Using TransCelerate/ Placebo & Standard of Care data, with Bayesian dynamic borrowing, to design a clinical study

Graeme Archer & Fi Guillard

Contents

RDIP Data Domain Overview

Integrating, simplifying and unlocking data to better support data reuse

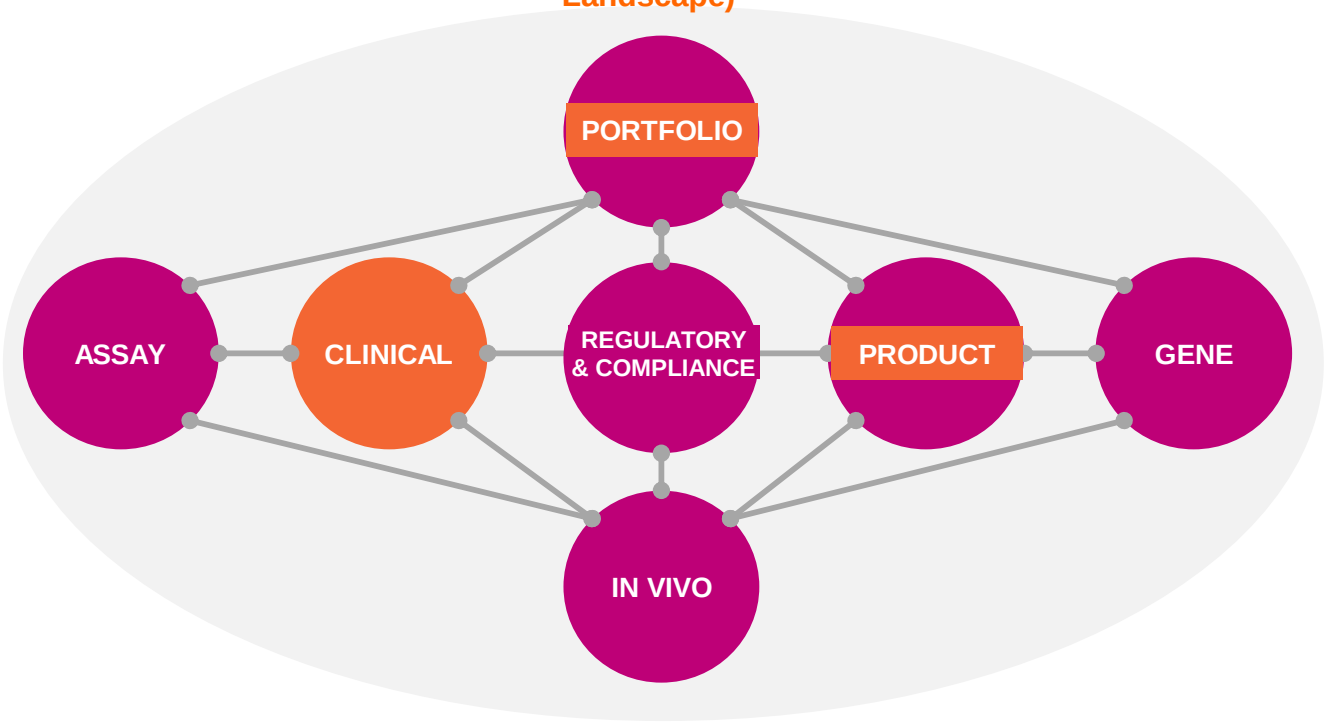
Domain Fundamentals

Why do we have data domains?
Data domains are a key enabler in the data to value continuum. They are used to organize, secure, structure and present data in a way that drives maximum value to multiple end user personas

What is a data domain?
A data domain is a construct used to support the indexing and cataloguing of information assets and intended to be relatively seamless to the end user.

