USC-SI_kJalpha
UVA-Ensemble
4 wk ahead
2020-07
2021-01
2021-07
2022-01
2022-07

Imputation



When missing, impute average (sensitivity analyses: baseline, best, worst).

Metrics

Using New York Times data as truth, we computed:

1. **Accuracy:** % correct
2. **Sensitivity/Specificity:** given true positive or negative status, how many correct?
3. **Positive predictive value/Negative predictive value:** given prediction class, how many correct?

Metrics

Using New York Times data as truth, we computed:

1. **Accuracy:** % correct
2. **Sensitivity/Specificity:** given true positive or negative status, how many correct?
3. **Positive predictive value/Negative predictive value:** given prediction class, how many correct?

In more preliminary results, we:

1. Characterize decision rules
2. Explore alternative ensembles

Extensions

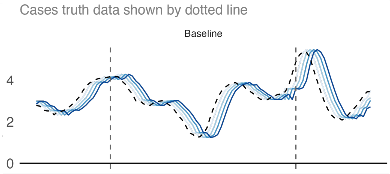
1. Limited decision points (in progress w/RIDOH)
2. Trade-offs between lead time and certainty

Methods

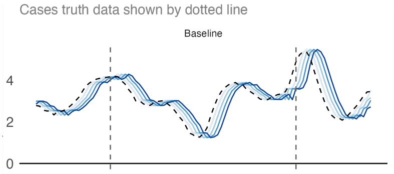
Results

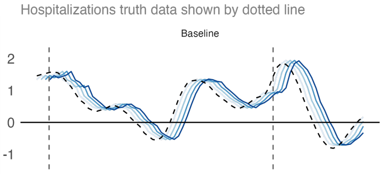
Discussion

Baseline

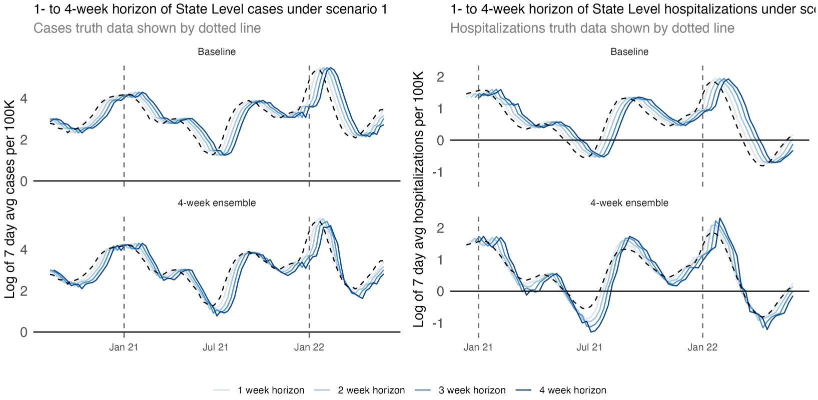


Baseline comparison

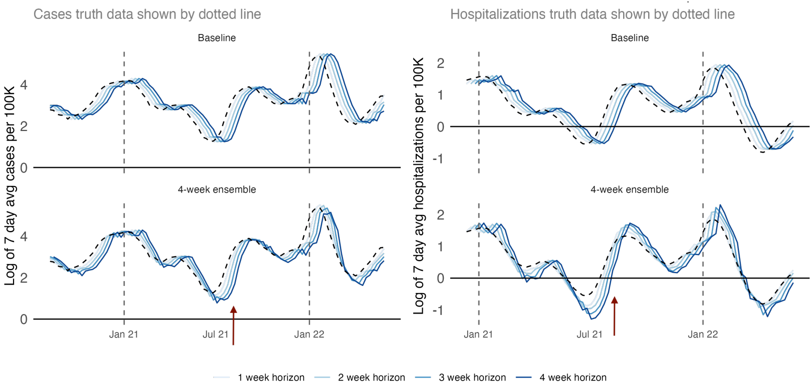




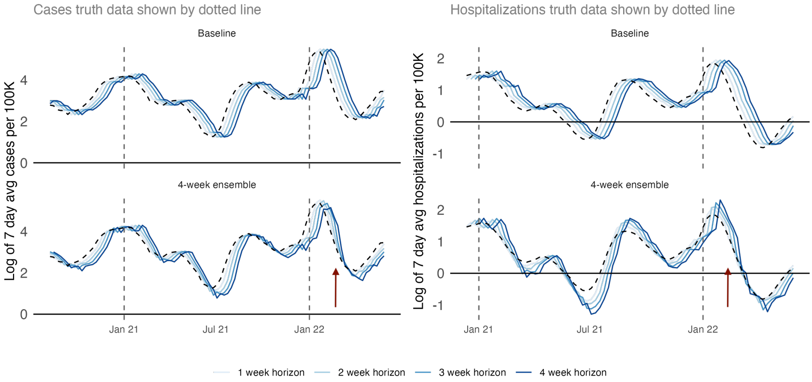
Baseline comparison



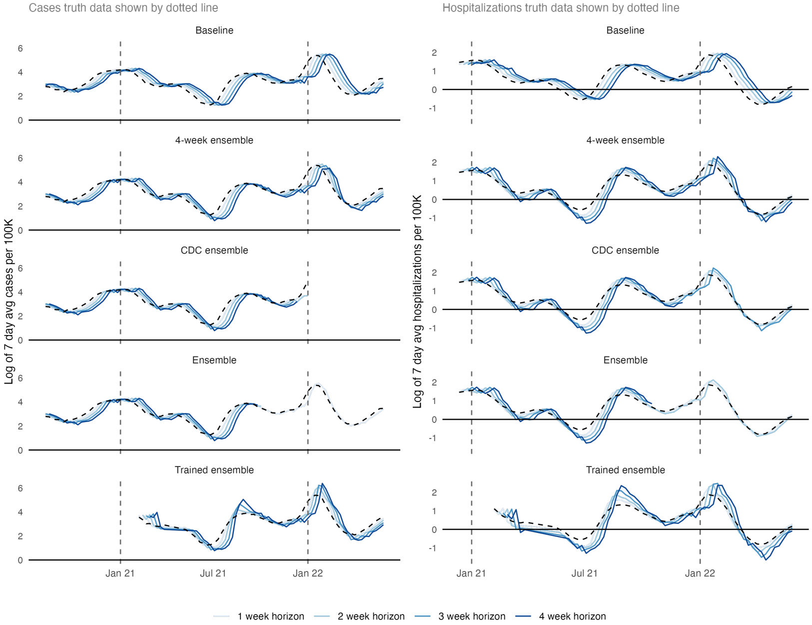
Baseline comparison



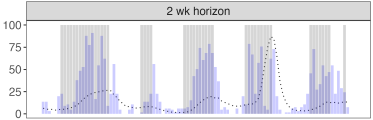
Baseline comparison



Baseline comparison

