Big data and big pharma - mining the wealth of data held by GSK from respiratory clinical trials to uncover potential new uses for medicines

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. 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.

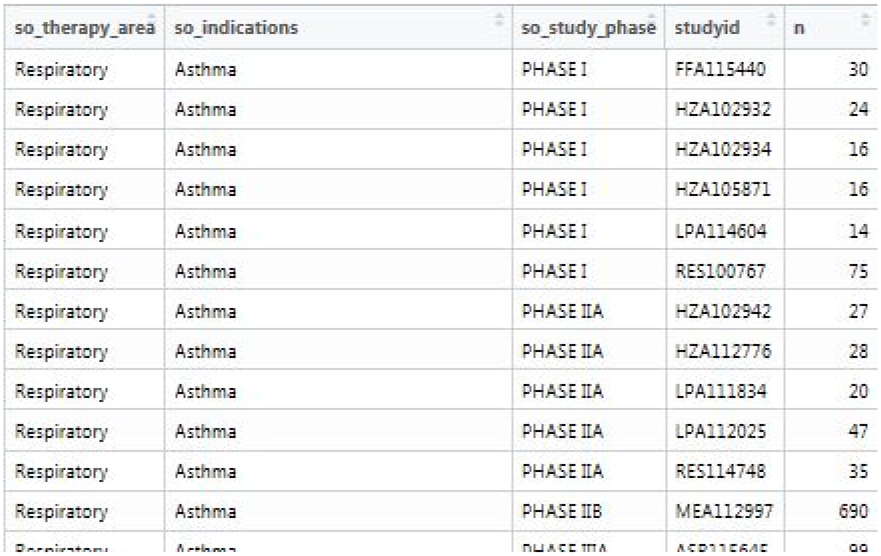
```
Summary of lbdy
Min.              1st Qu.    Median     Mean         3rd Qu.     Max.
-3.403e+38 -6.000e+00  5.800e+01 -1.605e+37  1.650e+02  4.480e+02

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.

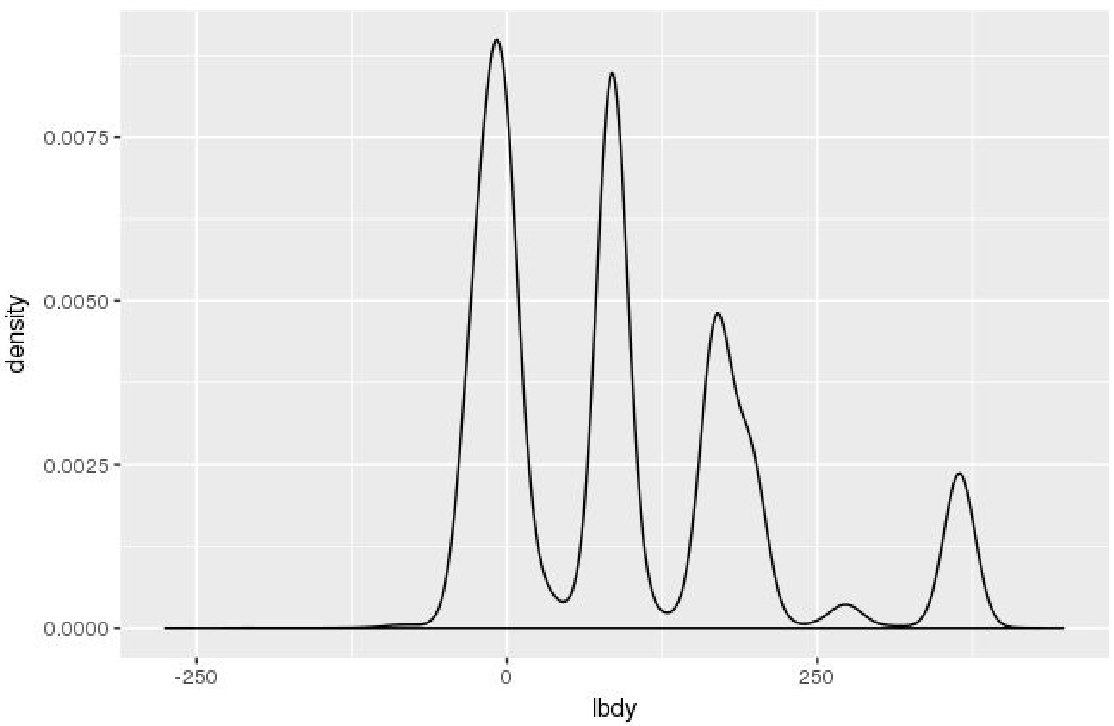
Therefore the first step was to take a look at the types of trials in the dataset:

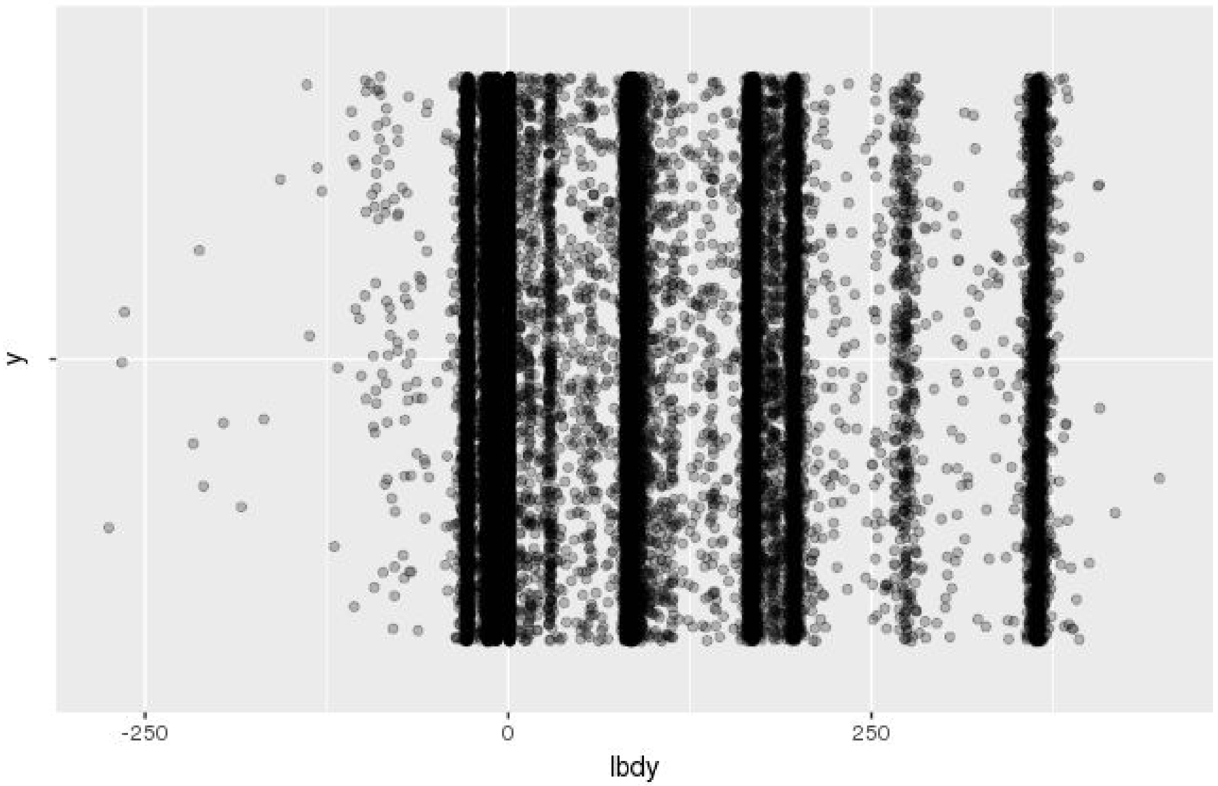


