Supplementary Appendix A: Study protocol

COVID-19 Low risk AI RCT		
CLINICALTRIALS. GOV	Clinicaltrials.gov status	Live
	Approval number, if assigned	NCT04570488
	Lay title	Predicting favorable outcomes in hospitalized COVID-19 patients
INTRODUCTION	Background	
	Problem analysis	Can we use predictive analytics to predict which COVID-19+ patients are at low risk for an adverse event (ICU transfer, intubation, mortality, hospice discharge, re-presentation to the ED, oxygen requirements exceeding nasal cannula at 6 L/min) in the next 96 h?
	Baseline data	Median 10 new green discharges per day; median LOS from green to discharge 3.3
	Observations	
	Objective(s)	To assess if display of low risk of adverse event can safely reduce length of stay and plan for discharge.
	Project location	NYU Langone Health
	Start date	5/15/20
	Planned end study date	200 d, ending in Dec
	Key stakeholders	Hospital leadership
	Project lead	Dr. Jonathan Austrian
	Team members	Yin Aphinyanaphongs, Narges Razavian, Vincent Major, Vuthy Nguy, Peter Stella, Michael Quinn
INTERVENTIONS	Intervention A versus Intervention B description	Display of risk score/ colored flag in Epic patient list column versus no display ("hidden"); will be viewable to all frontline workers
		b) Green: 90% chance of no adverse event @ 96 h
		c) Orange: 67% chance of no adverse event @ 96 h
		d) Red: 8% chance of no adverse event @ 96 h
		e) Missing Data: Incomplete vitals (including spO2) in last 12 h OR no CBC w diff/CRP/LDH/BMP
		f) Hidden: calculation not displayed at random (QI project)
		N/A: patient excluded from calculation: age < 18; no COVID+ infection flag
OUTCOMES	Primary outcome	Reduction in median days from first low-risk (green) score to discharge (GTD)
	Rationale	Once patient is low risk for an adverse event, care team can plan for discharge
	Is this outcome currently routinely captured in clinical care	Yes, using Epic (prior to QI project, green score did not exist)
	Baseline performance	Median 10 new green discharges per day; median LOS from green to discharge (GTD) 3.3
	Minimum clinically important effect size	Median −0.5 d
	Secondary outcome(s)	1. Reduction in LOS for green patients that have not been in the ICU

(*Continued*)

(*Continued*)

COVID-19 Low risk AI RCT		
		2. Reduction in GTD versus LOS for all green patients discharged alive vs all patients discharged alive
		3. No change in 30 d re-ED presentation or hospital admission rate for cohort
	Balancing outcome(s)	Unintended consequences (re-presentation to ED, readmission, mortality, post-discharge mortality)
	Sub analyses	Differences by site in the proportion of patients who turn green (all scored patients: 37% Tisch/Kimmel, 31% Long Island, 21% Brooklyn, 11% Orthopaedics) and their GTD vs. LOS.
	Demographic characteristics	Age, race, ethnicity, sex, admitting location
	Sensitivity analyses	Deaths as censor events (competing risk?)
	Unintentional consequences	Longer LOS for patients who are never green worsening readmissions, re-presentations
	Conditions for continuing/terminating the project	Unintended consequences (re-presentation to ED, readmission, mortality, post-discharge mortality)
	Specify factors to consider if no significant difference found for the primary outcome (e.g., any improvement in primary outcome, secondary outcomes).	
PARTICIPANTS	Study population definition	Adult hospitalized COVID19+ patients predicted to have no adverse event at 96 events with a threshold at 90% PPV, with at least one green score during their admission who are discharged alive and have not been in the ICU
	Exclusion criteria	Age <18 y.
	Expect *N*/wk	10
	N required to reach desired effect size with 80% power	500 per arm, 1,000 total
	Anticipated screening	2,000–4,000 (screening to end once 1,000 patients who have been discharged with at least one green are reached)
RANDOMIZATION	Unit of randomization (patient, provider, hospital-level)	Patient level randomization
	Allocation ratio	1:1, Odd OR even last digit of Epic Enterprise ID
	Sequence generation	
	Method for generating the random allocation sequence	Odd OR even last digit of Epic Enterprise ID
	Type of randomization, details of any restriction (e.g., blocking and blocking size)	Pseudo-randomization
	Allocation concealment mechanism	
	Mechanism for implementing random allocation sequence	Epic
	Implementation	
	Who will generate random allocation sequence	Epic