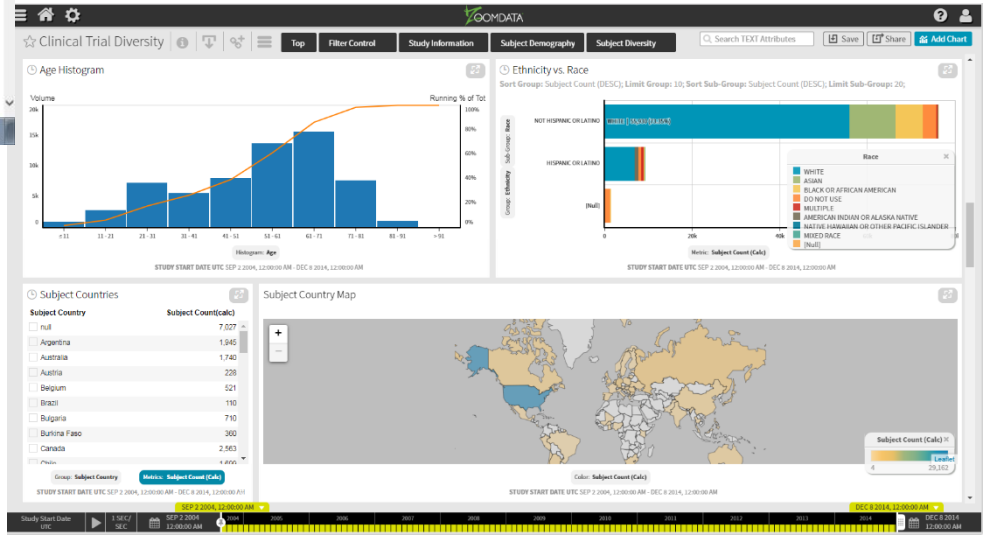
Domains are traversed using common metadata e.g. compound, disease, target, study, project….

Current domain organization (informed by use cases and R&D IT Data Landscape)



R&D Information Platform

Clinical: Use cases, tools and prototypes



The Clinical domain works…

With drug discovery and development partners in……

- PCPS/TAs
- Med Affairs / VEO
- PTS/DPUs
- TSci
- Tech
- Data CoE

on…..

- Clinical Trial Data
- Operational Data
- Claims Data
- EHR/EMR Data
- Industry Data
- Sensor/Device Data

