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: 2022-01-03.parquet

Total records in dataset: 42,489,148

Quick Summary

	all_fields	quasi_fields
total_fields	19	8
total_records	42,489,148	42,489,148
total_cells	807,293,812	339,913,184
missing_fields	176,132,592	25,516,907
missing_pct	22%	8%
complete_fields	631,161,220	314,396,277
complete_pct	78%	92%
unknown_fields	35,379,771	19,410,993
unknown_pct	4%	6%
suppressed_fields	23,865,475	20,914,179
suppressed_pct	3%	6%
available_fields	571,915,974	274,071,105
available_pct	71%	81%

Field-level Utility Summary

	suppressed	suppressed_percent	missing	missing_percent
case_month	4	0.0%	0	0.0%
res_state	983	0.0%	0	0.0%
res_county	2,950,313	6.9%	0	0.0%
age_group	457,993	1.1%	368,834	0.9%
sex	1,272,656	3.0%	49,462	0.1%
race	6,521,962	15.3%	2,957,179	7.0%
ethnicity	7,741,327	18.2%	2,517,729	5.9%
death_yn	1,968,941	4.6%	19,623,703	46.2%
records_with_any_field	11,356,132	26.7%	20,734,808	48.8%