PandemicPrediction: three-year temporal validation of SEEK-Cover models during the Covid pandemic

Egill A. Fridgeirsson[1], Jenna M. Reps[1,2]

[1]Department of Medical Informatics Erasmus University Medical Center, Rotterdam, the Netherlands

[2]Johnson & Johnson, Raritan, NJ, USA

Background

When SARS-CoV-2 emerged no disease-specific prognostic tools were available and data on patients infected with SARS-CoV-2 were limited. Models that were able to identify which patients were at risk of severe infection due to SARS-CoV-2 early in the pandemic were needed to help prioritize healthcare and interventions. One approach to enable the development of a model with limited data was to use a disease proxy such as influenza. Williams et al [1] developed the Seek-COVER models at the start of the pandemic. They used influenza as a proxy and developed three models. All of them used an outpatient visit with influenza as the index date and used features from the year prior to index. COVER-H predicted hospitalization, COVER-I predicted respiratory failure or insufficiency and COVER-F predicted death. These models were parsimonious and used a set of cohort covariates as features along with demographic information. Data driven counterparts were developed as well that used all available conditions and drugs. The limited amount of data for patients infected with SARS-CoV-2 early into the pandemic were used to validate the influenza models and showed that the models appeared to perform adequately.

Seek-COVER models did not include an updating strategy. During the COVID-19 pandemic, conditions evolved rapidly: SARS-CoV-2 spread quickly worldwide, vaccines were developed and deployed, and successive variants emerged and became predominant at different times. These dynamics can induce distributional shifts that affect model performance. It is therefore unclear whether the performance of the influenza model developed at the start of the pandemic deteriorated over time. In this study, we perform temporal validation to examine how performance changes under distribution shift.

Methods

The Seek-COVER models were developed on Optum's Clinformatics(R) Extended Data Mart, Date of Death (Optum Clinformatics) which we used as well for this study. Our target cohort was diagnosis of COVID-19 or a positive SARS-Cov-2 test. The outcome definitions were the same as in the original study: hospitalization, fatality and respiratory failure/insufficiency.

Optum Clinformatics is derived from a database of administrative health claims for members of large commercial and Medicare Advantage health plans. It is statistically de-identified under the Expert Determination method consistent with HIPAA and managed according to Optum's customer data use agreements. Administrative claims submitted for payment by providers and pharmacies are verified, adjudicated and de-identified prior to inclusion. This data, including patient-level enrollment information, is derived from claims submitted for all medical and pharmacy health care services with information related to health care costs and resource utilization. The population is geographically diverse, spanning all 50 states. The database currently contains 101M patients with data recorded between 2000 – 2024.

We created fourteen, non-overlapping 3-month COVID-19 cohorts spanning 1 Jan 2020–1 Jun 2023 (population per period 4,480–636,005; overall 4.2 million patients, see Table 1). Model discrimination (AUROC) and expected calibration error (Eavg) were calculated for each period and compared between Seek-COVER (parsimonious models with phenotype features) and data-driven (with condition/drug exposure features) models.

Results

Table 1: Validation samples and outcomes per period

Period	Sample size	Outcome events		
		Critical Illness (%)	Death (%)	Hospitalization
2020-Q1	5009	1,482 (33.1%)	90 (1.8%)	1,344 (28.0%)
2020-Q2	117703	11,687 (10.7%)	5,752 (4.9%)	11,528 (9.9%)
2020-Q3	187810	14,279 (8.0%)	3,859 (2.1%)	14,583 (7.8%)
2020-Q4	469100	34,646 (7.7%)	8,604 (1.8%)	36,080 (7.7%)
2021-Q1	355454	29,382 (8.6%)	7,178 (2.0%)	26,712 (7.5%)
2021-Q2	110868	9,258 (8.7%)	1,287 (1.2%)	8,408 (7.6%)
2021-Q3	361455	31,492 (9.1%)	5,585 (1.5%)	27,954 (7.8%)
2021-Q4	413284	23,285 (5.9%)	4,605 (1.1%)	21,798 (5.3%)
2022-Q1	636005	25,535 (4.2%)	7,498 (1.2%)	26,441 (4.2%)
2022-Q2	362324	5,554 (1.6%)	1,233 (0.3%)	6,460 (1.8%)
2022-Q3	491615	9,340 (2.0%)	2,359 (0.5%)	10,928 (2.2%)
2022-Q4	330082	8,970 (2.9%)	2,317 (0.7%)	9,977 (3.1%)
2023-Q1	258015	8,361 (3.5%)	2,181 (0.8%)	8,091 (3.2%)

