Capstone - milestone report
Data story

Introduction

To get any medicine to market, a pharmaceutical company must first test that medicine in the lab and then in humans, through clinical trials. Firstly to make sure it is safe and secondly to make sure it is efficacious.

GlaxoSmithKline (GSK) is one of the biggest pharmaceutical companies in the world and a leader in respiratory and infectious disease medications. They have carried out years of clinical trials and with clinical trials comes data collection.

The largest trials can include over 1000 patients who have many different measurements taken during the trial. The investigators are looking for a particular effect, for example a reduction in exacerbations of asthma attacks in a respiratory trial. The trial is split into different arms (typically 2-4) with each arm having a different dosing level (from placebo to a high concentration of the medicine being tested). But what if the medicine was affecting something else as well? What if there were another condition that the medicine was reducing or curing but that wasn't seen because it wasn't looked for?

This paper investigates the glucose measurements taken from patients in several GSK late stage or Phase 3 clinical trials. Glucose is of potential interest to GSK because if any trend of glucose reduction is seen it could indicate that the medicine may have potential as a cure for diabetes. It's expected that glucose will remain consistent for all patients on a trial regardless of whether they are given a placebo or the medicine. But there may be hidden trends in the glucose data that were not spotted because that's not what the investigators were looking for.

The most famous medicine that was discovered in this way is the blockbuster Viagra - it was on trial as a heart medication and what it's used for now was merely a side effect that was reported by the participants.

The dataset - creating a manageable dataframe for analysis

The GSK RDIP (research and development information platform)

In the last year, the Data Science Centre of Excellence at GSK has built a repository of many different data that exists across the company. This has been stored on a cloud platform called RDIP.

There are several databases housed on RDIP but the ones of interest for this project are the clinical trial test results (lb) and the demographic (dm) database. The former is the lab results collected from all trials conducted over the past 10 years and the latter is demographic data on each patient who has taken part in a GSK trial.

database	Table name
clntrl_e	t_anon_int_sdtm_dm_21jul2017
clntrl_e	t_anon_int_sdtm_ds_25sep2017
clntrl_e	t_anon_int_sdtm_dv_25sepl2017
clntrl_e	t_anon_int_sdtm_lb_25sep2017

The top of the list of the databases on the RDIP platform at GSK

Examining the data

As the data sits on RDIP, it's necessary to pull from the tables into R to conduct analysis. But it's important to make sure only the relevant data is pulled as the tables are huge and not all the data within them will be needed.

Size: A few SQL queries were run to get a feel for the size of these tables. There are 86,925 rows in the dm table, one for each patient. The lb table is huge, containing nearly 2.5million rows. Filtering on glucose shows just 92,585 rows which is easier to deal with so only glucose rows were pulled into R to create a dataframe to work with. A quick examination shows that glucose tests were conducted on both blood and urine samples. Urine should never contain glucose and it's only glucose tests on serum that are of interest so the dataset was refined further to filter out these results, leaving 71, 943 rows.

Variables of interest

A quick look at the lb table shows there are 97 variables. Of these, only 16 are actually required:

Studyid (the unique study identifier), usubjid (subject ID, unique patient identifier), lbdy (lab day in the trial that the test was conducted), lbdtc (lab day time code), visit (identifies which number visit this was), lbtest (what is being tested, in this case all Glucose), lbspec (how was the test done - in this case all Serum), lborres (result of the test), lborresu (units of the test), lbstresn (lab result in standard units), lbstresu (standard units for lab result), lbstnrlo (lower value of aveage glucose tests), lbstnrhi (higher value of glucose test), so_therapy_area (therapry area of the trial), so_indications (disease medicine is targeting), so_study_phase (stage of the study from Phase 1 to phase 3), so_abbreviated_title (study title).

The lbdy measures the day within the trial that the glucose test was conducted and the first day of dosing is lbdy = 1 (there is no day zero). Patients will routinely attend the clinic and have a glucose test before the trial begins (negative lbdy) and after they have finished taking the medicine (large lbdy). Therefore, to get a really accurate picture of whether the medicine is affecting glucose, it will be necessary to identify the date that the patient started the dosing and ended the dosing and pull the two glucose results for those days from the data set.

