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: 2022-02-18.parquet

Total records in dataset: 61,267,087

Quick Summary

	all_fields	**quasi_fields**
total_fields	19	8
total_records	61,267,087	61,267,087
total_cells	1,164,074,653	490,136,696
missing_fields	261,264,394	40,330,501
missing_pct	22%	8%
complete_fields	902,810,259	449,806,195
complete_pct	78%	92%
unknown_fields	59,024,439	29,404,375
unknown_pct	5%	6%
suppressed_fields	32,541,626	28,419,599
suppressed_pct	3%	6%
available_fields	811,244,194	391,982,221
available_pct	70%	80%

Field-level Utility Summary

	suppressed	**suppressed_percent**	**missing**	**missing_percent**
case_month	4	0.0%	0	0.0%
res_state	1,002	0.0%	0	0.0%
res_county	4,121,025	6.7%	0	0.0%
age_group	517,781	0.8%	702,958	1.1%
sex	1,549,248	2.5%	100,052	0.2%
race	9,439,099	15.4%	4,820,535	7.9%
ethnicity	10,624,299	17.3%	3,739,507	6.1%
death_yn	2,167,141	3.5%	30,967,449	50.5%
records_with_any_field	15,448,947	25.2%	32,608,784	53.2%