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: 2021-12-27.parquet

Total records in dataset: 41,648,536

Quick Summary

	all_fields	quasi_fields
total_fields	19	8
total_records	41,648,536	41,648,536
total_cells	791,322,184	333,188,288
missing_fields	172,269,066	24,966,807
missing_pct	22%	7%
complete_fields	619,053,118	308,221,481
complete_pct	78%	93%
unknown_fields	34,567,184	18,956,935
unknown_pct	4%	6%
suppressed_fields	23,583,694	20,681,912
suppressed_pct	3%	6%
available_fields	560,902,240	268,582,634
available_pct	71%	81%

Field-level Utility Summary

	suppressed	suppressed_percent	missing	missing_percent
case_month	12	0.0%	0	0.0%
res_state	1,036	0.0%	0	0.0%
res_county	2,900,746	7.0%	0	0.0%
age_group	458,714	1.1%	365,237	0.9%
sex	1,268,596	3.0%	48,103	0.1%
race	6,443,855	15.5%	2,892,243	6.9%
ethnicity	7,654,184	18.4%	2,460,689	5.9%
death_yn	1,954,769	4.7%	19,200,535	46.1%
records_with_any_field	11,212,321	26.9%	20,292,247	48.7%