This is the information in the dm data table which is why it's needed - it contains the time code for the patient receiving their first dose of the medicine (rfxstdtc) and the last (rfxendtc) so can identify which lbdy was their first and last day.

The lbstrens is used for the glucose results as these are all in the same units.

A quick look at both lbdy and lbstrens shows that there are outliers and strangely large or small results that will need to be filtered out.

Summary of lbdy

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3.403e+38	-6.000e+00	5.800e+01	-1.605e+37	1.650e+02	4.480e+02

Summary of lbstresn

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3.403e+38	5.000e+00	5.000e+00	-7.095e+34	6.000e+00	2.160e+02

Initial data wrangling - identifying three trials to look at in detail

GSK is most interested in analysing Phase 3 trials - these later stage trials have a lot more patients so there's more chance of identifying any trend.

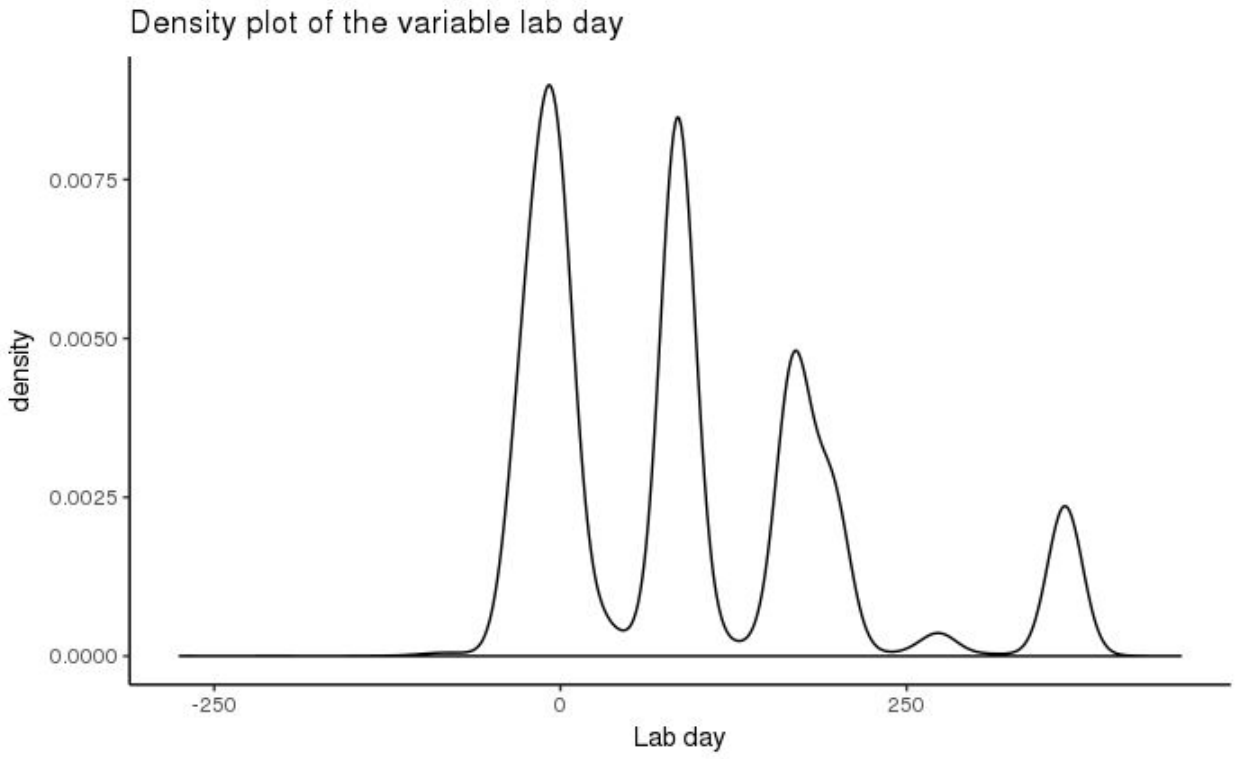
Therefore the first step was to take a look at the types of trials in the dataset:

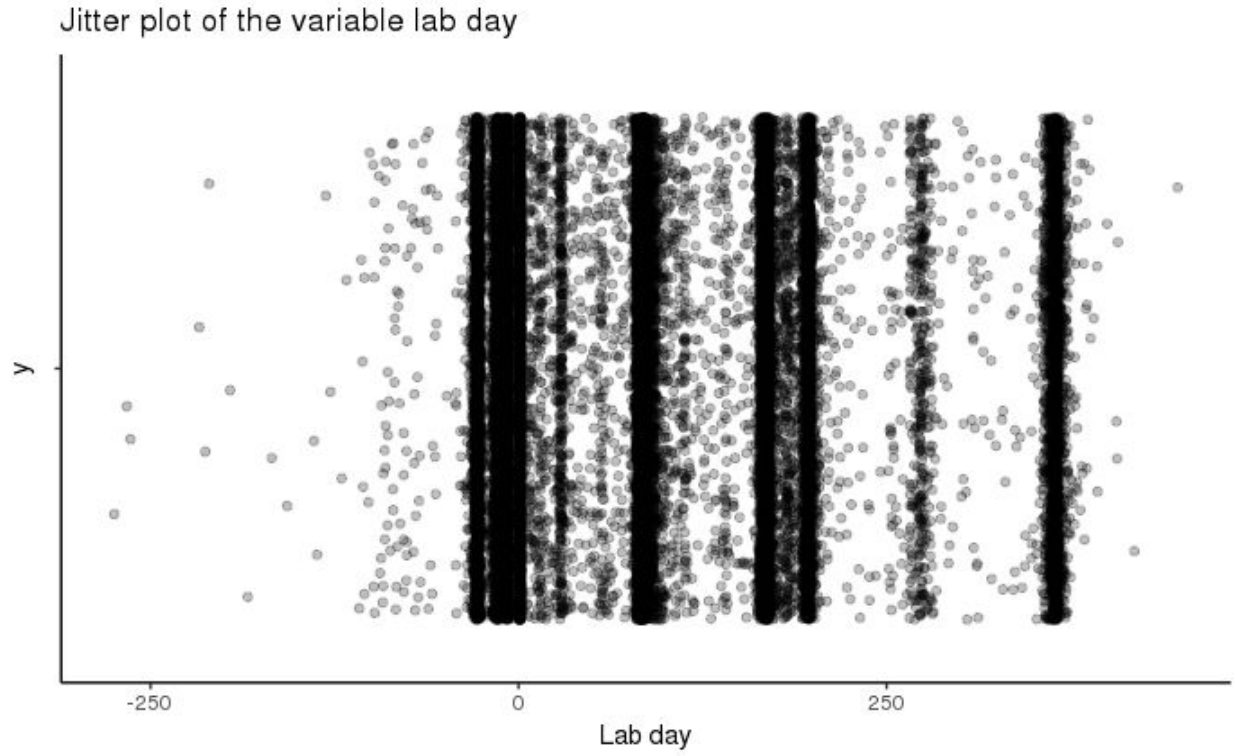
	so_therapy_area	so_indications	so_study_phase	n
1	Respiratory	Asthma	PHASE I	175
2	Respiratory	Asthma	PHASE IIA	157
3	Respiratory	Asthma	PHASE IIB	690
4	Respiratory	Asthma	PHASE IIIA	1663
5	Respiratory	Asthma	PHASE IIIB	792
6	Respiratory	Cystic Fibrosis	PHASE IIA	146
7	Respiratory	Dermatitis, Atopic	PHASE I	20
8	Respiratory	Dermatitis, Atopic	PHASE IIA	25
9	Respiratory	Lung Injury, Acute	PHASE I	3
10	Respiratory	Pulmonary Disease, Chronic Obstructive	PHASE I	416
11	Respiratory	Pulmonary Disease, Chronic Obstructive	PHASE IIA	265
12	Respiratory	Pulmonary Disease, Chronic Obstructive	PHASE IIIA	13029
13	Respiratory	Pulmonary Disease, Chronic Obstructive	PHASE IIIB	808

