

# QUERY COMPARISON REPORT

19dec2023\_Mini\_Query\_\_EX\_Recent |

Generated: Wednesday, February 28, 2024, at 12:30

## QUICK LINKS - REPORT SECTIONS

- [1. Report objectives](#)
- [2. General](#)
- [3. Censoring](#) | Percentage of censored values per variable.
- [4. Univariate comparison](#) | Comparing each original variable with its synthetic counterpart.

## 1. REPORT OBJECTIVES

This report compares MDClone's synthetic data with the original data.

## 2. GENERAL

The synthetic data file includes 208 rows, representing a cohort of that size. The corresponding original data file features a reasonably similar number of rows. Both original and synthetic files include 103 columns, which represent three types of variables:

- 46 categorical variables
- 55 numerical variables
- 2 date variables

## 3. CENSORING

Some values in the categorical variables are censored to ensure that patient privacy is maintained and no original information is disclosed about an individual or a small group of individuals. The following table presents the categorical variables with the highest ratio of censored values:

Categorical variable name	% Censored	% Full
blood pressure_measurement...	98.0%	98.06%
smok_health factors_smok...	93.0%	92.64%
ldl_copy-lab test	93.0%	92.64%
hba1c-lab test	91.0%	92.64%

Categorical variable name	% Censored	% Full
smok_health factors_smok...	91.0%	91.09%
hdl-lab test	90.0%	91.09%
triglycerides-lab test	90.0%	91.09%
hdl_copy-lab test	90.0%	91.09%
creatinine_copy-lab test	90.0%	91.09%
triglycerides_copy-lab test	90.0%	91.09%

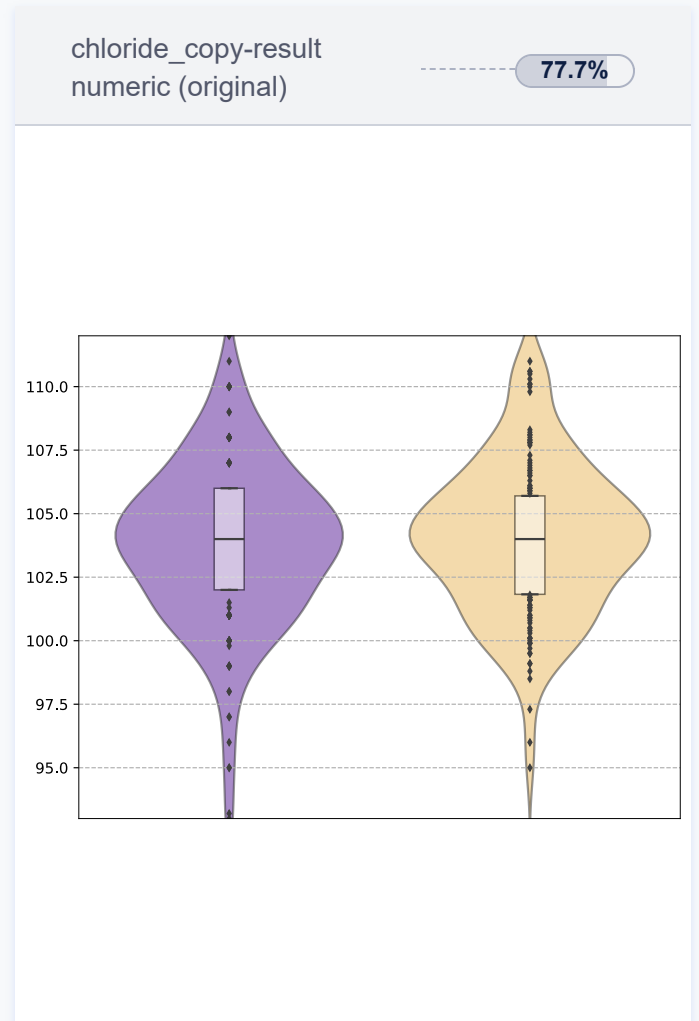
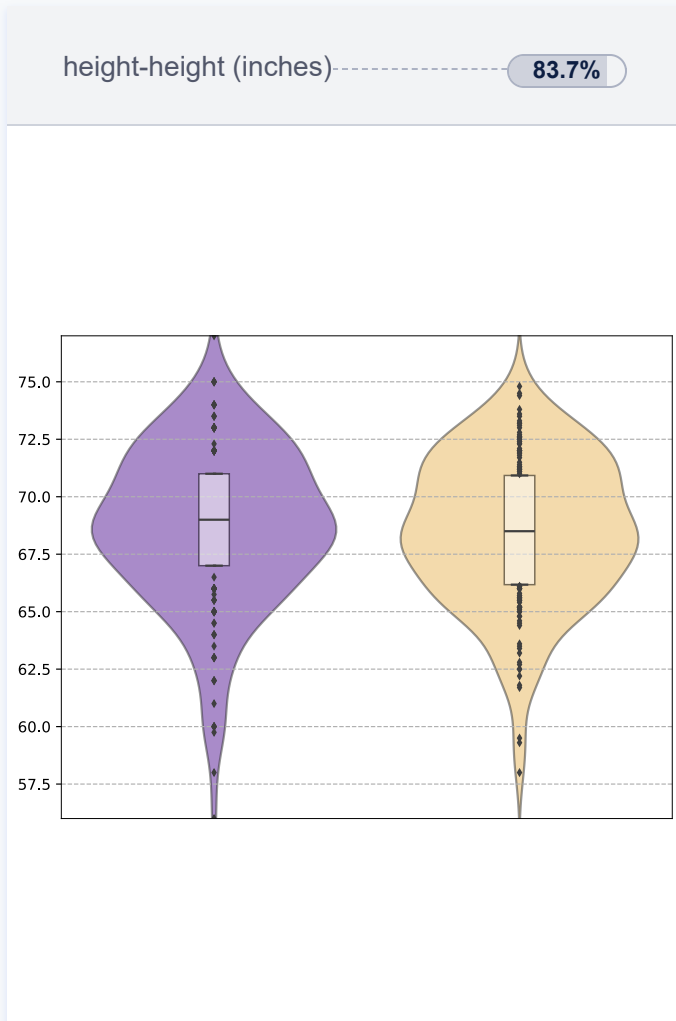
## 4. UNIVARIATE COMPARISON

