

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: 2021-06-07

Total records in dataset: 26,616,527

Quick Summary

	all_fields	quasi_fields
total_fields	19	8
total_records	26,616,527	26,616,527
total_cells	505,714,013	212,932,216
missing_fields	107,420,479	13,628,374
missing_pct	21%	6%
complete_fields	398,293,534	199,303,842
complete_pct	79%	94%
unknown_fields	25,307,914	13,268,263
unknown_pct	5%	6%
suppressed_fields	15,885,216	14,061,381
suppressed_pct	3%	7%
available_fields	357,100,404	171,974,198
available_pct	71%	81%

Field-level Utility Summary

	suppressed	suppressed_percent	missing	missing_percent
case_month	0	0.0%	0	0.0%
res_state	678	0.0%	0	0.0%
res_county	1,823,157	6.8%	0	0.0%
age_group	292,661	1.1%	168,849	0.6%
sex	775,476	2.9%	32,570	0.1%
race	4,493,073	16.9%	2,130,399	8.0%
ethnicity	5,391,156	20.3%	1,701,108	6.4%
death_yn	1,285,180	4.8%	9,595,448	36.1%
records_with_any_field	7,643,130	28.7%	10,736,589	40.3%