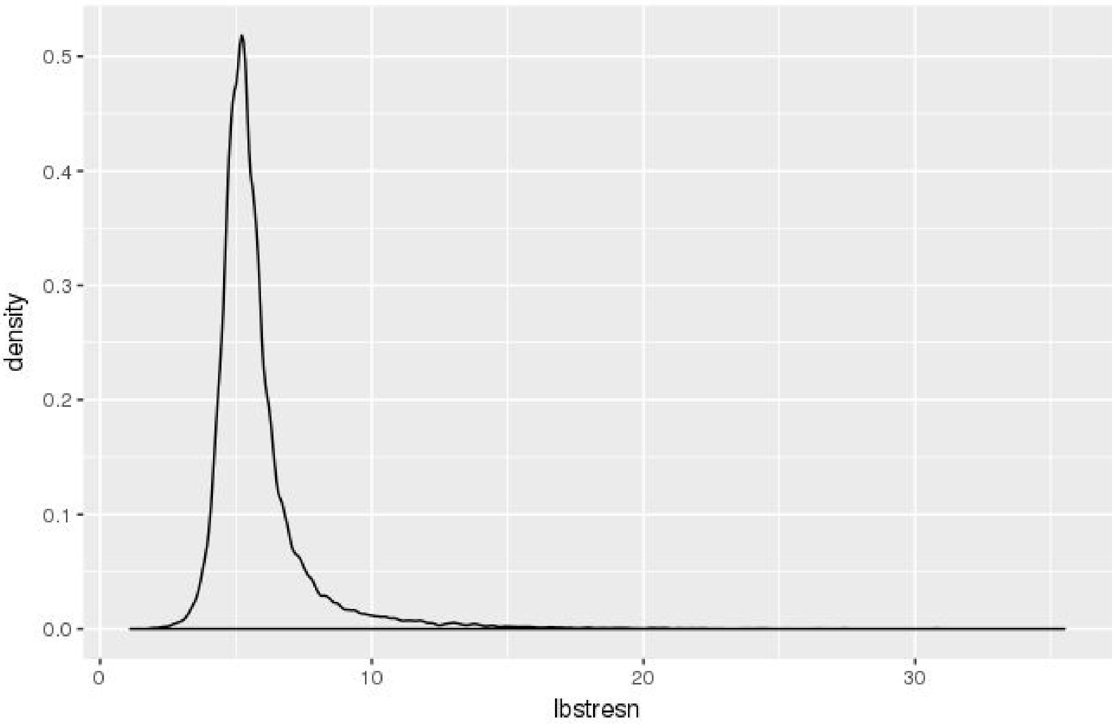
so_therapy_area	so_indications	so_study_phase	studyid	n
Respiratory	Asthma	PHASE I	FFA115440	30
Respiratory	Asthma	PHASE I	HZA102932	24
Respiratory	Asthma	PHASE I	HZA102934	16
Respiratory	Asthma	PHASE I	HZA105871	16
Respiratory	Asthma	PHASE I	LPA114604	14
Respiratory	Asthma	PHASE I	RES100767	75
Respiratory	Asthma	PHASE IIA	HZA102942	27
Respiratory	Asthma	PHASE IIA	HZA112776	28
Respiratory	Asthma	PHASE IIA	LPA111834	20
Respiratory	Asthma	PHASE IIA	LPA112025	47
Respiratory	Asthma	PHASE IIA	RES114748	35
Respiratory	Asthma	PHASE IIB	MEA112997	690
Respiratory	Asthma	PHASE IIIA	[illegible]	99

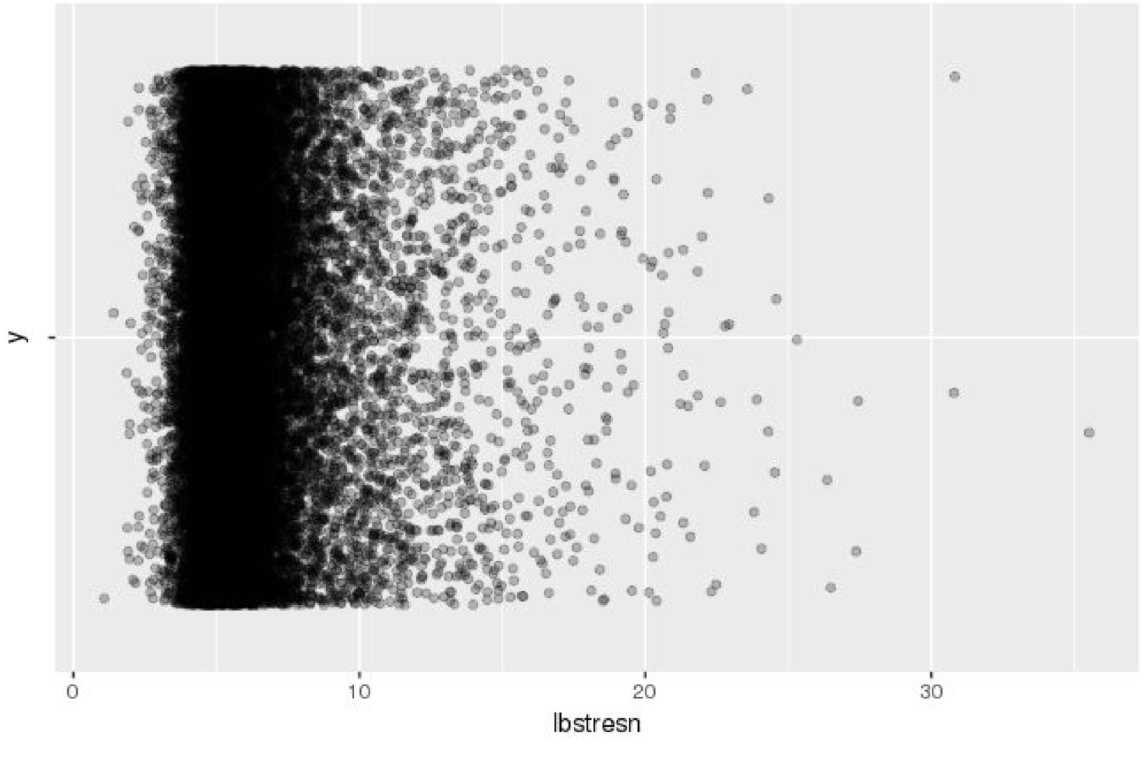