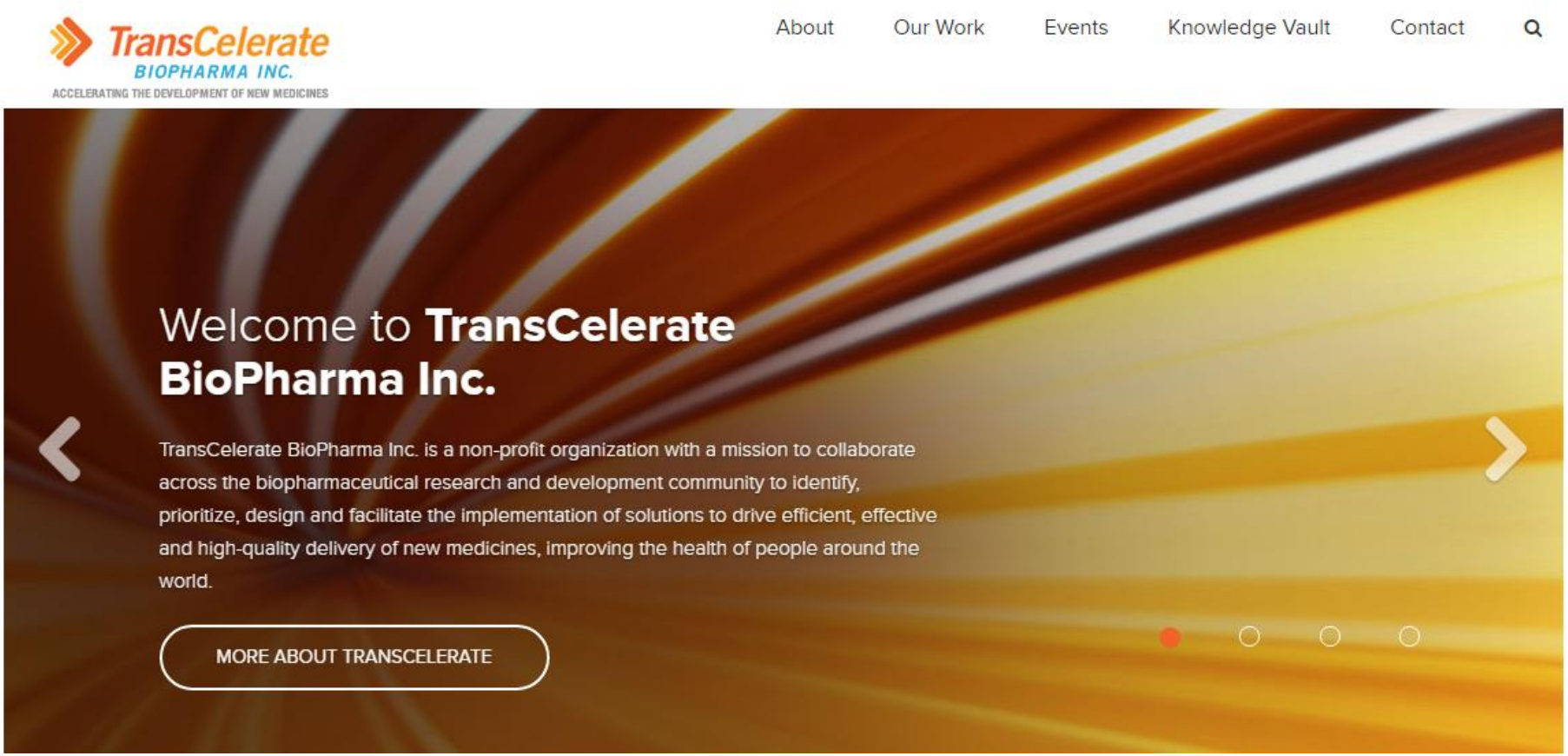
to enable…

- Improved Protocol Design
- Improved patient recruitment
- Increased quality
- Decrease cycle times
- Improved decision making

... by providing accessible, integrated and actionable data insights, state-of-the art tools, to improve drug development from protocol design, to data driven decision making in every stage of execution.

TransCelerate

A cross-pharma consortium with the objective of improving drug development times by increasing cross industry partnerships



TransCelerate Placebo Standard-of-Care workstream

We have access to placebo and standard-of-care data from dozens of partner organisations, helping in the design of studies for which GSK may have little prior experience

01

Provide an update on the progress of establishing a placebo & standard of care database

02

Share and get feedback on emerging industry best practices for using external controls in confirmatory clinical trials

03

Share upcoming trials where these methods are planned for use

04

Develop a shared plan for broader use of methodology: request for joint working group to develop framework

We believe there is value in the use of historical controls in confirmatory trials beyond rare diseases and pediatrics

Would the agency be willing to engage with us to progress this initiative?

TransCelerate BIOPHARMA INC.

Teaching Clinical Statistics about TransCelerate and RDIP

- R&D objective: reduce cycle times
- Clinical Statistics objective: enthuse R&D about the potential for data re-use with formal Bayesian dynamic borrowing methods
- RDIP objective: demonstrate value to R&D, and upskill Clinical Statistics with regard to data science

- Devised a “Hackathon” with a clinical development objective
  - Design a new schizophrenia study and estimate the number of AEs likely to be observed for any given sample size
- Under the “cover” of this realistic task, the Hackathon workshop is used to introduce participants to
  - The RDIP environment
  - How to conglomerate data
  - How to implement simple dynamic borrowing

SPDS Introduction to RDIP

The re-use of data as informative (mixture) priors with weights updated via dynamic borrowing

Graeme Archer, Statistical Innovation Group 6 November 2017

Imagine if we could use historical data directly in our new studies. This would open the opportunity to

- reduce sample size in new studies, or
- increase precision/PoS for new studies

But it also increases the chance of a false positive. We can't just bring the old patient data into the new study as though they were "the same" as the new study's patients.

To do this (in the Bayesian setting) would be identical to choosing the posterior distribution (from the old data) and using this as an informative prior for the new study.

But we know:

- between-study heterogeneity exists
- patients aren't necessarily directly exchangeable -- depression patients in 1980s clinical trials aren't necessarily "the same" as depression patients in 2010s clinical trials.
- (the same applies to endpoint measurements).

So: **it's hopeless?** No. There exist (many) methods for incorporating information -- in the form of historical data -- into a new study, in the form of an informative prior distribution, but doing it in such a way as to reduce the impact of that information according to how non-exchangeable it is with the new patient data.

In other words (and heuristically):

- old data + new data are "the same" $\longrightarrow$ *use the old data to increase inferential precision*
- old data + new data are not "the same" $\longrightarrow$ *downweight the influence of the old data on posterior inference, according to the degree of "unlikeness".*

The model is introduced for simple dynamic borrowing…

Likelihood

Let R be the number of events on the control arm in the new study. Then

$$[R \mid \theta] \sim \text{Binomial}(n, \theta).$$

Prior

If we knew nothing about θ, then we'd probably assume a vague, Jeffrey's style prior

$$\theta \sim \text{Beta}(\tfrac{1}{2}, \tfrac{1}{2}).$$

Alternatively, if immediately prior to the new study's commencement, using identical patient inclusion/exclusion criteria (and so on), we had witnessed r_H AE events from a total of n_H patients, we'd likely be tempted to use an informative prior:

$$\theta \sim \text{Beta}(\tfrac{1}{2} + r_H, \tfrac{1}{2} + n_H - r_H), \text{ say.}$$

Certain uncertainty

The honest truth - even in the presence of historical data - is that we're uncertain which of those two priors ought to pertain.

