

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: 2021-07-19

Total records in dataset: 27,185,885

Quick Summary

	all_fields	quasi_fields
total_fields	19	8
total_records	27,185,885	27,185,885
total_cells	516,531,815	217,487,080
missing_fields	108,484,843	12,930,774
missing_pct	21%	6%
complete_fields	408,046,972	204,556,306
complete_pct	79%	94%
unknown_fields	25,469,702	13,216,311
unknown_pct	5%	6%
suppressed_fields	16,011,582	14,486,167
suppressed_pct	3%	7%
available_fields	366,565,688	176,853,828
available_pct	71%	81%

Field-level Utility Summary

	suppressed	suppressed_percent	missing	missing_percent
case_month	0	0.0%	0	0.0%
res_state	1,137	0.0%	0	0.0%
res_county	1,524,278	5.6%	0	0.0%
age_group	325,273	1.2%	177,080	0.7%
sex	848,955	3.1%	31,315	0.1%
race	4,763,444	17.5%	2,154,125	7.9%
ethnicity	5,688,796	20.9%	1,611,770	5.9%
death_yn	1,334,284	4.9%	8,956,484	32.9%
records_with_any_field	7,645,931	28.1%	10,110,702	37.2%