Next, all but the phase 3 trials (PHASE IIIA and PHASE IIIB) were filtered out. GSK is also not interested in analysis on mixed dosing trials in this project, where each patient is given three different medicines, just in a different order (so ABC or BCA or CAB etc). To figure out if a trial is mixed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

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):


density
0.0075
0.0050
0.0025
0.0000
-250
0
250
lbdy

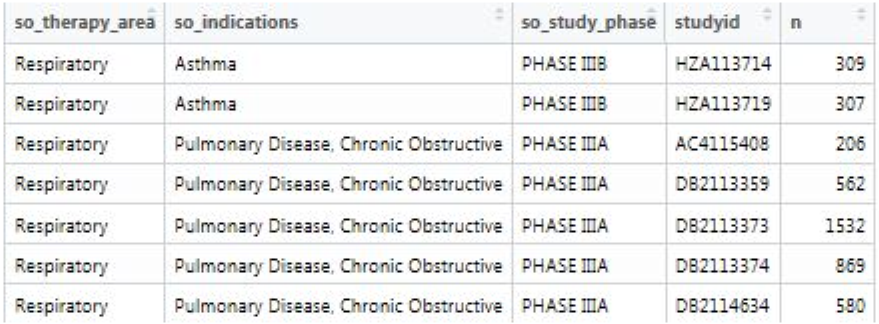

y
-250
0
250
lbdy





Analysis of glucose levels in the trials

The three trials with the most patients were selected for further examination of glucose levels.



