

# COVID-19 Case Surveillance Public Use Data with Geography Utility Summary

Users should consider the level of completeness, including suppression levels when planning their analyses and use of public datasets. Privacy protections will suppress field values to reduce reidentification risks. Completeness varies by jurisdiction (i.e., state, local, and territorial) and time period. Variables are consistently coded to the value “Unknown” when jurisdictions specify in the case data submitted to CDC that the value is unknown, the value “Missing” when jurisdictions do not provide a value, and the value “NA” when the value is suppressed as part of privacy protections.

Dataset version: 2022-01-10.parquet

Total records in dataset: 44,083,363

## Quick Summary

	all_fields	quasi_fields
total_fields	19	8
total_records	44,083,363	44,083,363
total_cells	837,583,897	352,666,904
missing_fields	182,153,168	26,407,279
missing_pct	22%	7%
complete_fields	655,430,729	326,259,625
complete_pct	78%	93%
unknown_fields	38,031,752	20,467,822
unknown_pct	5%	6%
suppressed_fields	24,476,129	21,397,166
suppressed_pct	3%	6%
available_fields	592,922,848	284,394,637
available_pct	71%	81%

## Field-level Utility Summary

	suppressed	suppressed_percent	missing	missing_percent
case_month	4	0.0%	0	0.0%
res_state	1,006	0.0%	0	0.0%
res_county	3,077,957	7.0%	0	0.0%
age_group	457,830	1.0%	381,883	0.9%
sex	1,285,457	2.9%	50,866	0.1%
race	6,689,854	15.2%	3,107,173	7.0%
ethnicity	7,896,524	17.9%	2,634,993	6.0%
death_yn	1,988,534	4.5%	20,232,364	45.9%
records_with_any_field	11,649,424	26.4%	21,404,694	48.6%