(*Continued*)

COVID-19 Low risk AI RCT		
	Who will enroll participants	Epic
	Who will assign participants to interventions	Epic
	Blinding	
	Patient? Y/N	Y
	Provider? Y/N	N
	Investigator? Y/N	N
	Data Analyst? Y/N	Y
DATA ANALYSIS	Analytical approach	
	Rationale	

Supplementary Appendix B: Clinical Decision Support

The appearance of the risk score within clinical decision support (CDS) is restricted by the functionality of the Epic EHR. When added to a patient list (►**Appendix Fig. B1**), the risk score appears as a color-coded oval (green, orange, or red) containing the numeric risk on a scale of 0 to 100 for patients allocated to the intervention arm. Those in the control arm are uncolored displaying the word "Hidden." Patients with missing data cannot have a score created and are uncolored displaying "Missing data."

When a user hovers over the score, an explanatory bubble expands (►**Appendix Fig. B2**). The bubble contains the current score, a trendline of recent scores, and a table containing the nine largest contributing factors (in magnitude), their contribution and current value.

The second channel users can view model risk scores is from within a COVID-19 specific summary report where the risk scores are only one component (►**Appendix Fig. B3**).

Supplementary Appendix C: Supplementary Results

Length of stay is a continuous, nonzero quantity that is often non-uniformly distributed. The primary and secondary outcomes of this study, gLOS and LOS are non-uniform (►**Appendix Fig. C1**). The pre-specified decision to use Mann-Whitney tests to evaluate any intervention effect compares the median rather than the shape of the distribution. A planned Gamma regression would be more sensitive to distributional differences away from the median. However, ►**Appendix Fig. C1** suggests little difference in the distributions of gLOS and LOS.

Planned Gamma Regression

The planned Gamma regression evaluation would help control for residual imbalance between treatment groups or temporal effects for example. Possible confounding variables were investigated by comparing the medians and interquartile ranges (IQR) for demographic, geographic, and temporal variables (►**Appendix Table C1**).

Assembling sex, age (2nd order polynomial), location, month of first green score, and primary symptoms into a gamma regression found the intervention was not a significant predictor of gLOS (►**Appendix Table C2**). Age (nonlinear), later months, and one hospital location were significant.

Unplanned Secondary Analyses

Stratification by Location

Stratifying the primary cohort into the four hospital locations reveals large location-specific trends in the intervention effect on gLOS (►**Appendix Table C3**) where the median effect at hospital Tisch/Kimmel exceeds 0.5 days but is underpowered (control, $n = 140$: 3.82 [1.89–6.70] vs. intervention, $n = 130$: 3.17 [1.38–5.82], $p = 0.07$). Operational differences between hospital locations such as resourcing, role composition of care teams, cohorting of patients with COVID-19 as well as gaps in outreach of this AI system could cause this observed trend. Note that location Orthopaedics consisted of a very small percentage of the total study patients (1.8%) but the median gLOS was very long as this location is an Orthopaedics surgical hospital without an emergency department.

Prolongation of the study period to enroll 1,000 patients at Tisch/Kimmel hospital (extending to March 31, 2021) reduces the median gLOS, suggesting pronounced temporal effects, but no detectable difference is observed (control, $n = 478$: 2.61 [1.33–4.77], intervention, $n = 522$: 2.42 [1.23–4.83], $p = 0.3$).