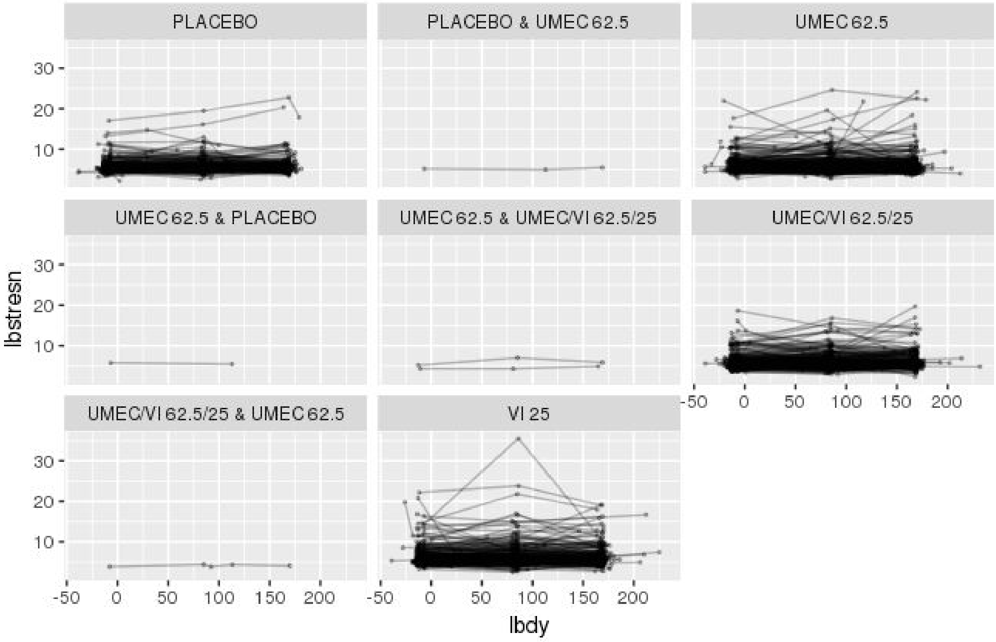
so_therapy_area	so_indications	so_study_phase	studyid	n
Respiratory	Asthma	PHASE IIIB	HZA113714	309
Respiratory	Asthma	PHASE IIIB	HZA113719	307
Respiratory	Pulmonary Disease, Chronic Obstructive	PHASE IIIA	AC4115408	206
Respiratory	Pulmonary Disease, Chronic Obstructive	PHASE IIIA	DB2113359	562
Respiratory	Pulmonary Disease, Chronic Obstructive	PHASE IIIA	DB2113373	1532
Respiratory	Pulmonary Disease, Chronic Obstructive	PHASE IIIA	DB2113374	869
Respiratory	Pulmonary Disease, Chronic Obstructive	PHASE IIIA	DB2114634	580

The three trials chosen are:

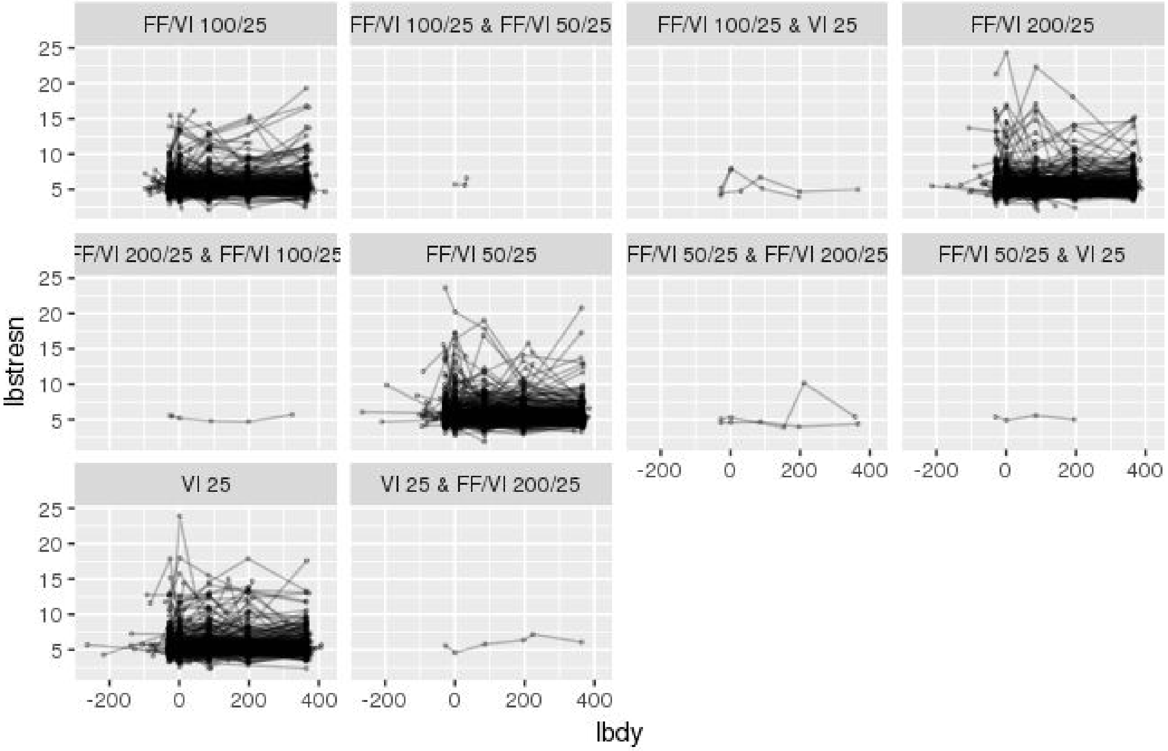
Therapy Area	Indication	Phase	Study ID	No of patients
Respiratory	Chronic Obstrucitve Pulmonary Disease	Phase IIIA	DB2113373	1532
Respiratory	Chronic Obstrucitve Pulmonary Disease	Phase IIIA	HZC102970	1633
Respiratory	Chronic Obstrucitve Pulmonary Disease	Phase IIIA	HZC102871	1622

Initial plots were made of these three trials. The variable actarm was used to split the data into separate plots. Actarm is the variable for the arm of the trial the patient was assigned to - it could be placebo or a certain level of dosing.

