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: 2022-01-24.parquet

Total records in dataset: 51,447,566

Quick Summary

	all_fields	**quasi_fields**
total_fields	19	8
total_records	51,447,566	51,447,566
total_cells	977,503,754	411,580,528
missing_fields	214,954,576	32,405,057
missing_pct	22%	8%
complete_fields	762,549,178	379,175,471
complete_pct	78%	92%
unknown_fields	48,010,205	24,181,315
unknown_pct	5%	6%
suppressed_fields	28,053,842	24,584,818
suppressed_pct	3%	6%
available_fields	686,485,131	330,409,338
available_pct	70%	80%

Field-level Utility Summary

	suppressed	**suppressed_percent**	**missing**	**missing_percent**
case_month	4	0.0%	0	0.0%
res_state	990	0.0%	0	0.0%
res_county	3,468,034	6.7%	0	0.0%
age_group	493,743	1.0%	488,236	0.9%
sex	1,415,143	2.8%	64,990	0.1%
race	7,956,261	15.5%	3,711,361	7.2%
ethnicity	9,167,465	17.8%	3,056,830	5.9%
death_yn	2,083,178	4.0%	25,083,640	48.8%
records_with_any_field	13,324,018	25.9%	26,433,889	51.4%