Temporal Shift

Related works have found temporal changes in LOS and survival[24,25] through the pandemic that may extend to gLOS. Splitting the cohort into four groups by the month of their first green score reveals a general decrease in gLOS and

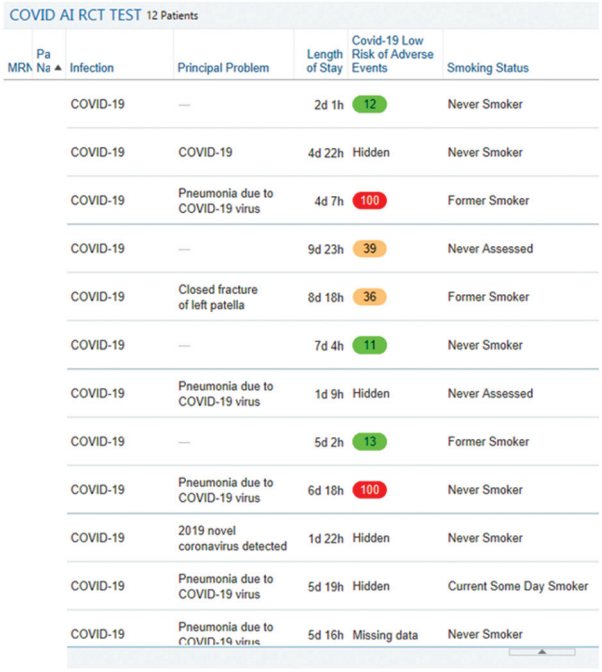
a *U*-shape in LOS (►Appendix Table C1). The gamma regression highlights the latter two groups as significant predictors. Grouping into these two halves, early (May–August, $n = 482$) and late (September–December, $n = 528$) reveals a complex relationship between time and study group. The control group experiences a negligible decrease in gLOS over time. But, the intervention group starts higher and decreases dramatically (►Appendix Table C4; May–Aug: 3.66 [2.00–6.40] vs. Sep–Dec: 2.82 [1.52–5.14], $p = 0.006$). In both early and late periods, the intervention had no detectable effect as the interquartile ranges were similar.

Together these results suggest that any effect of the intervention reducing gLOS is smaller than temporal changes yielded by factors such as reduced surge burden, expanded testing, improved treatment, and shifts in hospitalization rates by hospital location. A non-randomized study, such as a pre-post analysis, would have been susceptible to wrongly attributing these changes to the intervention.

Primary Respiratory Symptoms

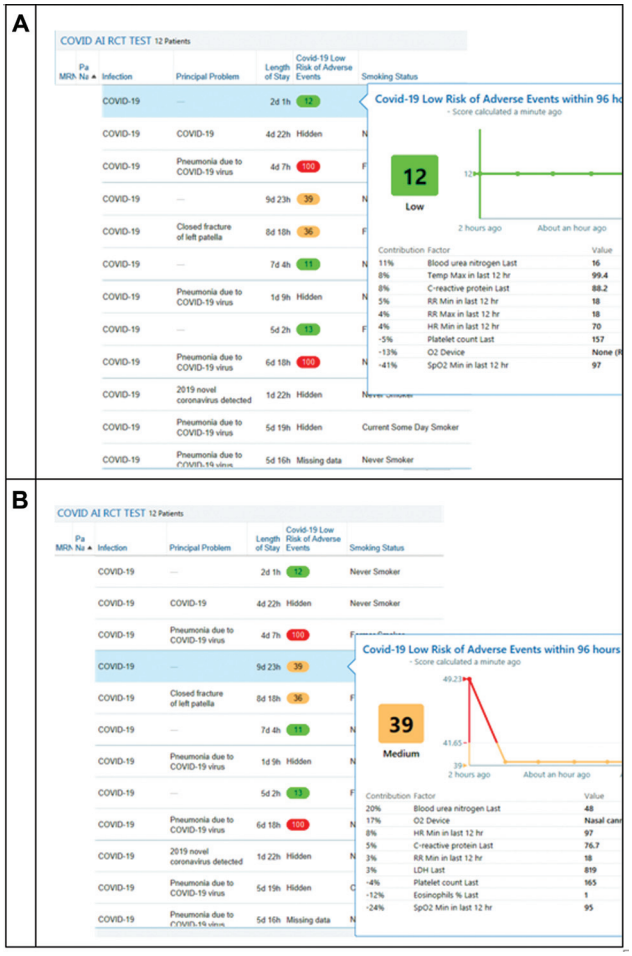
During the study, the weekly census of patients with COVID-19 dropped and rose again (►Fig. 2). As the COVID-19 burden waned, more patients with secondary COVID-19 infection were observed as elective surgeries resumed, for example. ICD-10 diagnosis codes were used to identify admissions with a primary diagnosis consistent with the typical symptoms of COVID-19 (e.g., acute respiratory failure = J96.01, pneumonia = J18.9 or U07.1, shortness of breath = R06.02, see ►Appendix Table C5 for complete list).