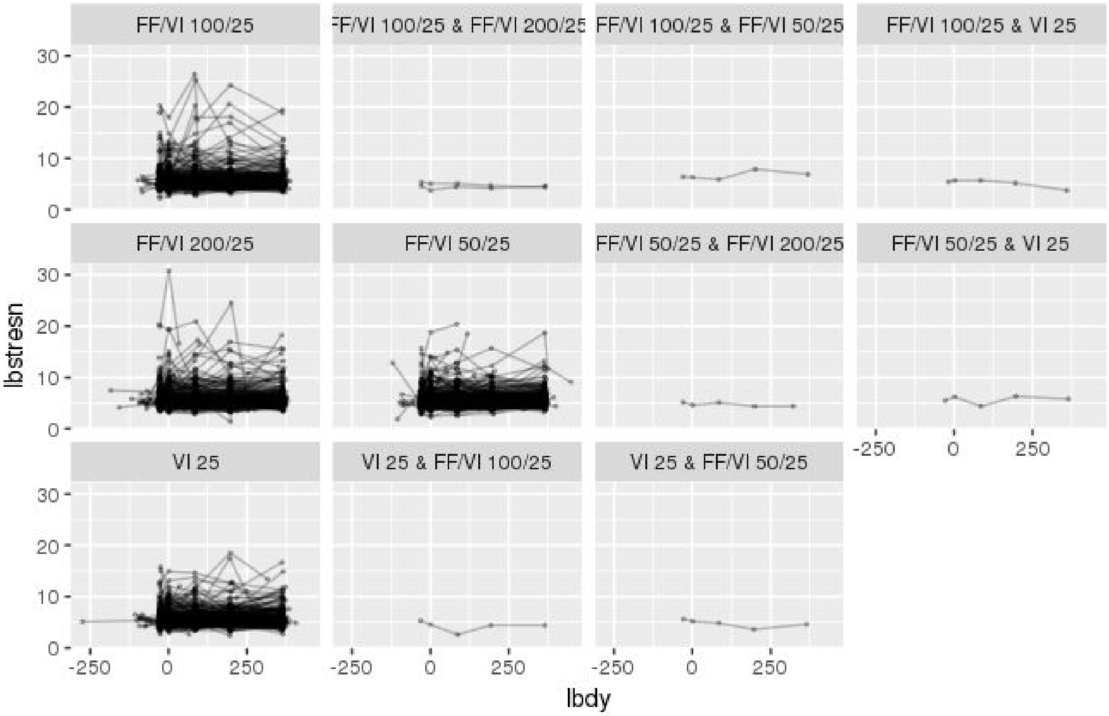
Trial 1 - DB2113373



Trial 2 - HZC102970

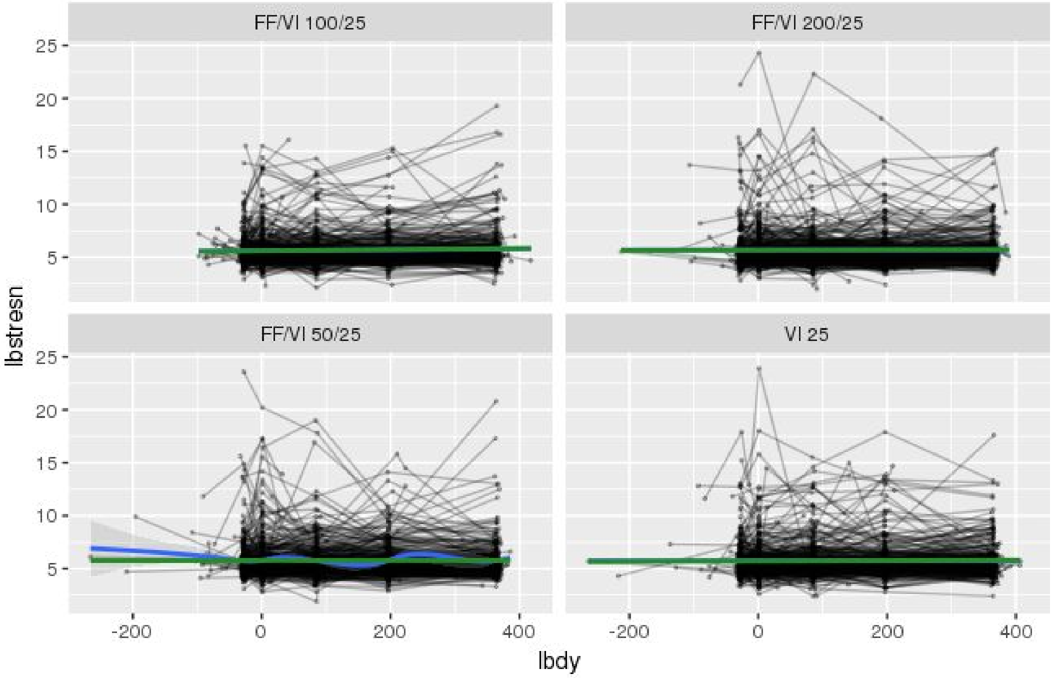


Trial 3 - HZC102871

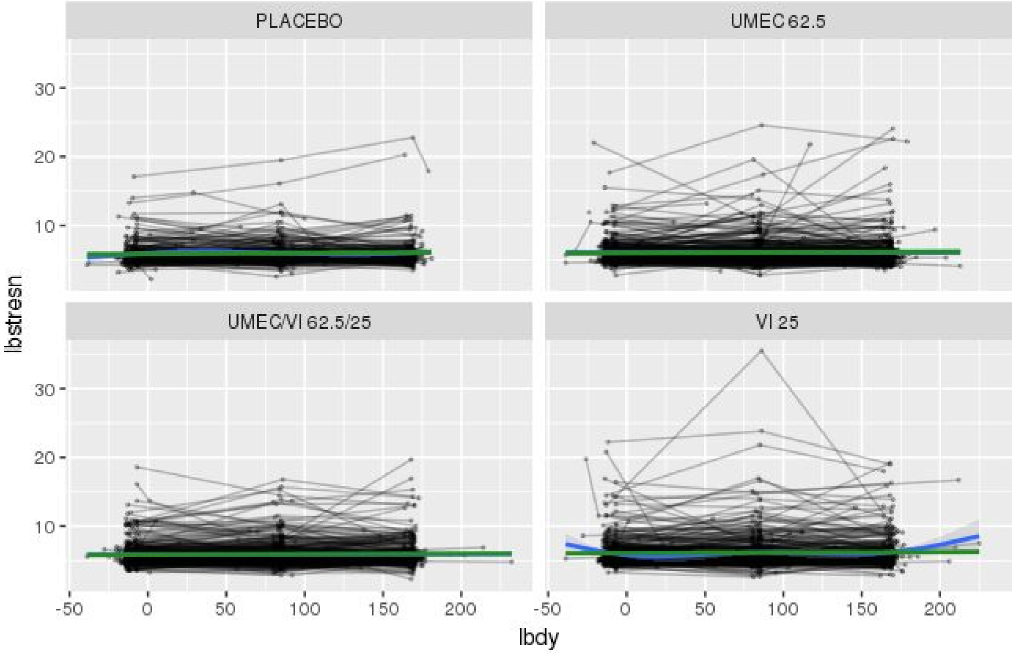


From the initial plots it’s clear there are some arms of the trials that only have one patient. These will not be useful for modelling so are removed. A crude linear