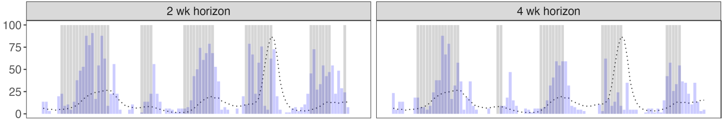


Performance over time



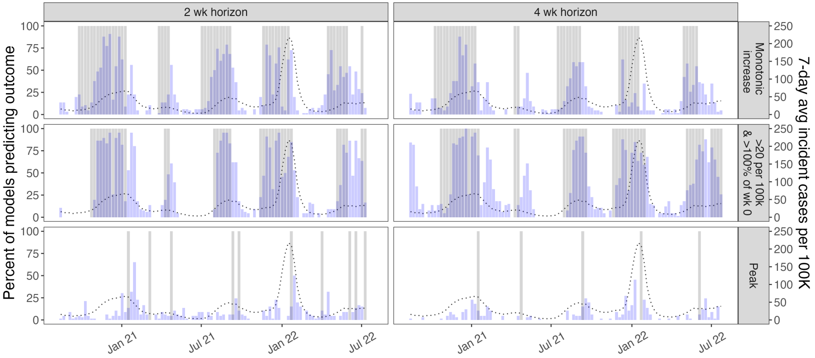
Hospitalizations

Performance over time



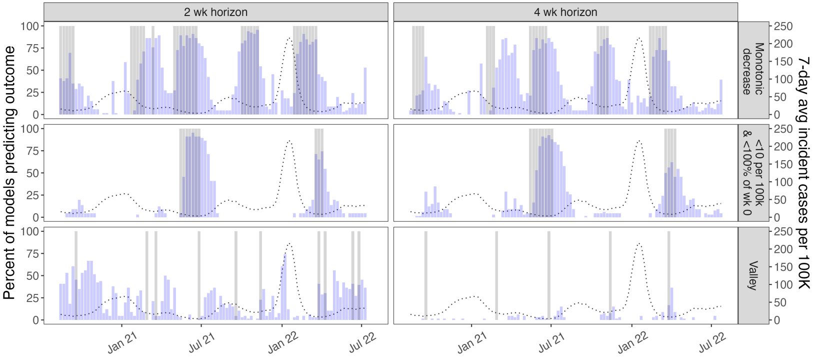
Hospitalizations

Performance over time



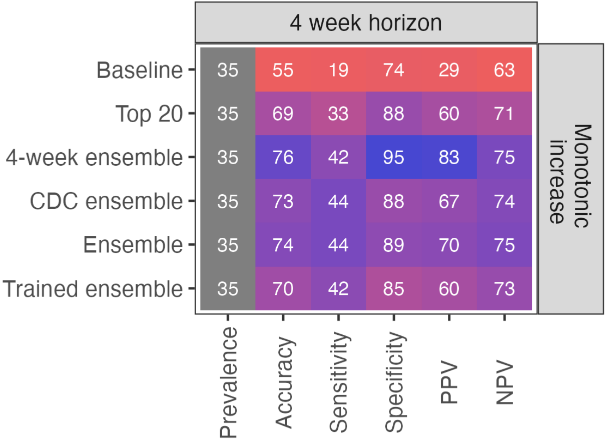
Hospitalizations

Performance over time

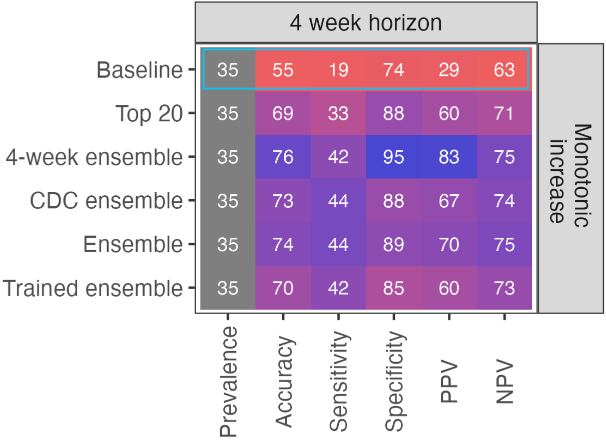


Hospitalizations

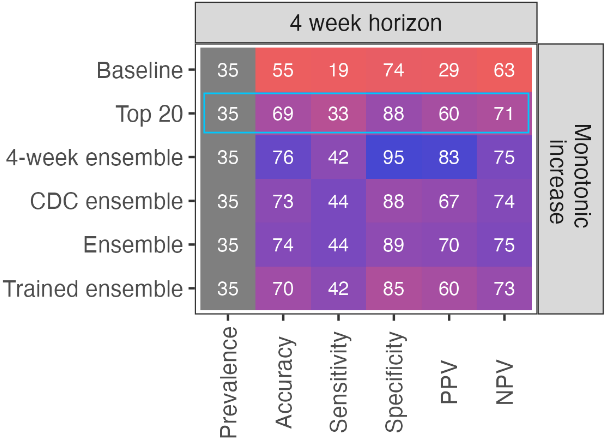
Metrics


4 week horizon
Monotonic increase
Baseline
Top 20
4-week ensemble
CDC ensemble
Ensemble
Trained ensemble
Prevalence
Accuracy
Sensitivity
Specificity
PPV
NPV
35 55 19 74 29 63
35 69 33 88 60 71
35 76 42 95 83 75
35 73 44 88 67 74
35 74 44 89 70 75
35 70 42 85 60 73

