Overview

This replication package contains the code and publicly available data, but not the confidential data necessary for replication. The code in this replication package constructs the analysis file from the original data files, obtained from the Regenstrief institute, using Stata. The code also generates all tables and figures included in the manuscript. The code will not run, however, without the confidential data. The code ran in 75 minutes on a high performance server. Much of the time is spent estimated bootstrapped confidence intervals.

Data Availability and Provenance Statements

The project primarily uses confidential and proprietary Indiana Network for Patient Care (INPC) and the Regenstrief Institute COVID-19 registry. These data were shared with Regenstrief Institute by health care providers from across the state of Indiana. The data cannot be shared by us because they contain sensitive health information, and because the terms under which the original data was provided preclude further data sharing. Researchers interested in working with INPC data, or replicating our results should submit a feasibility request through the url https://www.regenstrief.org/feasibility-request/ or a data request through the url https://www.regenstrief.org/data-request/.

The project also uses publicly available Census data describing population counts, and author-collected data from the Indiana COVID dashboard,.

Statement about Rights

I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data and programs contained within this replication package.

Appropriate permission are documented in the LICENSE.txt file.

Summary of Availability

Some data **cannot be made** publicly available.

Details on each Data Source

Publicly available data

- data/covid_report_date.xlsx
    - Daily count of COVID cases and deaths, downloaded from state dashboard, by authors.
- data/ili_cli_codes.csv
    - List of ICD codes describing influenza or COVID-like illnesses, compiled by authors
- data/nhgis0001_ds239_20185_tract.csv

    - Population counts at census tract-age-sex cell, obtained from Manson et al. (2021).
- dxccsr/DXCCSR_v2020-3.CSV
    - Groups diagnosis codes into broader categories,, obtained from the Health Care Utilization Project (CCSR 2021).

Confidential data not provided in the replication package

These data are confidential but may be obtained with a Data Use Agreement with the Regenstrief Institute. Researchers interested in access to the data should submit a feasibility request through the url https://www.regenstrief.org/feasibility-request/ or a data request through the url https://www.regenstrief.org/data-request/.

- rdrp3367_covid_de_v5.csv
    - COVID test-level data describing test date and outcome, extracted from Regenstrief Institute's COVID-19 registry. All COVID-19 tests in the database between January 1, 2020 and December 18 2020.
- rdrp3367_dx_de_v5.csv
    - Diagnostic information – all diagnosis codes associated with hospitalizations in the INPC hospitals, between January 1, 2020 and December 18, 2020, extracted from INPC data hosted by the Regenstrief Institute. Each row is a hospitalization-by-diagnosis code.
- rdrp3367_inp_de_v5
    - Information on all hospitalizations in the INPC (date of admission and discharge, patient identifier, etc.) between January 1, 2020 and December 18, 2020, extracted from INPC data hosted by the Regenstrief Institute. Each row is a hospitalization.
- rdrp3367_demo_de_v5
    - Demographic information, where available, for all patients (a) in the INPC or COVID-19 registry and (b) appearing at least once in the hospitalization or test data above. Data are collected from health care providers in the INPC and from COVID-19 tests (when reported by the testing agency), and passed on to Regenstrief Institute. Each row is a patient.

Dataset list

Data file	Source	Provided
`health/rdrp3367_covidde_v5.csv`	Regenstrief	No
`health/rdrp367_dx_de_v5.csv`	Regenstrief	No
`health/rdrp367_inp_de_v5.csv`	Regenstrief	No
`health/rdrp367_demo_de_v5.csv`	Regenstrief	No
`data/covid_report_date.csv`	IN Dashboard	Yes
`data/nhgis0001_ds239_20185_tract.csv`	Manson et al. (2021)	Yes
`dxccsr/DXCCSR_v2020-3.CSV`	CCSR (2021)	yes

Descriptions of the columns of each data file are available in the tables at the end of this readme.

Computational requirements

- **Software:** Stata (code was last run with Stata MP version 17.0)
- **Controlled randomness:** Random seed is set at line 7 of program main.do
- **Memory and runtime:** Less than 5gb of memory and 24 hours of runtime. However 90% of this time is spent in bootstrap estimation of confidence intervals.
- **Computing environment:** The code was last run on a Intel Xeon Gold 5118 CPU @ 2.30GHz, 48 processors, 12 CPU cores, 755 GBs RAM. However it could run on a desktop.

Description of programs/code

- All programs are in scripts/. The program scripts/main.do executes all code.
- The programs scripts/import* and scripts/clean* prepare the analysis sample. The remaining programs construct all tables and figures in the manuscript (as well as appendix exhibits and additional results).
- The table below lists the tables and figures in the manuscript, the programs that generate them, and the name of the output corresponding to the manuscript exhibit.