Primary respiratory symptoms were found for half of patients ($n = 525$) who had longer gLOS and LOS (►Appendix Table C1). The intervention had no detectable difference among patients who did or did not have a primary symptom consistent with COVID-19 (►Appendix Table C6).

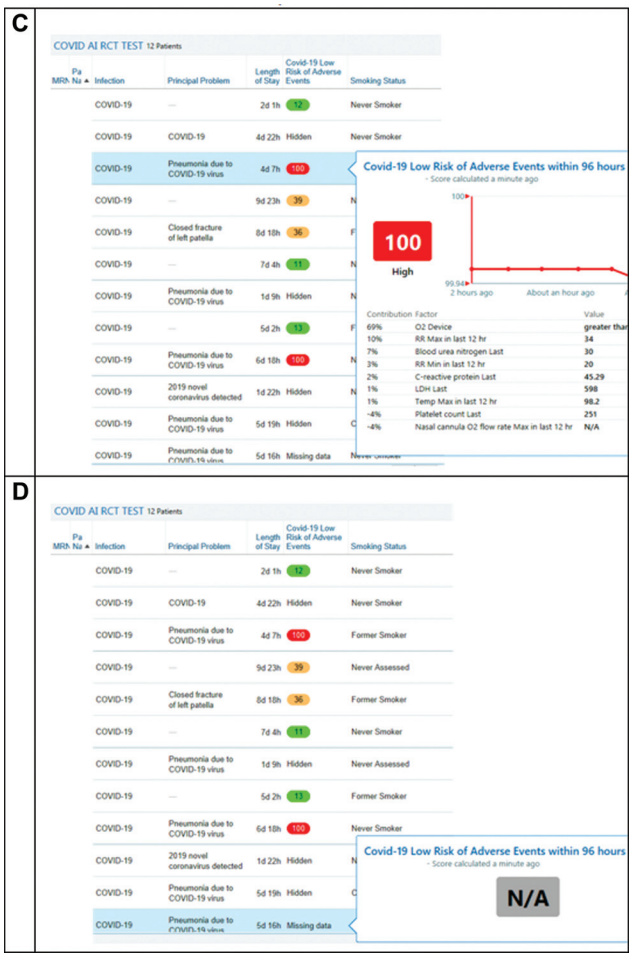

COVID AI RCT TEST 12 Patients

MRN	Pa Na ▲	Infection	Principal Problem	Length of Stay	Covid-19 Low Risk of Adverse Events	Smoking Status
		COVID-19	—	2d 1h	12	Never Smoker
		COVID-19	COVID-19	4d 22h	Hidden	Never Smoker
		COVID-19	Pneumonia due to COVID-19 virus	4d 7h	100	Former Smoker
		COVID-19	—	9d 23h	39	Never Assessed
		COVID-19	Closed fracture of left patella	8d 18h	36	Former Smoker
		COVID-19	—	7d 4h	11	Never Smoker
		COVID-19	Pneumonia due to COVID-19 virus	1d 9h	Hidden	Never Assessed
		COVID-19	—	5d 2h	13	Former Smoker
		COVID-19	Pneumonia due to COVID-19 virus	6d 18h	100	Never Smoker
		COVID-19	2019 novel coronavirus detected	1d 22h	Hidden	Never Smoker
		COVID-19	Pneumonia due to COVID-19 virus	5d 19h	Hidden	Current Some Day Smoker
		COVID-19	Pneumonia due to COVID-19 virus	5d 16h	Missing data	Never Smoker