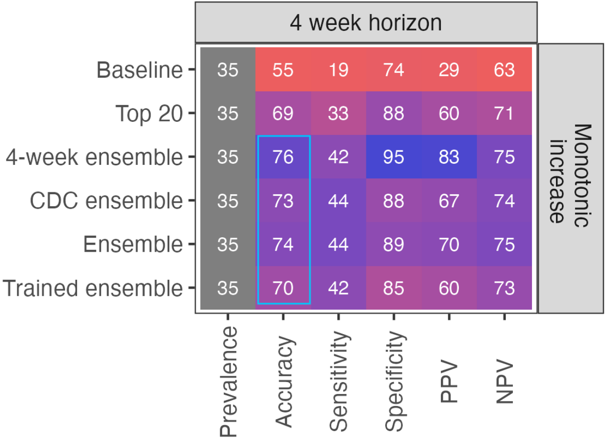
Metrics


4 week horizon
Monotonic increase
Baseline
Top 20
4-week ensemble
CDC ensemble
Ensemble
Trained ensemble
Prevalence
Accuracy
Sensitivity
Specificity
PPV
NPV
35 55 19 74 29 63
35 69 33 88 60 71
35 76 42 95 83 75
35 73 44 88 67 74
35 74 44 89 70 75
35 70 42 85 60 73

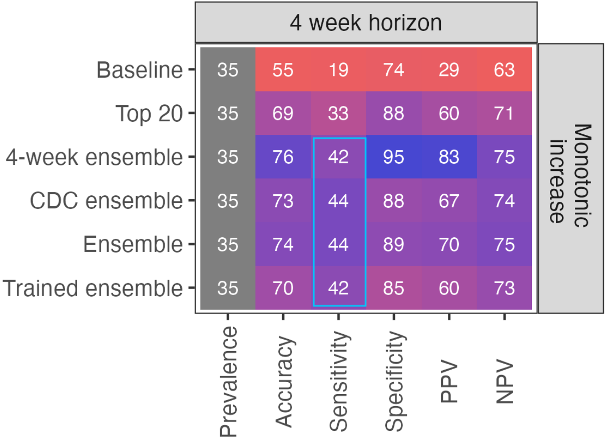
Metrics


4 week horizon
Baseline
Top 20
4-week ensemble
CDC ensemble
Ensemble
Trained ensemble
35 55 19 74 29 63
35 69 33 88 60 71
35 76 42 95 83 75
35 73 44 88 67 74
35 74 44 89 70 75
35 70 42 85 60 73
Prevalence
Accuracy
Sensitivity
Specificity
PPV
NPV
Monotonic increase

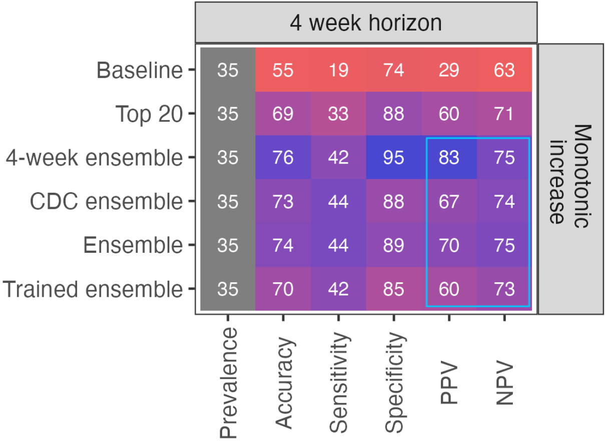
Metrics


4 week horizon
Baseline
Top 20
4-week ensemble
CDC ensemble
Ensemble
Trained ensemble
35 55 19 74 29 63
35 69 33 88 60 71
35 76 42 95 83 75
35 73 44 88 67 74
35 74 44 89 70 75
35 70 42 85 60 73
Prevalence
Accuracy
Sensitivity
Specificity
PPV
NPV
Monotonic increase

