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: 2022-04-04.parquet

Total records in dataset: 69,664,983

Quick Summary

	all_fields	quasi_fields
total_fields	19	8
total_records	69,664,983	69,664,983
total_cells	1,323,634,677	557,319,864
missing_fields	301,168,362	46,493,269
missing_pct	23%	8%
complete_fields	1,022,466,315	510,826,595
complete_pct	77%	92%
unknown_fields	65,761,438	34,006,360
unknown_pct	5%	6%
suppressed_fields	36,379,553	31,395,263
suppressed_pct	3%	6%
available_fields	920,325,324	445,424,972
available_pct	70%	80%

Field-level Utility Summary

	suppressed	suppressed_percent	missing	missing_percent
case_month	4	0.0%	0	0.0%
res_state	997	0.0%	0	0.0%
res_county	4,983,293	7.2%	0	0.0%
age_group	556,257	0.8%	743,357	1.1%
sex	1,685,060	2.4%	149,440	0.2%
race	10,255,624	14.7%	5,615,647	8.1%
ethnicity	11,530,512	16.6%	4,169,379	6.0%
death_yn	2,383,516	3.4%	35,815,446	51.4%
records_with_any_field	17,313,328	24.9%	37,713,595	54.1%