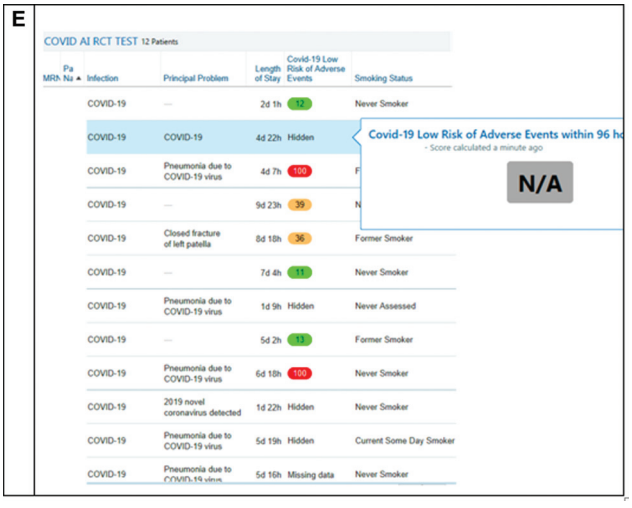
Appendix Fig. B1 Display of the risk score via the patient list column.



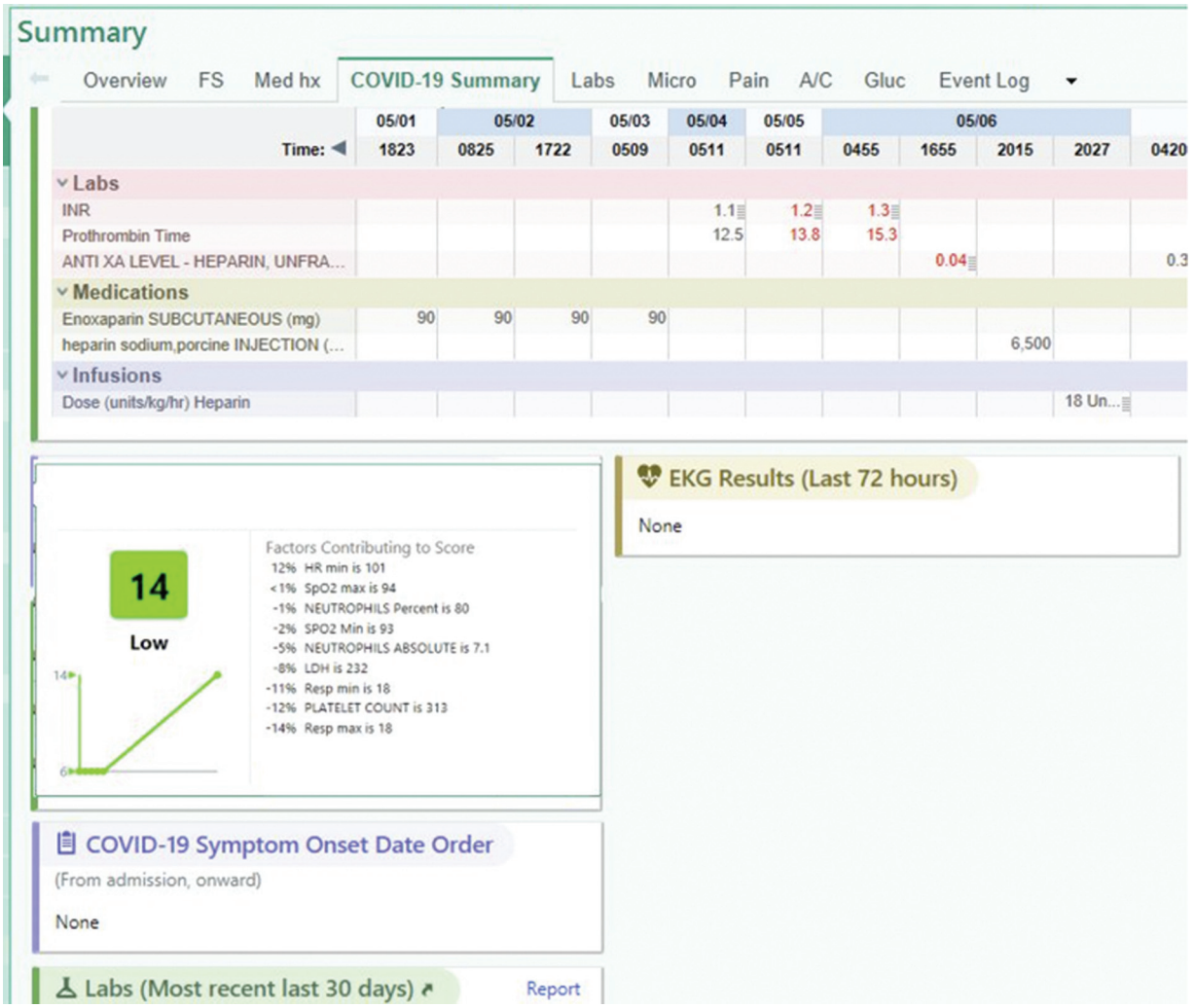
Appendix Fig. B2 Explanatory bubble for patients with **(A)** green, **(B)** orange, **(C)** red, **(D)** missing scores for patients allocated to the intervention group, and **(E)** patients allocated to the control arm.