Metrics

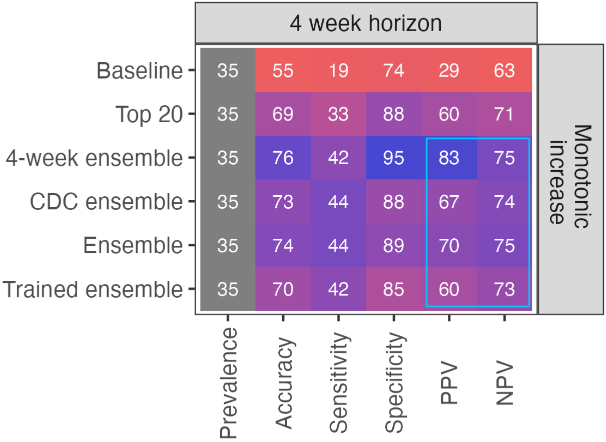

4 week horizon
Monotonic increase
Baseline
Top 20
4-week ensemble
CDC ensemble
Ensemble
Trained ensemble
Prevalence
Accuracy
Sensitivity
Specificity
PPV
NPV
35 55 19 74 29 63
35 69 33 88 60 71
35 76 42 95 83 75
35 73 44 88 67 74
35 74 44 89 70 75
35 70 42 85 60 73

Metrics



Cases Hospitalizations

What to do with this?



⟶ NPV and PPV are about 60-80%. Given a result (and the distribution of outcomes), about 1 in 3 chance it is correct.

Decision rules

When should we change behavior?

For "increasing" metrics:

- *Act on "increase" signal if:* willing to accept one false alarm (false positive) for every 2-4 correct calls

Decision rules

When should we change behavior?

For “increasing” metrics:

- *Act on “increase” signal if:* willing to accept one false alarm (false positive) for every 2-4 correct calls

Decision rules

When should we change behavior?

For "increasing" metrics:

- *Act on "increase" signal if:* willing to accept one false alarm (false positive) for every 2-4 correct calls

For "decreasing" metrics:

- *Act on "decrease" signal if:* willing to accept 50-50% chance correct call

Decision rules

When should we change behavior?

For "increasing" metrics:

- *Act on "increase" signal if:* willing to accept one false alarm (false positive) for every 2-4 correct calls

For "decreasing" metrics:

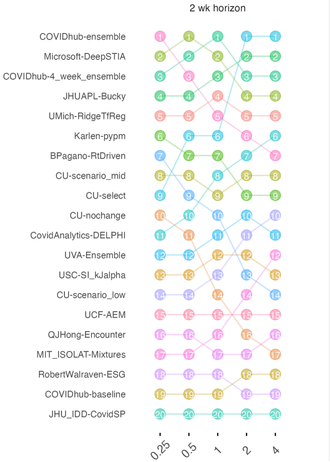
- *Act on "decrease" signal if:* willing to accept 50-50% chance correct call

For both:

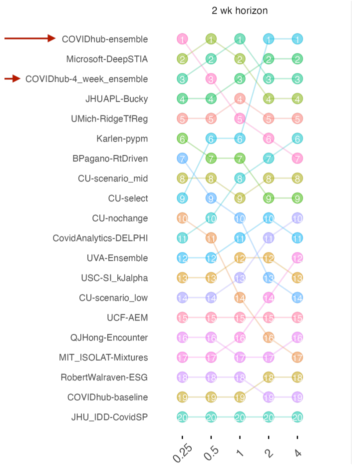
- *Stay put on "no change" signal:* 75-95% chance correct

There are caveats, but...**much appreciation for the work of state and local officials!**