### 4.2 VIOLIN PLOTS

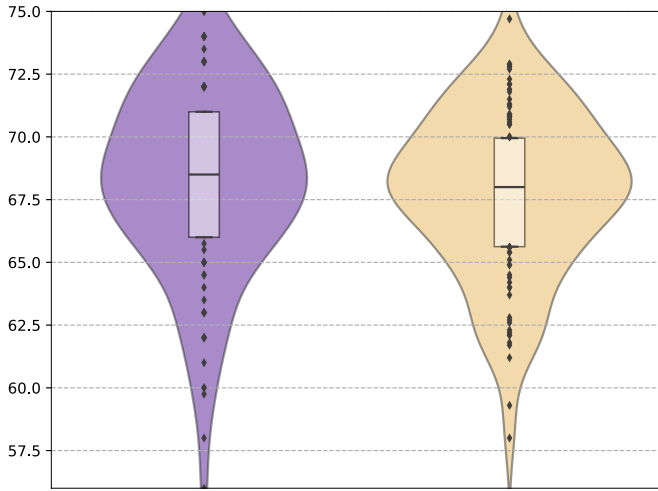
■ synthetic data

■ original data

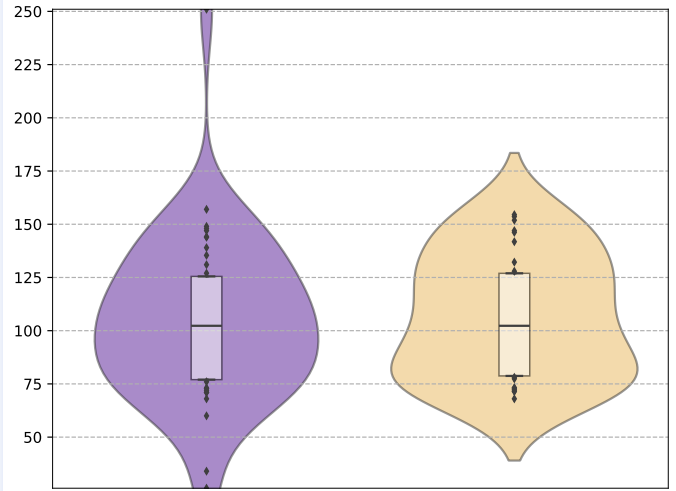
20% % full = % data that exists for this variable



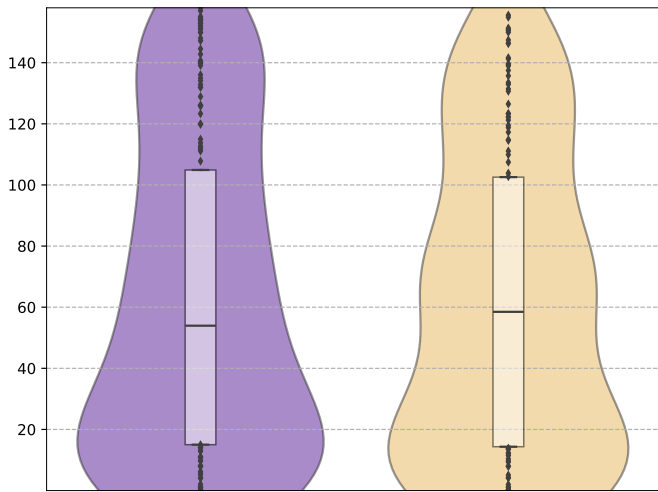
bmi >= 30-height (inches)----- 50.0%



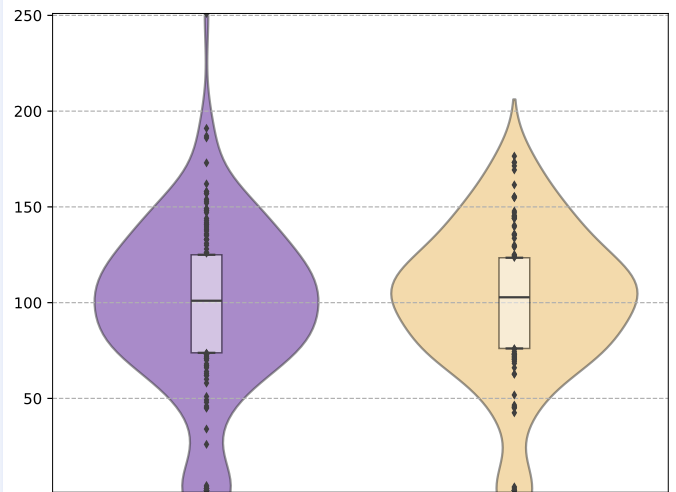
ldl-result numeric (original)----- 15.9%



