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: 2021-12-14

Total records in dataset: 40,314,956

Quick Summary

	all_fields	quasi_fields
total_fields	19	8
total_records	40,314,956	40,314,956
total_cells	765,984,164	322,519,648
missing_fields	166,493,514	24,271,942
missing_pct	22%	8%
complete_fields	599,490,650	298,247,706
complete_pct	78%	92%
unknown_fields	33,354,832	18,328,928
unknown_pct	4%	6%
suppressed_fields	22,720,880	19,914,174
suppressed_pct	3%	6%
available_fields	543,414,938	260,004,604
available_pct	71%	81%

Field-level Utility Summary

	suppressed	suppressed_percent	missing	missing_percent
case_month	12	0.0%	0	0.0%
res_state	944	0.0%	0	0.0%
res_county	2,805,762	7.0%	0	0.0%
age_group	437,950	1.1%	361,336	0.9%
sex	1,217,395	3.0%	47,098	0.1%
race	6,186,258	15.3%	2,804,444	7.0%
ethnicity	7,377,166	18.3%	2,392,743	5.9%
death_yn	1,888,687	4.7%	18,666,321	46.3%
records_with_any_field	10,819,870	26.8%	19,734,757	49.0%