

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: 2021-03-31

Total records in dataset: 22,507,139

Quick Summary

	all_fields	quasi_fields
total_fields	19	8
total_records	22,507,139	22,507,139
total_cells	427,635,641	180,057,112
missing_fields	90,389,202	11,807,853
missing_pct	21%	7%
complete_fields	337,246,439	168,249,259
complete_pct	79%	93%
unknown_fields	22,581,935	11,725,242
unknown_pct	5%	7%
suppressed_fields	13,410,092	11,793,921
suppressed_pct	3%	7%
available_fields	301,254,412	144,730,096
available_pct	70%	80%

Field-level Utility Summary

	suppressed	suppressed_percent	missing	missing_percent
case_month	0	0.0%	0	0.0%
res_state	682	0.0%	0	0.0%
res_county	1,615,489	7.2%	0	0.0%
age_group	242,770	1.1%	159,186	0.7%
sex	645,024	2.9%	31,530	0.1%
race	3,754,054	16.7%	1,961,121	8.7%
ethnicity	4,492,415	20.0%	1,664,232	7.4%
death_yn	1,043,487	4.6%	7,991,784	35.5%
records_with_any_field	6,443,392	28.6%	9,120,358	40.5%