model is also added to the remaining plots to see if it looks like there are any trends in the glucose levels.

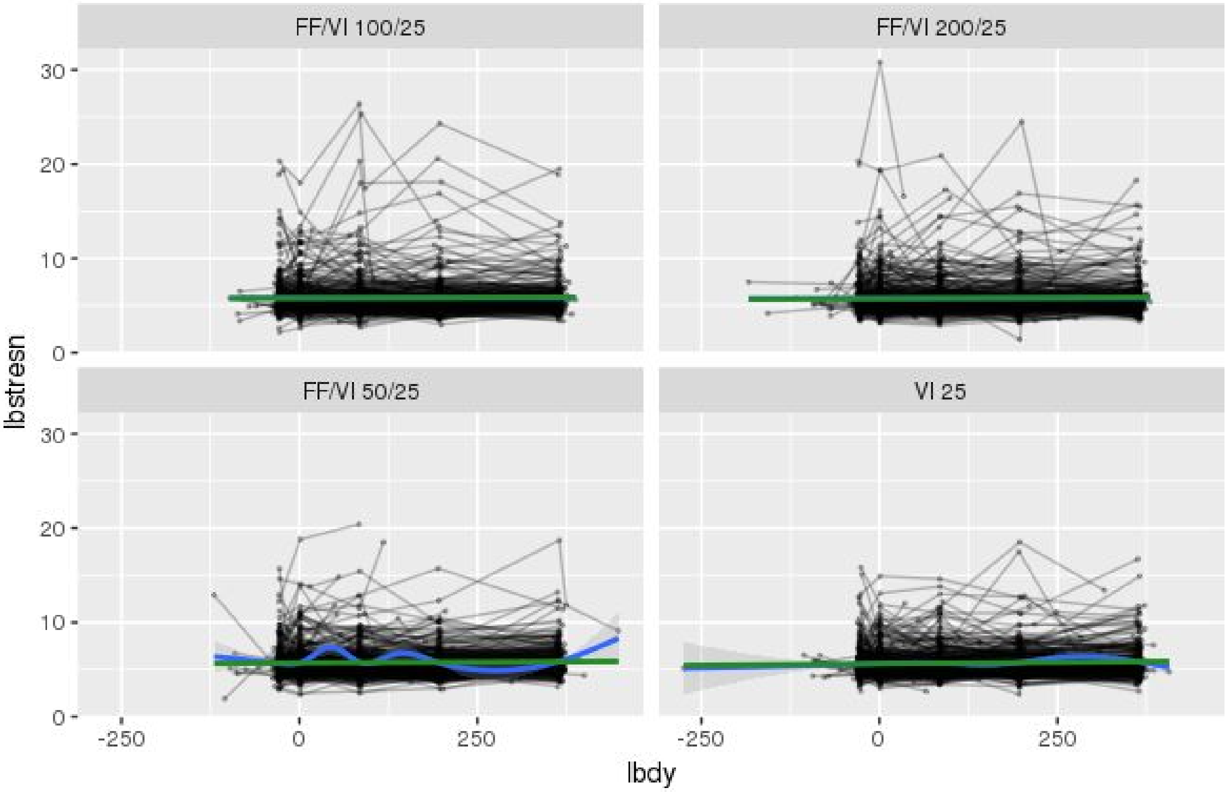
Trial 1 - DB2113373



Trial 2 - HZC102970



Trial 3 - HZC102871



These crude models seem to show there is nothing strange happening with the glucose but models are created for these trials to check.