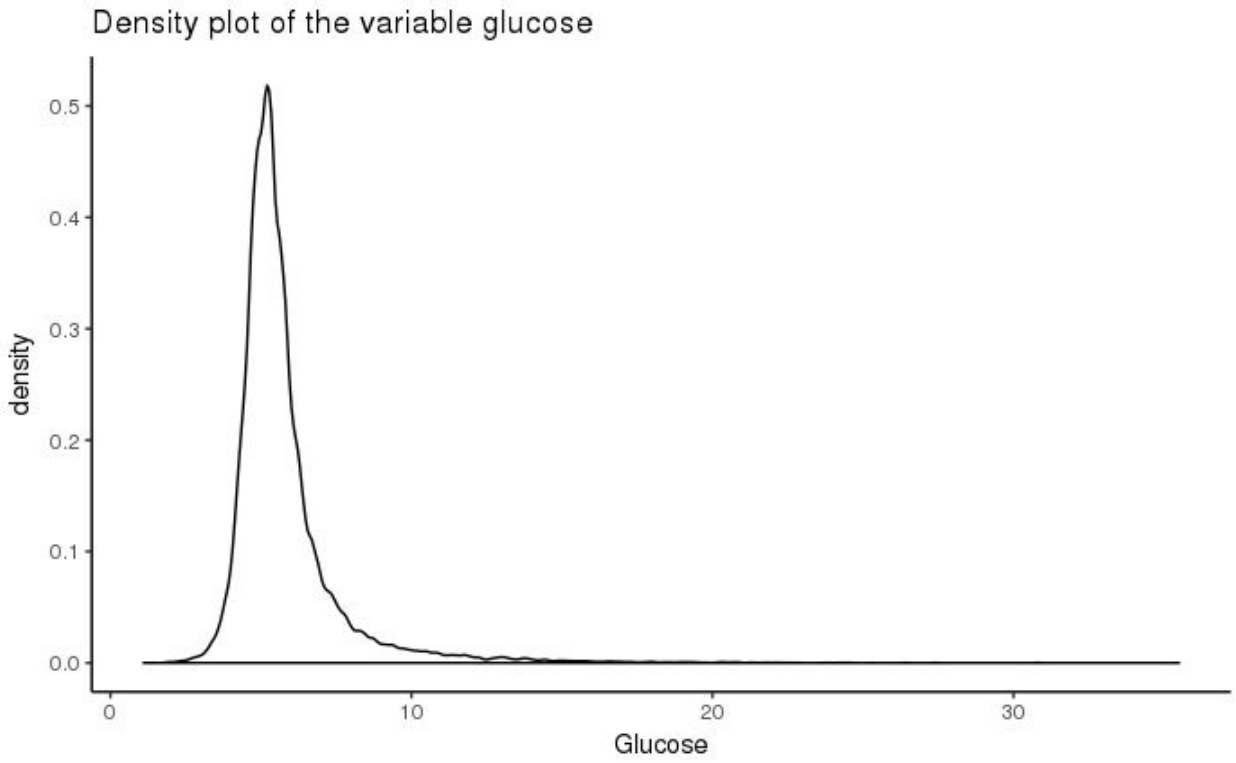
Next, all but the phase 3 trials (PHASE IIIA and PHASE IIIB) were filtered out. GSK is also not interested in analysis on mixed dosing trials in this project, where each patient is given three different medicines, just in a different order (so ABC or BCA or CAB etc). To figure out if a trial is