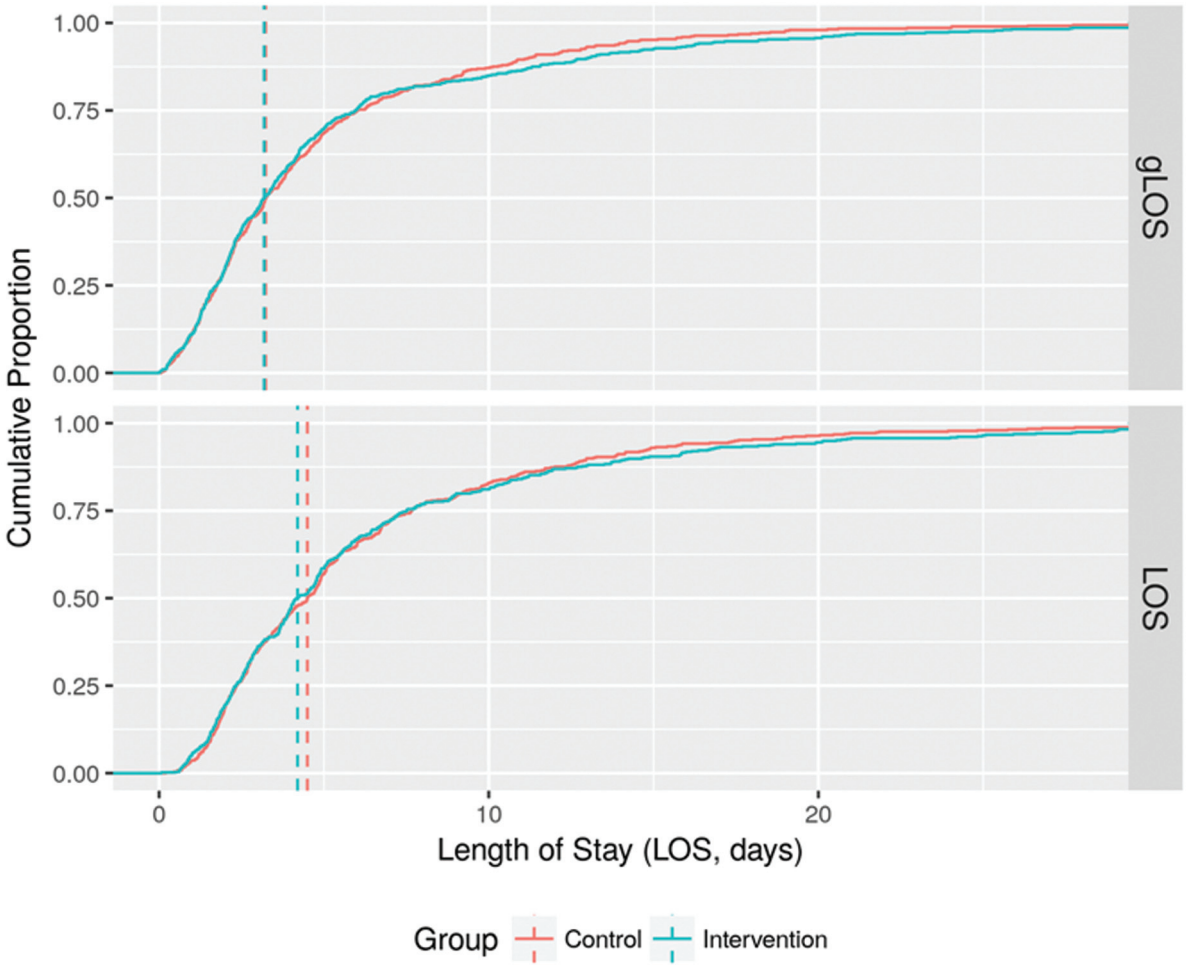
Appendix Fig. B2 (contd.)