What we'd really like: use the informative prior when we should, discard it when we shouldn't.

This motivates the use of a mixture prior for θ, which arises from a hierarchical model (with an extra parameter), which is normally written in shorthand like this:

$$\theta \sim \omega\text{Beta}(\tfrac{1}{2} + r_H, \tfrac{1}{2} + n_H - r_H) + (1 - \omega)\text{Beta}(\tfrac{1}{2}, \tfrac{1}{2}).$$

The choice of $\omega \in [0, 1]$ is up to us: it controls the *a priori* degree of confidence we have in the informative component of the posterior distribution.

Dynamic mixture priors

In general, with likelihoods in the exponential family, we can write

$$\pi(\theta) = \omega\pi_I(\theta) + (1 - \omega)\pi_V(\theta,$$

where the subscripts on the components of the mixture denote *Informative* and *Vague*.

How does it work?

Bayes' Theorem (combine historical belief with new data) is our natural tool

– Historical data used to generate predictive prior distribution for mean response in new trial

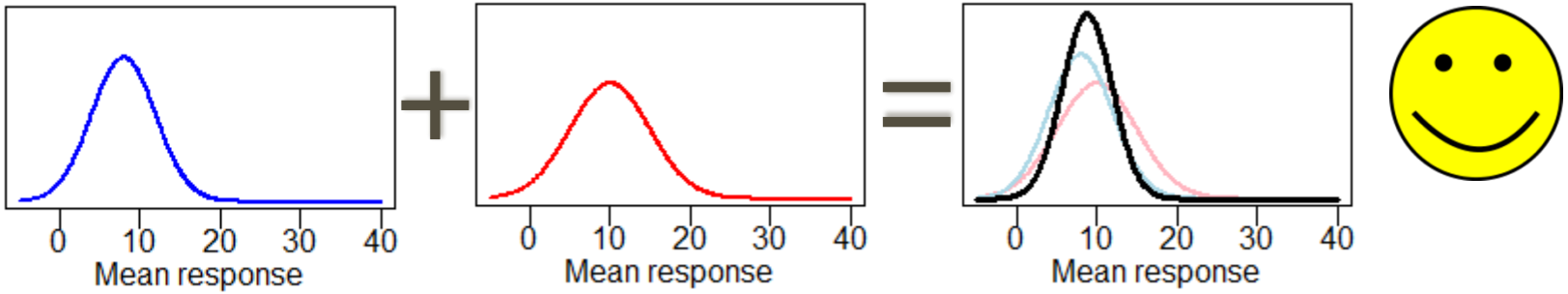
Predictive distribution for what we believe about future responses based on the historical studies ("prior")

What we see in the new study ("sampling distribution")

"Posterior" distribution: weighted average of prior and new trial data

Scenario 1

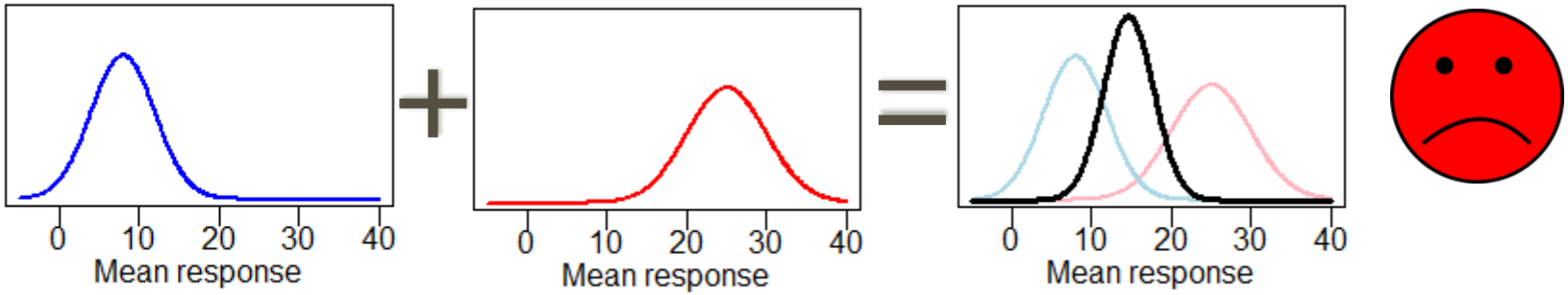
Historical and new data are consistent



– But, can result in potentially unrealistic estimates if historical data conflicts with new data