License for Code

Appropriate permission are documented in the LICENSE.txt file.

Instructions to Replicators

- Unzip all material.
- Obtain the confidential data and place it in the directory health/.
- Edit `scripts/main.do` line 3 so that it points to the path of the unzipped material.

- Run scripts/main.do to run all programs in sequence.

Details of programs

- age_sex_counts.do
  - Gets state-level population counts in age-sex bins
- import_ili_cli.do
  - Loads the list of ICD-10 codes corresponding to influenza-like and covid-like illness and saves in state format
- import_state_counts.do
  - loads the state's reported covid case counts and saves in state format
- clean_dxccsr.do
  - makes diagnosis groupings for ICD-10 codes
- clean_tests.do
  - cleans up formatting of raw test data, making person-day test panel
- clean_dx.do
  - cleans up formatting of diagnosis information, making hospitalization-level data set with flags for selected diagnosis croups
- clean_inp.do
  - cleans up hospitalization data set
- clean_demo.do
  - cleans up demographic information
- check_sums.do
  - checks that samples sizes are correct
- sum_demo .do
  - reports summary statistics on demographics
- inp_group_age.do
  - reports age distribution of patients admitted for different diagnosis groups
- show_hospital_test_times.do
  - reports distribution of time between test date and admission dates
- weekly_test_rates_wtd.do
  - reports age-weighted distribution of test rates in different populations
- weekly_test_rates.do
  - reports age-unweighted distribution of test rates in different populations
- bootstrap_bounds.do
  - Estimates bounds in 500 bootstrap iterations (age weighted)
- calculate_bs_ci.do
  - Calculates bootstrap-based confidence intervals for weekly intersection bounds (age weighted)
- plot_bounds .do
  - plots the weekly bounds
- width_bounds.do
  - calculates the width of bounds under different assumptions
- outsheet_bounds.do
  - outsheets tables of weekly bounds under different assumptions/samples
- show_small_bias.do
  - shows small finite sample bias in intersction bounds

- bootstrap_bounds_unwtd.do
    - Estimate bounds in 500 bootstrap iterations (age unweighted)
- calculate_bs_ci_unwtd.do
    - Calculates bootstrap-based confidence intervals for weekly intersection bounds (age unweighted)
- outsheet_bounds_unwtd.do
    - outsheets tables of weekly bounds under different assumptions/samples (age unweighted)
- bootstrap_causes.do
    - calculates prevalence bounds in bootstrap samples for specific causes of admission
- table_narrow_causes.do
    - reports prevalence bounds by cause of admission
- study_retests.do
    - estimates statistics on test-retest agreement
- community_rate_validation.do
    - compares county prevalence of hospitalized individuals and full population
- prior_rate_validation.do.do
    - compares prior test rate of hospitalized and non-hospitalized individuals
- compare_to_state.do
    - compares positivity count in data to reports from state
- icli_hosps_pop.do
    - estimates the share of ICLI-hospitalized patients out of total population
- hospital_test_share.do
    - Calculates the share of all tests that are conducted on hospitalized patients
- ado/scalarout.ado
    - A utility for reporting scalars

List of tables and programs

The provided code reproduces all numbers provided in text in the paper, and all tables and figures in the paper (except figure 1, which is not based on a data). The table below lists the programs that produce each exhibit, as well as the line number producing the exhibit, and the name of the relevant output.

All numbers reported in the text also appear in the tables or figures, with three exceptions. First, the average reduction in the width of the identified set is calculated by `width_shrinkage.do`, and reported in the output `tables/width_shrinkage.tex`. Second, the estimated statistics on test reliability are calculated by study_retests.do and reported in `tables/test_retest_rates.csv`. Third, the share of all tests accounted for by hospitalized patients is calculated by `hospital_test_share.do` and reported in `tables/hospital_test_share.tex`.