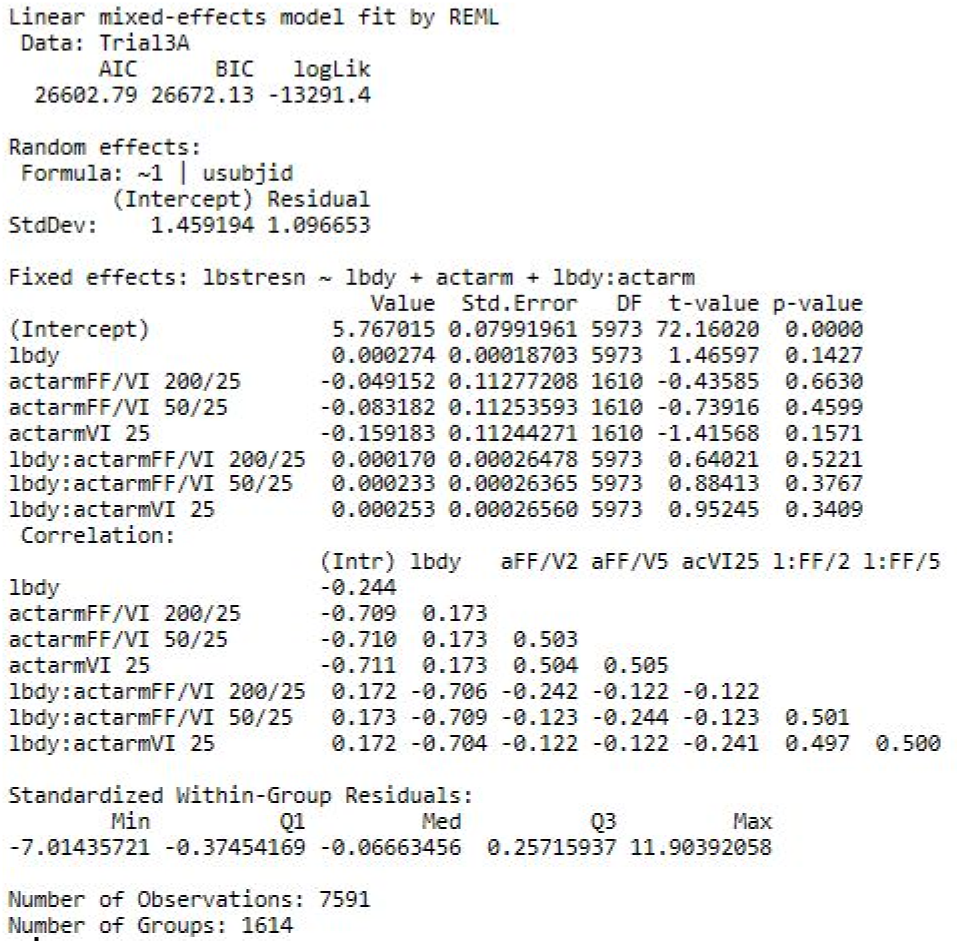
Mixed effects linear modelling

This type of modelling was recommended by the director of data science at GSK. A mixed effect linear model allows for random as well as fixed effects on the variable of interest whereas a standard linear model only allows for fixed effects.

In this data set, each arm of the trial is a subset of patients. Not every patient has been in each arm of the trial which means the way the patients have been split up is a random effect. Patients also didn't have their glucose taken on exactly the same day, also a random effect.

The model allows for these random effects and the variance that may be happening because of them (for example, there may be an unusual amount of people with naturally high glucose in one arm of the trial) to be taken into account.

