

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: 2021-05-10

Total records in dataset: 25,009,120

Quick Summary

	all_fields	quasi_fields
total_fields	19	8
total_records	25,009,120	25,009,120
total_cells	475,173,280	200,072,960
missing_fields	100,876,371	13,035,829
missing_pct	21%	7%
complete_fields	374,296,909	187,037,131
complete_pct	79%	93%
unknown_fields	23,695,617	12,137,182
unknown_pct	5%	6%
suppressed_fields	15,161,753	13,317,696
suppressed_pct	3%	7%
available_fields	335,439,539	161,582,253
available_pct	71%	81%

Field-level Utility Summary

	suppressed	suppressed_percent	missing	missing_percent
case_month	0	0.0%	0	0.0%
res_state	690	0.0%	0	0.0%
res_county	1,843,367	7.4%	0	0.0%
age_group	267,288	1.1%	162,487	0.6%
sex	712,954	2.9%	32,925	0.1%
race	4,243,731	17.0%	2,042,651	8.2%
ethnicity	5,085,158	20.3%	1,671,564	6.7%
death_yn	1,164,508	4.7%	9,126,202	36.5%
records_with_any_field	7,308,861	29.2%	10,277,686	41.1%