The discrimination performance for the models is in Figure 1. For fatality prediction the parsimonious model (Cover F) was superior to its data driven analog in 8/14 periods but performance converged in 2022. For hospitalization the parsimonious model (Cover H) outperformed the data driven model in all periods. For respiratory failure the models had the same performance until late year 2021 when the data driven model outperformed the parsimonious model (Cover I)

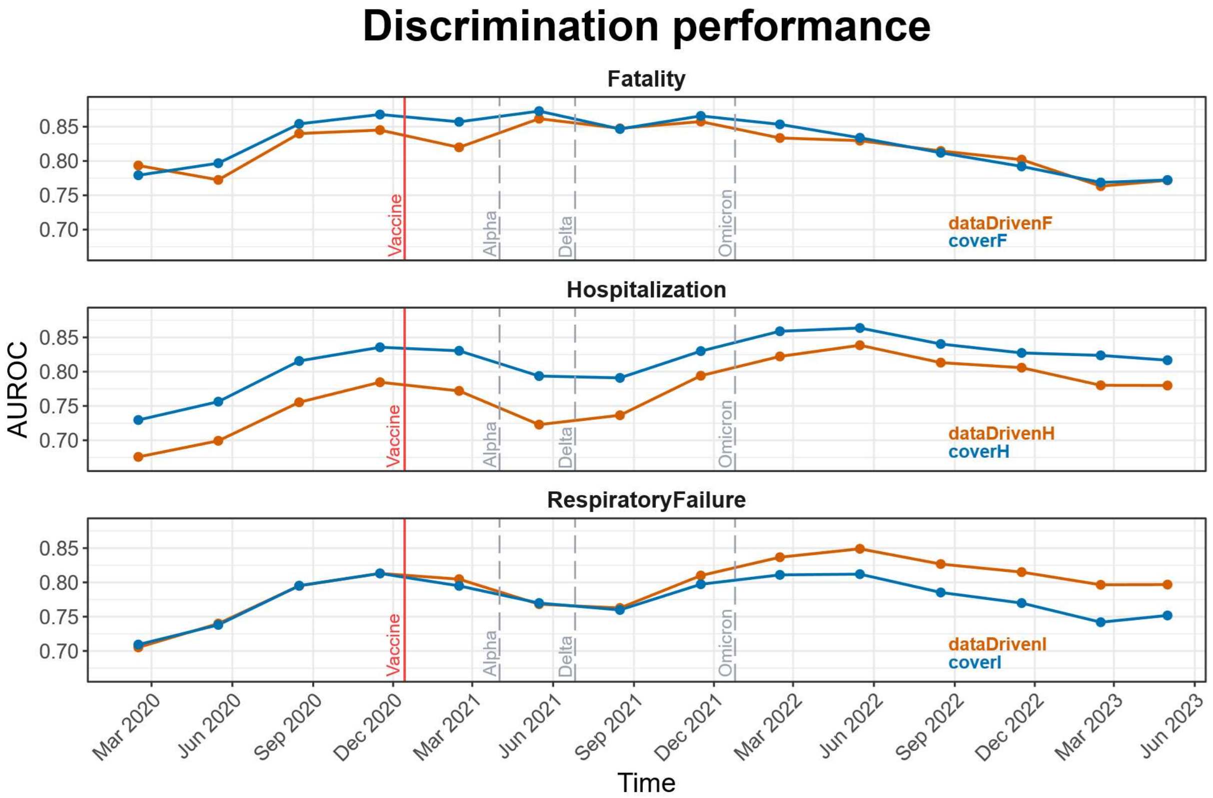

Figure 1: Discrimination performance in terms of the area under the receiver operating characteristics curve (AUROC) for both data driven and parsimonious models. The red vertical line shows when the vaccine first became available to the US population, and the grey lines show when new variants of the virus took became dominant (>50% of cases).

The calibration performance is summarized in Figure 2 using expected calibration error (Eavg; lower is better). For fatality, Eavg was <0.025 in every time window (median 0.003), with similar calibration for the data-driven and parsimonious models. For hospitalization, calibration was worse early in 2020 (ECE 0.26 in the first period) but improved thereafter (median 0.025); the data-driven model showed lower ECE than the parsimonious model. For respiratory failure there was a similar early spike (ECE 0.25 in the first period) with low values thereafter (median 0.032), and calibration was similar between models.

All models exhibited temporal drift: discrimination peaked during the Alpha/Delta waves (Q4-2020/Q1-2021) and fell by ≈0.07 by early-2023, in parallel with mass vaccination, therapeutic advances and the emergence of the Omicron lineage.

