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: 2021-11-09

Total records in dataset: 37,532,072

Quick Summary

	all_fields	quasi_fields
total_fields	19	8
total_records	37,532,072	37,532,072
total_cells	713,109,368	300,256,576
missing_fields	154,821,314	23,000,058
missing_pct	22%	8%
complete_fields	558,288,054	277,256,518
complete_pct	78%	92%
unknown_fields	32,759,802	17,463,303
unknown_pct	5%	6%
suppressed_fields	22,963,332	20,401,128
suppressed_pct	3%	7%
available_fields	502,564,920	239,392,087
available_pct	70%	80%

Field-level Utility Summary

	suppressed	suppressed_percent	missing	missing_percent
case_month	12	0.0%	0	0.0%
res_state	920	0.0%	0	0.0%
res_county	2,561,284	6.8%	0	0.0%
age_group	406,634	1.1%	354,913	0.9%
sex	1,244,276	3.3%	45,821	0.1%
race	6,597,416	17.6%	2,751,821	7.3%
ethnicity	7,837,672	20.9%	2,185,088	5.8%
death_yn	1,752,914	4.7%	17,662,415	47.1%
records_with_any_field	10,910,088	29.1%	18,725,950	49.9%