Appendix Fig. B2 (contd.)



Appendix Fig. B3 Display of the risk score as one component of a larger COVID-19 summary report.



Appendix Fig. C1 Cumulative distribution of gLOS and LOS by allocated study group with median highlighted.

Appendix Table C1 gLOS and LOS stratified by demographics, location, month of first green score, and primary symptoms

Strata		gLOS	LOS
Sex			
Female	$n = 526$	3.02 [1.78–5.52]	3.78 [2.15–6.78]
Male	$n = 484$	3.53 [1.66–6.35]	4.90 [2.75–8.78]
Age			
Age $\leq$52	$n = 340$	2.31 [1.56–3.76]	2.85 [1.99–5.01]
52 <Age $\leq$69	$n = 337$	3.60 [1.62–5.89]	4.71 [2.52–7.81]
Age >69	$n = 333$	4.52 [2.24–9.03]	5.48 [3.10–11.01]
Ethnicity			
Hispanic	$n = 258$	2.70 [1.55–4.54]	3.22 [2.02–5.56]
Not Hispanic	$n = 752$	3.47 [1.83–6.44]	4.79 [2.55–8.34]
Race			
African American (Black)	$n = 122$	3.04 [1.43–5.43]	4.04 [2.05–7.24]
Asian	$n = 61$	2.47 [1.24–5.80]	4.78 [2.49–6.87]
Native American	$n = 12$	1.92 [1.60–3.79]	3.03 [1.82–4.99]
Other	$n = 295$	2.98 [1.68–5.04]	3.75 [2.21–7.10]
Pacific Islander	$n = 9$	2.15 [1.56–9.90]	2.61 [1.99–10.24]
White	$n = 511$	3.64 [1.96–6.41]	4.79 [2.60–7.81]
Location			
Tisch/Kimmel	$n = 270$	3.38 [1.64–6.21]	4.7 [2.52–7.76]
Orthopaedics	$n = 18$	15.69 [10.93–19.94]	15.83 [10.95–20.20]
Brooklyn	$n = 416$	2.70 [1.70–5.00]	3.33 [2.10–6.61]
Long Island	$n = 306$	3.83 [2.00–6.03]	5.02 [2.82–7.72]
Month of first green score			
May-June	$n = 285$	3.81 [2.03–8.42]	4.70 [2.33–9.21]
July-August	$n = 197$	3.17 [1.92–4.97]	3.86 [2.29–6.73]
Sep-October	$n = 193$	3.20 [1.75–5.37]	4.40 [2.27–6.29]
Nov-December	$n = 335$	3.05 [1.38–5.77]	4.76 [2.54–8.06]
Primary symptom			
Respiratory/COVID-19	$n = 525$	3.64 [1.88–6.42]	5.01 [2.84–8.10]
Other	$n = 485$	2.90 [1.71–5.11]	3.44 [2.08–6.76]

Note: Age was binned into tertiles, month of first green score into bins of two calendar months. Underlined values denote the largest gLOS and LOS for each category.