resulting weight-measurement date-time-days from reference----- 68.7%

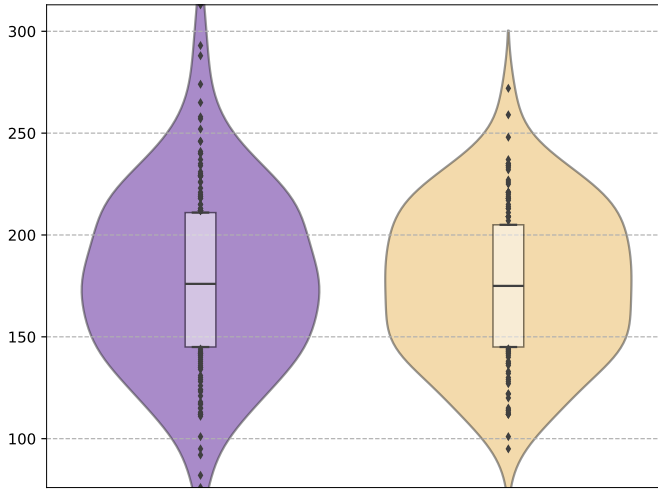


ldl\_copy-result numeric (original)----- 61.2%



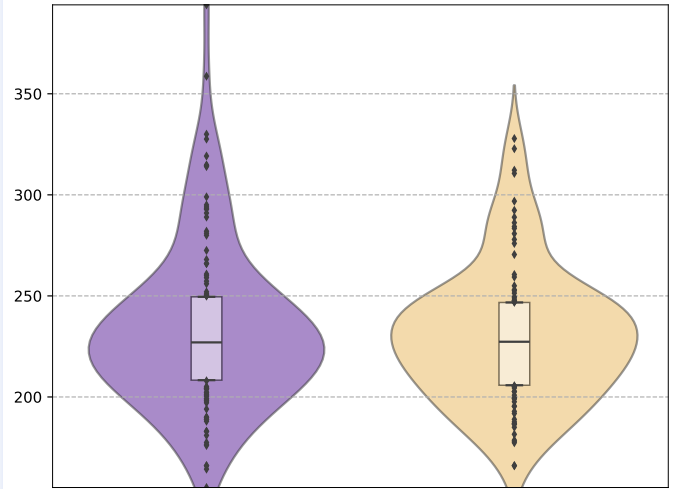
total cholesterol-result  
numeric (original)

64.4%



bmi >= 30-weight (pounds)

50.0%

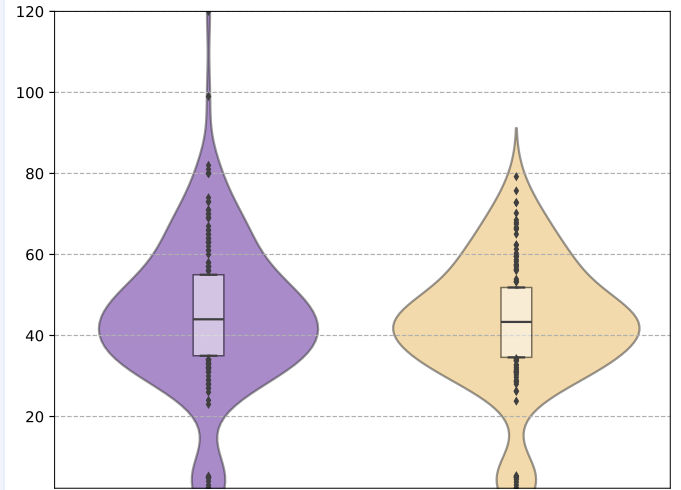
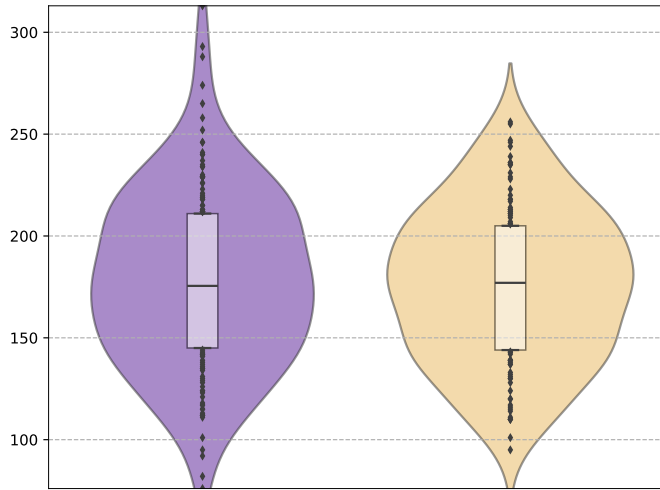


total cholesterol\_copy-result  
numeric (original)

62.4%

hdl\_copy-result numeric  
(original)

63.9%



Note! variables with low %full can hinder the synthetic algorithm and impact statistic distribution similarity