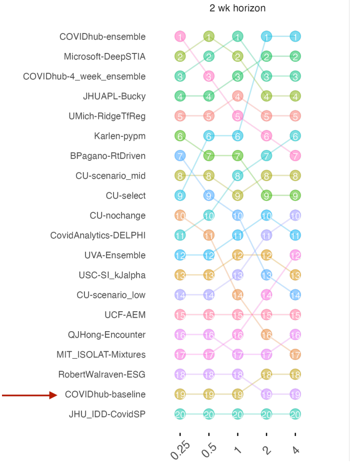
Model rankings



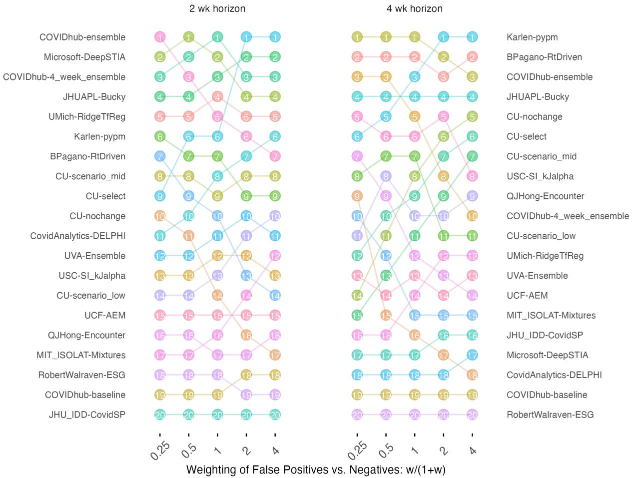
Model ranking and aggregation



Model ranking and aggregation



Model ranking and aggregation



Extensions

Alternative ensemble

- We have experimented with alternative ensembles but not really found a clearly better-performing set of weights.

Extensions

Alternative ensemble

- We have experimented with alternative ensembles but not really found a clearly better-performing set of weights.
- Distributional challenges

Extensions

Alternative ensemble

- We have experimented with alternative ensembles but not really found a clearly better-performing set of weights.
- Distributional challenges

States

- Qualitatively similar results

Methods

Results

Discussion

Conclusions

1. We hope this work encourages thinking about the best ways to link predictive models to actions.

Conclusions

1. We hope this work encourages thinking about the best ways to link predictive models to actions.
2. There are remains an unmet need to model changes in trajectory (and clearly communicate corresponding uncertainty).

Conclusions

1. We hope this work encourages thinking about the best ways to link predictive models to actions.
2. There are remains an unmet need to model changes in trajectory (and clearly communicate corresponding uncertainty).
3. For our metrics, ensemble models continue to perform best, but have considerable uncertainty.

Limitations and next steps

1. **Next steps**
   - Optimized ensemble
   - More complex decision rules
   - Fire alarms rather than forecasts
   - Regression discontinuity for when actions change trajectory

Limitations and next steps

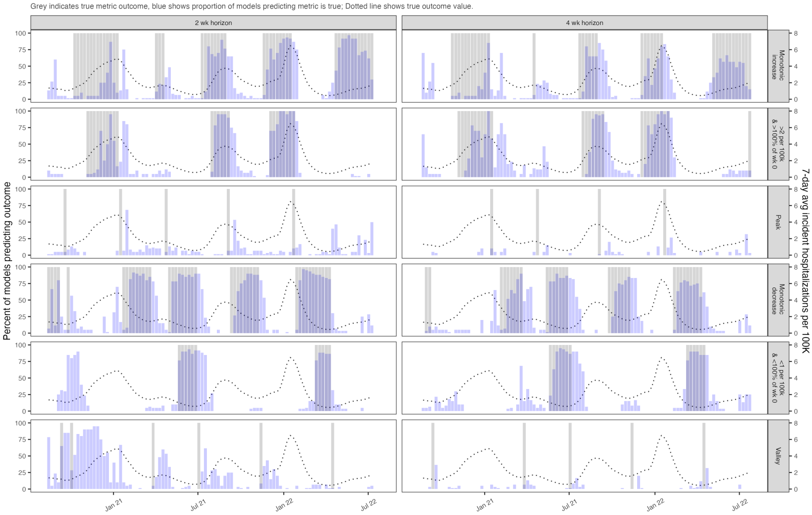
1. **Next steps**
   - Optimized ensemble
   - More complex decision rules
   - Fire alarms rather than forecasts
   - Regression discontinuity for when actions change trajectory
2. **Limitations**
   - Not a proper scoring rule – best way to fit models?
   - Unusual data!

Thank you!

Questions?