Appendix Table C2 Gamma regression coefficients

Covariate	Coefficient	Std. error	*p*-Value	
Intercept	0.55	0.068	<0.001	
Study group				
Control	Reference			
Intervention	−0.0098	0.012	0.4	
Sex				
Female	Reference			
Male	−0.0073	0.012	0.5	
Age				
Linear	−0.0087	0.0021	<0.001	*
2nd order polynomial	0.000047	0.000016	0.004	*
Location				
Tisch/Kimmel	Reference			
Orthopaedics	−0.091	0.02	<0.001	*
Brooklyn	−0.0077	0.016	0.6	
Long Island	−0.012	0.016	0.5	
Month of first green score				
May-June	Reference			
July-August	0.036	0.019	0.05	.
Sep-October	0.048	0.019	0.01	*
Nov-December	0.035	0.015	0.02	*
Primary symptoms				
Non-COVID-19	Reference			
COVID-19	−0.0022	0.012	0.9	

Note: Statistical significance is indicated (*) for $p < 0.05$ where (.) for $p = 0.05$.

Appendix Table C3 Intervention effect on gLOS stratified by hospital location of admission

Hospital	Control	Intervention	Mann-Whitney *p*-value
Tisch/Kimmel	3.82 [1.89–6.70]	3.17 [1.38–5.82]	0.07
Orthopaedics	14.61 [7.28–17.95]	15.84 [12.31–24.58]	0.4
Brooklyn	2.68 [1.71–4.76]	2.82 [1.63–5.19]	0.5
Long Island	3.99 [1.63–6.22]	3.59 [2.11–6.03]	0.8

Appendix Table C4 Intervention effect on gLOS stratified by month of first green score into early or late period

Month	Control	Intervention	Mann-Whitney *p*-value
May–Aug	3.29 [1.96–6.33]	3.66 [2.00–6.40]	0.7
Sep–Dec	3.21 [1.62–5.84]	2.82 [1.52–5.14]	0.4
Mann-Whitney *p*-value	0.2	0.006	

Appendix Table C5 ICD-10 diagnosis codes used to label diagnoses as consistent with symptoms of COVID-19

ICD-10 Code	ICD-10 Description	%	*N*
U07.1	COVID-19	43.8	442
A41.89	Other specified sepsis	8.6	87
R06.02	Shortness of breath	4.3	43
J96.01	Acute respiratory failure with hypoxia	2.8	28
A41.9	Sepsis, unspecified organism	2.0	20
J18.9	Pneumonia, unspecified organism	1.7	17
J44.1	Chronic obstructive pulmonary disease with (acute) exacerbation	0.3	3
J12.89	Other viral pneumonia	0.1	1
J96.91	Respiratory failure, unspecified with hypoxia	0.1	1
R05	Cough	0.1	1
A41.9	Sepsis, unspecified organism	0	0
B97.29	Other coronavirus as the cause of diseases classified elsewhere	0	0
J06.9	Acute upper respiratory infection, unspecified	0	0
J12.81	Pneumonia due to SARS-associated coronavirus	0	0
J12.9	Viral pneumonia, unspecified	0	0
J21.9	Acute bronchiolitis, unspecified	0	0
J22	Unspecified acute lower respiratory infection	0	0
J80	Acute respiratory distress syndrome	0	0
J98.8	Other specified respiratory disorders	0	0
O99.512O99.52	Diseases of the respiratory system complicating pregnancy, second trimester/childbirth	0	0
R09.2	Respiratory arrest	0	0

Appendix Table C6 Intervention effect on gLOS stratified by primary diagnosis code consistent with the symptoms of COVID-19

Primary symptom	Control	Intervention	Mann-Whitney *p*-value
Non-COVID-19	2.75 [1.76–4.90]	2.98 [1.54–5.36]	0.7
COVID-19	3.97 [1.75–7.13]	3.42 [1.92–6.03]	0.4
Mann-Whitney *p*-value	0.01	0.2	