The summaries of the three models are shown below:



```
Linear mixed-effects model fit by REML
 Data: Trial3A
       AIC      BIC   logLik
  26602.79 26672.13 -13291.4

Random effects:
 Formula: ~1 | usubjid
        (Intercept) Residual
StdDev:    1.459194 1.096653

Fixed effects: lbstresn ~ lbdy + actarm + lbdy:actarm
                            Value  Std.Error   DF  t-value p-value
(Intercept)              5.767015 0.07991961 5973 72.16020  0.0000
lbdy                     0.000274 0.00018703 5973  1.46597  0.1427
actarmFF/VI 200/25      -0.049152 0.11277208 1610 -0.43585  0.6630
actarmFF/VI 50/25       -0.083182 0.11253593 1610 -0.73916  0.4599
actarmVI 25             -0.159183 0.11244271 1610 -1.41568  0.1571
lbdy:actarmFF/VI 200/25  0.000170 0.00026478 5973  0.64021  0.5221
lbdy:actarmFF/VI 50/25   0.000233 0.00026365 5973  0.88413  0.3767
lbdy:actarmVI 25         0.000253 0.00026560 5973  0.95245  0.3409
 Correlation:
                        (Intr) lbdy   aFF/V2 aFF/V5 acVI25 1:FF/2 1:FF/5
lbdy                    -0.244
actarmFF/VI 200/25      -0.709  0.173
actarmFF/VI 50/25       -0.710  0.173  0.503
actarmVI 25             -0.711  0.173  0.504  0.505
lbdy:actarmFF/VI 200/25  0.172 -0.706 -0.242 -0.122 -0.122
lbdy:actarmFF/VI 50/25   0.173 -0.709 -0.123 -0.244 -0.123  0.501
lbdy:actarmVI 25         0.172 -0.704 -0.122 -0.122 -0.241  0.497  0.500

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-7.01435721 -0.37454169 -0.06663456  0.25715937 11.90392058

Number of Observations: 7591
Number of Groups: 1614
```

It’s clear from the intercepts and also the p values that there is no difference in glucose through the trial.

Detailed statistical examination of one trial

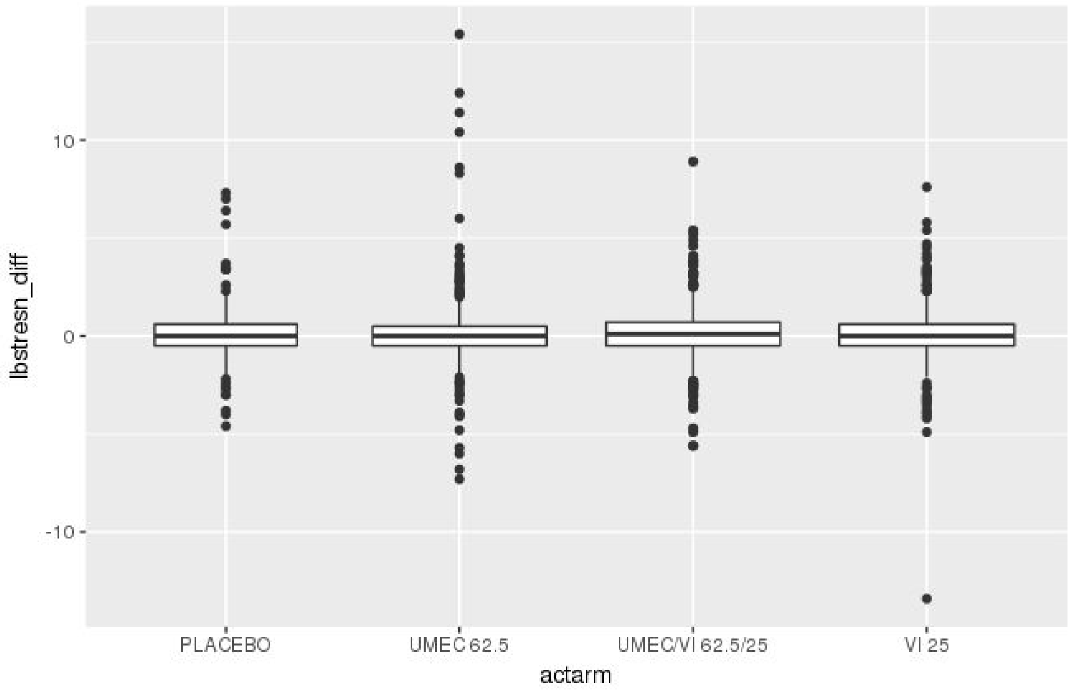
As well as using the model above, GSK asked for one trial to be examined statistically, looking at the absolute differences in glucose values for patients on the first and last days of dosing. This was why the dm data was added to the dataset. The first trial, DB2113373 was chosen.

The first dose was identified by pulling the glucose result for the lbdy that was closest to zero. The mean glucose was assigned to any patients where there were two tests done on the same lbdy.

The last dose was identified by subtracting the first dose date (rfxendtc) from the last dose date (rfxstdtc) to calculate which lbdy the last dose was given. Then the glucose result for that lbdy was pulled out.

The differences between these first and last glucose measures were taken for all patients and various statistics were calculated:

Bar plot of the differences in glucose on first and last day of dosing for all arms of the trial:



Numerical analysis of he differences in glucose on first and last day of dosing for all arms of the trial:

actarm	No of patients	Mean of differences	SD of differences	Median of differences
PLACEBO	279	0.08956833	1.347520	0.0000000
UMEC 62.5	416	0.17142856	1.947642	0.0000000
UMEC/VI 62.5/25	413	0.15496368	1.448704	0.0999999
VI 25	421	0.06380953	1.530213	0.0000000

As both the bar chart and the table show, there is no difference in glucose in any of the trial arms.

Conclusions and further work

It’s clear that there was no hidden trend in the glucose levels of the patients in these three trials.

However there are 61 other trials in the dataset that could be examined through applying the code developed for this report.

There are also other tests, not just glucose, that could be examined.

It may also be interesting to split the data using demographics - age, sex, race - and see if there are changes to glucose in those smaller subsets that are masked in these larger cuts of the data.

This work would not have been possible without the constant support of Tilo Blenk, director of data science at GSK who was always ready to go through lines of code to help refine it and to explain new concepts like SQL and mixed effects linear models.