Exhibits in paper and the programs that create them

Exhibit	Program that creates it	line	output
Panel A. Figures and tables in manuscript			
Figure 2	weekly_test_rates_wtd.do	187	figures/tests_all_groups_wtd.pdf
Figure 3	plot_bounds.do	82	figures/bounds_all_groups_wtd_ci.pdf
table 1	sum_demo.do	156-157	tables/demographics_by_test_and_hopsital.tex
table 2	table_narrow_causes.do	84-85	tables/groups_pool_time.tex
Panel B. Appendix Figures and tables			
Figure A1	compare_to_state.do	49	figures/compare_to_state.pdf
Figure A2	show_hospital_test_times.do	34	figures/hospital_test_times.pdf
Figure E1	icli_hosps_pop.do	34	figures/icli_hosp_rate.pdf
Figure F1	show_small_bias.do	45	figures/flip_rate.pdf
Figure G1	prior_rate_validation.do	201	figures/prior_rates.pdf
Table A1	weekly_test_rates_wtd.do	142-143	tables/test_rate_by_group_wtd.tex
Table A2	weekly_test_rates.do	118-119	tables/test_rate_by_group.tex
Table A3	outsheet_bounds.do	68-73	tables/bounds_by_group_wtd_Mar_Jul.tex
Table A4	outsheet_bounds.do	84-88	tables/bounds_by_group_wtd_Aug_Dec.tex
Table A5	outsheet_bounds_unwtd.do	56-60	tables/bounds_by_group_Mar_Jul.tex
Table A6	outsheet_bounds_unwtd.do	62-66	tables/bounds_by_group_Aug_Dec.tex
Table A7	inp_group_age.do	64-65	tables/hosp_ss_by_group.tex
Table F1	show_small_bias.do	27-28	tables/bias.tex
Table G1	outsheet_bounds.do	126-127	tables/nir_bounds_{b}.tex, b in pop, non_m, non_i. clear_m, clear_i
Table G2	community_rate_validation.do	115-116	tables/county_ind_test_rates.tex

References

Clinical Classifications Software Refined (CCSR). 2021 Healthcare Cost and Utilization Project (HCUP). October 2021. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp.

Steven Manson, Jonathan Schroeder, David Van Riper, Tracy Kugler, and Steven Ruggles. IPUMS National Historical Geographic Information System: Version 16.0 [dataset]. Minneapolis, MN: IPUMS. 2021. http://doi.org/10.18128/D050.V16.0

Acknowledgements

This readme file is based on the AEA Data Editor's template.

Appendix: Details of provided data sets

Fields in health/rdrp3367_covid_de_v5.csv

Field	Meaning
sid	Encrypted patient identifier
lab_date	Date of lab processing of COVIID PCR test
lab_name	Lab type identifier
lab_results	Positive/negative/ambiguous
in_other_cohort	=1 if SID appears in inpatient data

Fields in health/rdrp3367_dx_de_v5.csv

Field	Meaning
sid	Encrypted patient identifier
encounter_de	Identifier for hospitalization
dx_sys_id_text	Description of diagnosis
dx_priority	Priority of diagnosis (primary, etc)
dx_code	ICD-10 code
dx_text	Description of ICD code diagnosis
dx_type_code	Initial or final diagnosis

Fields in health/rdrp3367_inp_de_v5.csv

Field	Meaning
sid	Encrypted patient identifier
encounter_de	Identifier for hospitalization
inst_id_de	Encrypted hospital identifier
admit_time	Time and date of admission
discharge_time	Time and date of discharge
in_other_cohort	=1 if in test data

Fields in health/rdrp3367_demo_de_v5.csv

Field	Meaning
sid	Encrypted patient identifier
race	Patient race
ethnicity	Patient ethnicity
gender	Patient gender
date_of_birth	Patient date of birth
date_of_death	Patient date of death (if known and dead)
zip_code	Patient zip code
geocode_census_block_2010	2010 Census block (derived from patient address)

Fields in dxccsr/covid_report_date.xlsx

Field	Meaning
DATE	Date
COVID_TEST	# COVID-19 tests
COVID_COUNT_BY_SPECIMEN_COLLECTI	# Positive cases
(note: unlisted fields not used in analysis)	

Fields in dxccsr/DXCCSR_v2020-3.CSV

Field	Meaning
icd10cmcode	ICD 10 code
icd10cmcodedescription	Decscription
defaultccsrcategory	Category of ICD 10 code
Defaultccsrcategorydescription	Description of category
Ccsrcategory1	First category
Ccsrcaregory1description	Description of first category
Ccsrcategory2	Second category
Ccsrcaregory2description	Description of second category
Ccsrcategory3	Third category
Ccsrcaregory3description	Description of third category
Ccsrcategory4	Fourth category
Ccsrcaregory4description	Description of fourth category
Ccsrcategory5	Fifth category
Ccsrcaregory5description	Description of fifth category

Fields in data/indianaCTract_age_sex.dta

Field	Meaning
Totpop	# in census tract
totMale	# male in census tract
m_LT5,	# male younger 5, in census tract
f_LT5	# female younger than 5, in census tract
m_{a}_{b}	# male in age range a,b, in census tract
f_{a}_{b}	# female in age range a,b, in census tract
m_gte85	# male 85 or older in census tract
f_gte85	# female 85 or older in census tract