Feel free to reach out: alyssa_bilinski@brown.edu

Performance over time



Return

Metrics

Monotonic increase

	2 week horizon						4 week horizon					
	Prevalence	Accuracy	NPV	Specificity	PPV	Sensitivity	Prevalence	Accuracy	NPV	Specificity	PPV	Sensitivity
Baseline	47	50	52	60	46	38	35	55	63	74	29	19
Top 20	47	64	62	83	70	43	35	69	71	88	60	33
4-week ensemble	47	69	65	89	79	47	35	76	75	95	83	42
CDC ensemble	47	73	69	91	83	53	35	73	74	88	67	44
Ensemble	47	70	66	91	81	47	35	74	75	89	70	44
Trained ensemble	47	66	65	79	69	51	35	70	73	85	60	42

>20 per 100k & >100% of wk 0

	2 week horizon						4 week horizon					
	Prevalence	Accuracy	NPV	Specificity	PPV	Sensitivity	Prevalence	Accuracy	NPV	Specificity	PPV	Sensitivity
Baseline	37	63	70	71	50	49	43	60	63	69	54	48
Top 20	37	75	79	83	68	63	43	68	70	78	65	56
4-week ensemble	37	75	78	84	69	59	43	73	70	90	79	50
CDC ensemble	37	78	81	86	73	65	43	77	76	88	80	64
Ensemble	37	79	82	86	74	68	43	77	76	88	80	64
Trained ensemble	37	78	84	81	69	73	43	73	76	76	68	68

Peak

	2 week horizon						4 week horizon					
	Prevalence	Accuracy	NPV	Specificity	PPV	Sensitivity	Prevalence	Accuracy	NPV	Specificity	PPV	Sensitivity
Baseline	10	73	89	80	5	10	5	93	96	97	25	20
Top 20	10	82	90	91	9	8	5	91	95	95	9	9
4-week ensemble	10	82	89	91	0	0	5	94	96	98	33	20
CDC ensemble	10	83	89	92	0	0	5	93	95	98	0	0
Ensemble	10	83	89	92	0	0	5	94	95	99	0	0
Trained ensemble	10	88	90	98	0	0	5	87	95	92	0	0

Return

Metrics

Monotonic decrease

	2 week horizon						4 week horizon					
	Prevalence	Accuracy	NPV	Specificity	PPV	Sensitivity	Prevalence	Accuracy	NPV	Specificity	PPV	Sensitivity
Baseline	33	62	67	87	31	12	23	76	78	97	33	4
Top 20	33	71	80	75	55	63	23	69	84	75	37	49
4-week ensemble	33	74	92	67	57	88	23	70	91	67	41	78
CDC ensemble	33	76	92	70	59	88	23	72	92	70	43	78
Ensemble	33	76	92	70	59	88	23	74	92	72	45	78
Trained ensemble	33	80	86	84	69	73	23	75	85	84	46	48

<10 per 100k & <100% of wk 0

	2 week horizon						4 week horizon					
	Prevalence	Accuracy	NPV	Specificity	PPV	Sensitivity	Prevalence	Accuracy	NPV	Specificity	PPV	Sensitivity
Baseline	10	91	91	100	100	10	12	88	88	100		0
Top 20	10	90	96	93	51	67	12	86	94	91	44	54
4-week ensemble	10	94	98	96	67	80	12	87	94	91	47	58
CDC ensemble	10	93	97	96	64	70	12	87	92	93	45	42
Ensemble	10	93	97	96	64	70	12	87	92	93	45	42
Trained ensemble	10	95	99	96	69	90	12	88	96	90	50	75

Valley

	2 week horizon						4 week horizon					
	Prevalence	Accuracy	NPV	Specificity	PPV	Sensitivity	Prevalence	Accuracy	NPV	Specificity	PPV	Sensitivity
Baseline	10	67	90	71	10	30	5	90	95	95	0	0
Top 20	10	71	90	76	10	25	5	93	95	98	14	7
4-week ensemble	10	79	90	87	8	10	5	92	96	96	20	20
CDC ensemble	10	84	93	89	29	40	5	95	96	99	50	20
Ensemble	10	79	92	84	18	30	5	95	96	99	50	20
Trained ensemble	10	66	89	71	7	20	5	88	95	93	0	0

Return