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: 2022-02-28.parquet

Total records in dataset: 63,272,363

Quick Summary

	all_fields	quasi_fields
total_fields	19	8
total_records	63,272,363	63,272,363
total_cells	1,202,174,897	506,178,904
missing_fields	270,462,577	41,723,275
missing_pct	22%	8%
complete_fields	931,712,320	464,455,629
complete_pct	78%	92%
unknown_fields	60,556,670	30,367,849
unknown_pct	5%	6%
suppressed_fields	33,536,333	29,263,069
suppressed_pct	3%	6%
available_fields	837,619,317	404,824,711
available_pct	70%	80%

Field-level Utility Summary

	suppressed	suppressed_percent	missing	missing_percent
case_month	4	0.0%	0	0.0%
res_state	1,010	0.0%	0	0.0%
res_county	4,272,254	6.8%	0	0.0%
age_group	523,754	0.8%	724,039	1.1%
sex	1,578,810	2.5%	106,265	0.2%
race	9,723,225	15.4%	4,954,310	7.8%
ethnicity	10,937,497	17.3%	3,826,751	6.0%
death_yn	2,226,515	3.5%	32,111,910	50.8%
records_with_any_field	15,935,925	25.2%	33,776,169	53.4%