

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: 2021-04-30

Total records in dataset: 24,441,351

Quick Summary

	all_fields	quasi_fields
total_fields	19	8
total_records	24,441,351	24,441,351
total_cells	464,385,669	195,530,808
missing_fields	98,422,541	12,735,724
missing_pct	21%	7%
complete_fields	365,963,128	182,795,084
complete_pct	79%	93%
unknown_fields	23,284,836	11,875,097
unknown_pct	5%	6%
suppressed_fields	14,845,678	13,056,345
suppressed_pct	3%	7%
available_fields	327,832,614	157,863,642
available_pct	71%	81%

Field-level Utility Summary

	suppressed	suppressed_percent	missing	missing_percent
case_month	0	0.0%	0	0.0%
res_state	679	0.0%	0	0.0%
res_county	1,788,654	7.3%	0	0.0%
age_group	266,550	1.1%	160,416	0.7%
sex	705,216	2.9%	32,440	0.1%
race	4,163,723	17.0%	2,025,185	8.3%
ethnicity	4,995,992	20.4%	1,665,217	6.8%
death_yn	1,135,531	4.6%	8,852,466	36.2%
records_with_any_field	7,149,181	29.3%	10,002,040	40.9%