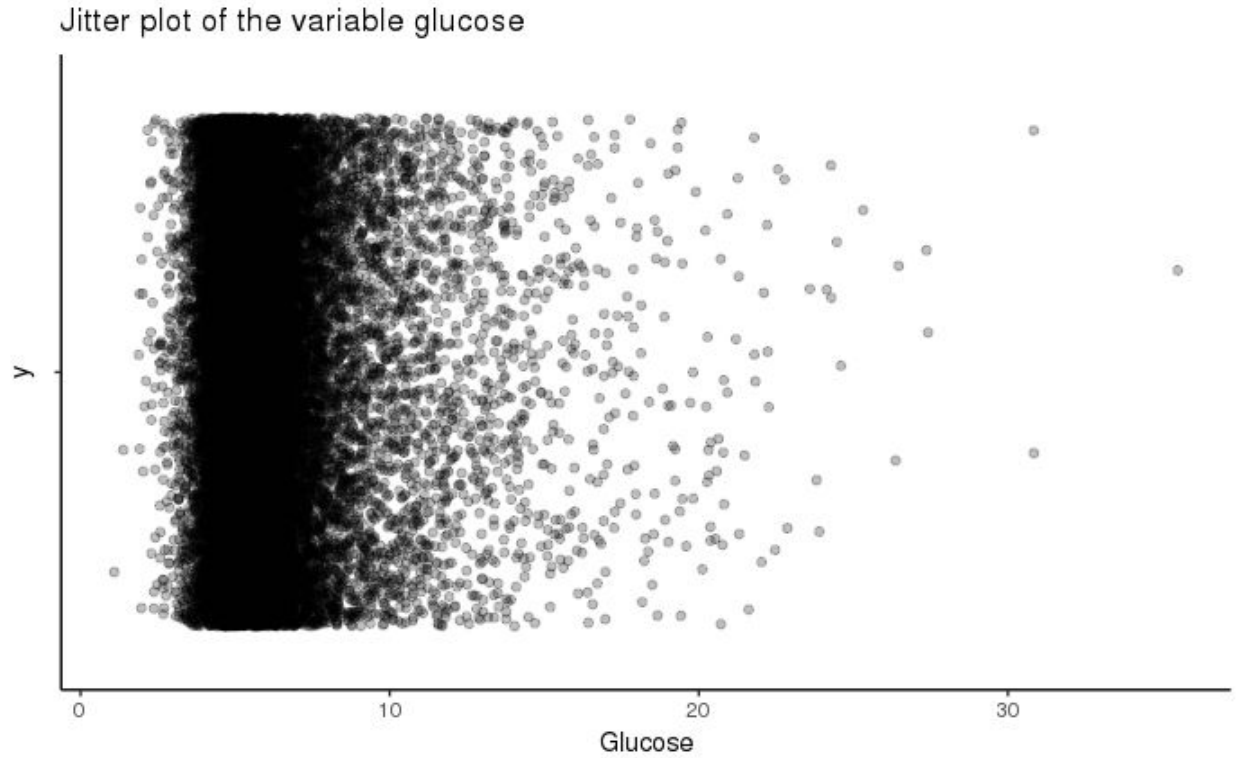
mixed dosing, the studyid had to be crossed checked against the full trial documentation on clinicaltrials.gov. Each study has a unique identifier and the website allows a user to search for the title and a few details on any study. Using the identifiers and the company name GSK, each trial was searched on the website. One of the pieces of information is whether the trial was mixed or parallel (where patients get just one medicine dosage in the trial) so the mixed ones can easily be identified and removed.

Once this is done, there are 14 trials left. The demographic data is joined on to those trials for statistical analysis later. The large and small values for lbdy and lbstren are filtered out and the data is visualised to check it looks reasonable (jitter and density plots).
For the lab days, it's expected to see several peaks as a patient will attend a trial 4-5 times. There will be a peak around zero for the first day they went and then several others that are regularly spaced out. A trial can last more than a year so it's possible to see peaks up to day 365.
For glucose, most of the data should be concentrated between zero and ten, the standard normal range. There may be some larger figures for patients who have high glucose due to other factors but not too many.


Density plot of the variable lab day
density
0.0075
0.0050
0.0025
0.0000
-250
0
250
Lab day


Jitter plot of the variable lab day
y
-250
0
250
Lab day





The next step will be to identify the biggest three Phase 3 trials and examine them in more detail using machine learning and statistics.