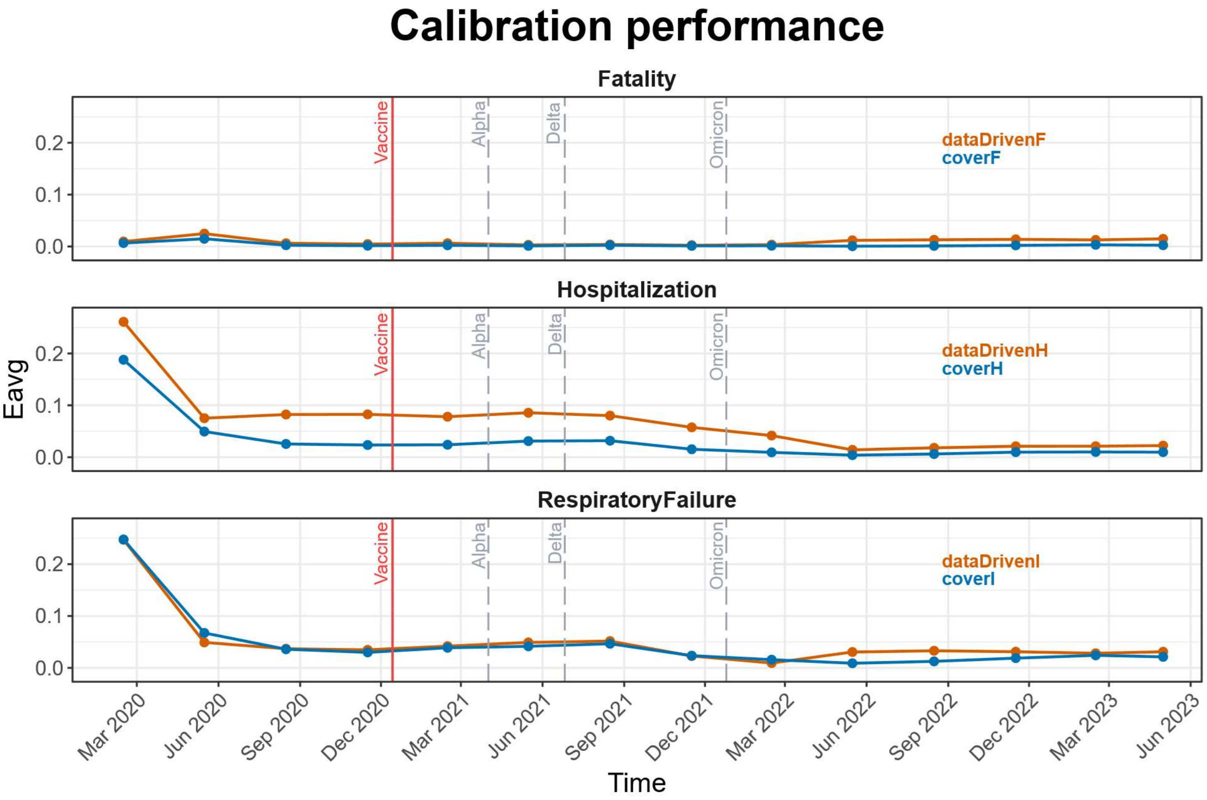

Figure 2: Calibration in terms of the expected calibration error (Eavg). The red vertical line shows when the vaccine became available to the US population and the grey lines show when new variants of the virus became dominant (>50% of cases).

Conclusion

Simple influenza-based COVER scores transported unexpectedly well to COVID-19, offering ready-to-use, well-calibrated risk stratification from the first day of the pandemic. Their stable performance for death and hospitalization suggests that a small set of pathogen-agnostic predictors captures much of the host susceptibility signal, while richer feature sets can add value as disease presentation evolves (observed for critical care after 2021). Continuous monitoring and occasional recalibration—not wholesale re-training—may therefore be sufficient when repurposing pre-existing respiratory-severity models, an insight that can inform preparedness for future outbreaks.

High calibration error for 2020Q1 could be explained by very high outcome rates during that period for hospitalizations and respiratory failure. Much higher than seen during model development.

In future work it would be interesting to investigate whether a model developed using patients infected with SARS-CoV-2, instead of the influenza proxy, would have better performance over time and at what point in time there would have been sufficient data to develop a model using SARS-CoV-2 infected patient data.

References

1. Williams, R.D., Markus, A.F., Yang, C., Duarte-Salles, T., DuVall, S.L., Falconer, T., Jonnagaddala,

J., Kim, C., Rho, Y., Williams, A.E. and Machado, A.A., 2022. Seek COVER: using a disease proxy to rapidly develop and validate a personalized risk calculator for COVID-19 outcomes in an international network. *BMC Medical Research Methodology*, *22*(1), p.35