**SUMMARY FOR DATA**

This is a dataset about 500 patients of which 350 patients participated in a clinical trial. None of the patients were using Novodra (a popular injectable insulin) or Auralin (the oral insulin being researched) as their primary source of insulin before. All were experiencing elevated HbA1c levels.

All 350 patients were treated with Novodra to establish a baseline HbA1c level and insulin dose. After 4 weeks, which isn’t enough time to capture all the change in HbA1c that can be attributed by the switch to Auralin or Novodra:

1. 175 patients switched to Auralin for 24 weeks
2. 175 patients continued using Novodra for 24 weeks

Data about patients feeling some adverse effects is also recorded.

**COLUMN DESCRIPTIONS**

**TABLE : PATIENTS**

**patient\_id**: the unique identifier for each patient in the Master Patient Index (i.e. patient database) of the pharmaceutical company that is producing Auralin

**assigned\_sex**: the assigned sex of each patient at birth (male or female)

**given\_name**: the given name (i.e. first name) of each patient

**surname**: the surname (i.e. last name) of each patient

**address**: the main address for each patient

**city**: the corresponding city for the main address of each patient

**state**: the corresponding state for the main address of each patient

**zip\_code**: the corresponding zip code for the main address of each patient

**country**: the corresponding country for the main address of each patient (all United states for this clinical trial)

**contact**: phone number and email information for each patient

**birthdate**: the date of birth of each patient (month/day/year). The inclusion criteria for this clinical trial is age >= 18 (there is no maximum age because diabetes is a growing problem among the elderly population)

**weight**: the weight of each patient in pounds (lbs)

**height**: the height of each patient in inches (in)

**bmi**: the Body Mass Index (BMI) of each patient. BMI is a simple calculation using a person's height and weight. The formula is BMI = kg/m2 where kg is a person's weight in kilograms and m2 is their height in metres squared. A BMI of 25.0 or more is overweight, while the healthy range is 18.5 to 24.9. The inclusion criteria for this clinical trial is 16 >= BMI >= 38.

**TABLE : TREATMENTS AND TREATMENT\_CUT**

**given\_name**: the given name of each patient in the Master Patient Index that took part in the clinical trial

**surname**: the surname of each patient in the Master Patient Index that took part in the clinical trial

**auralin**: the baseline median daily dose of insulin from the week prior to switching to Auralin (the number before the dash) and the ending median daily dose of insulin at the end of the 24 weeks of treatment measured over the 24th week of treatment (the number after the dash). Both are measured in units (shortform 'u'), which is the international unit of measurement and the standard measurement for insulin.

**novodra**: same as above, except for patients that continued treatment with Novodra

**hba1c\_start**: the patient's HbA1c level at the beginning of the first week of treatment. HbA1c stands for Hemoglobin A1c. The HbA1c test measures what the average blood sugar has been over the past three months. It is thus a powerful way to get an overall sense of how well diabetes has been controlled. Everyone with diabetes should have this test 2 to 4 times per year. Measured in %.

**hba1c\_end**: the patient's HbA1c level at the end of the last week of treatment

hba1c\_change: the change in the patient's HbA1c level from the start of treatment to the end, i.e., hba1c\_start - hba1c\_end. For Auralin to be deemed effective, it must be "noninferior" to Novodra, the current standard for insulin. This "noninferiority" is statistically defined as the upper bound of the 95% confidence interval being less than 0.4% for the difference between the mean HbA1c changes for Novodra and Auralin (i.e. Novodra minus Auralin).

**TABLE : ADVERSE\_REACTIONS**

**given\_name**: the given name of each patient in the Master Patient Index that took part in the clinical trial and had an adverse reaction (includes both patients treated Auralin and Novodra)

**surname**: the surname of each patient in the Master Patient Index that took part in the clinical trial and had an adverse reaction (includes both patients treated Auralin and Novodra)

**adverse\_reaction**: the adverse reaction reported by the patient

**ADDITIONAL INFORMATION**

1.Insulin resistance varies person to person, which is why both starting median daily dose and ending median daily dose are required, i.e., to calculate change in dose.

2.It is important to test drugs and medical products in the people they are meant to help. People of different age, race, sex, and ethnic group must be included in clinical trials. This diversity is reflected in the patients table.

**ISSUES WITH THE DATASET**

**1. DIRTY DATA**

**TABLE : PATIENTS**

patient\_id = 9 has misspelled name Dsvid instead of David accuracy issue

state column sometimes contain full name and some times abbreviation consistency issue

zip code column has entries with 4 digit validity issue

data missing for 12 patients in address, city, state, zip\_code , country, contact completion issue

incorrect data type assigned to sex, zip code, birthdate validity issue

duplicate entries by the name of John Doe accuracy issue

one patient has weight = 48 pounds accuracy issue

one patient has height = 27 inches accuracy issue

**TABLE : TREATMENTS AND TREATMENTS\_CUT**

given\_name and surname col is all lower case consistency issue

remove u from Auralin and Novadra cols validity issue

in novadra and Auralin col treated as nan validity issue

missing values in hba1c\_change col completion issue

1 duplicate entry by the name Joseph day accuracy issue

in hba1c\_change 9 instead of 4 accuracy issue

**TABLE : ADVERSE\_REACTIONS**

given\_name and surname are all in lower case `consistency issue

**2. MESSY DATA**

**TABLE : PATIENTS**

contact col contains both phone and email

**TABLE : TREATMENTS AND TREATMENTS\_CUT**

Auralin and Novadra col should be split into 2 cols start and end dose

merge both the tables

**TABLE : ADVERSE\_REACTIONS**

This table should not exist independently