Scenario 2

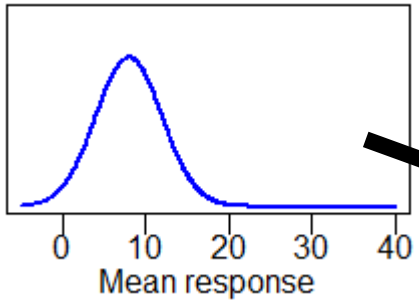
Historical and new data in conflict



"I beseech you, in the bowels of Christ, think it possible that you may be mistaken" (Cromwell's rule for historical data)

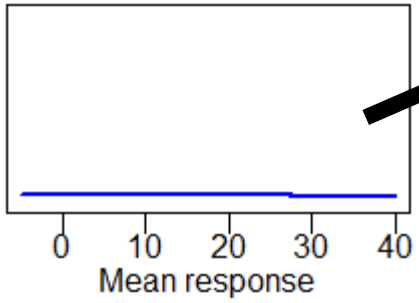
– Innovative design using robust (dynamic) Bayesian prior can be used to address the problem of conflict between historical and new data

Prior assuming historical data are relevant



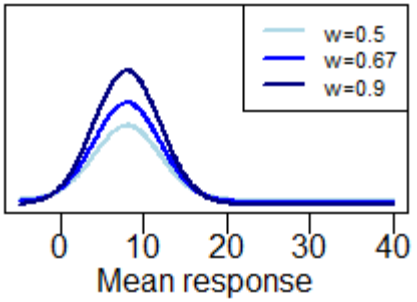
w

"In case we are mistaken" prior i.e. assuming historical data are not relevant



1-w

Robust prior = weighted mixture of these 2 priors



Schmidli et al (2014)

```r
# Next, use SQL language to pull out various SDTM datasest from the database where the
# Transcelerate data lives

trialSummary.df <-  sqlQuery(conn,
                             'SELECT studyid, domain, tsparm, tsval, etl_therapeuticarea
                              FROM placeboAndSocInsights.transcelerate_ts
                              ORDER BY studyid, tsparm')

# how many studies are there in this data-frame?
cat("No. of unique studies: ", length(unique(trialSummary.df$studyid)))

cat("\n", "in these therapeutic areas: ", "\n")
print(unique(as.character(trialSummary.df$etl_therapeuticarea)))

cat("\n", "First ten rows of trialSummary.df", "\n")
head(trialSummary.df, 10)

cat("\n", "With these unique datasets", "\n")
print(unique(as.character(trialSummary.df$tsparm)))
```

```
No. of unique studies:  83
 in these therapeutic areas:
 [1] "Amgen_Diabetes_NCT00093015_20170731" "Rheumatoid Arthritis"
 [3] "COPD"                                "Asthma"
 [5] "Diabetes"                            "Hypercholesterolemia"
 [7] "Vaccine"                             "Alzheimers Disease"
 [9] "Cardiovascular Disease"              "Stroke"
[11] "Ulcerative Colitis"                  "Schizophrenia"
[13] "Hidradenitis Suppurativa"
```

Next, the participants are shown how to use SQL queries in R to access and summarise the available data

We narrow-in on Schizophrenia studies…

```
# There are 15 studies which recruited Schizophrenia patients. Let's make a dataframe of only those
# Make a dataframe with Schizophrenia subjects, the phase of the study, and the number of subjects

schiz.df <- subset(trialSummary.df, etl_therapeuticarea == 'Schizophrenia'
                  & tsparm %in% c('Trial Phase Classification', 'Actual Number of Subjects'))

head(schiz.df, 20)
```

studyid	domain	tsparm	tsval	etl_therapeuticarea
M10503	TS	Actual Number of Subjects	72	Schizophrenia
M10503	TS	Trial Phase Classification	PHASE II TRIAL	Schizophrenia
M10854	TS	Actual Number of Subjects	68	Schizophrenia
M10854	TS	Trial Phase Classification	PHASE II TRIAL	Schizophrenia
M10855	TS	Actual Number of Subjects	144	Schizophrenia
M10855	TS	Trial Phase Classification	PHASE II TRIAL	Schizophrenia
M13608	TS	Actual Number of Subjects	51	Schizophrenia
M13608	TS	Trial Phase Classification	PHASE II TRIAL	Schizophrenia
R076477BIM3001	TS	Actual Number of Subjects	122	Schizophrenia
R076477BIM3001	TS	Trial Phase Classification	PHASE III TRIAL	Schizophrenia
R076477SCA3001	TS	Actual Number of Subjects	107	Schizophrenia
R076477SCA3001	TS	Trial Phase Classification	PHASE III TRIAL	Schizophrenia
R076477SCA3002	TS	Actual Number of Subjects	95	Schizophrenia

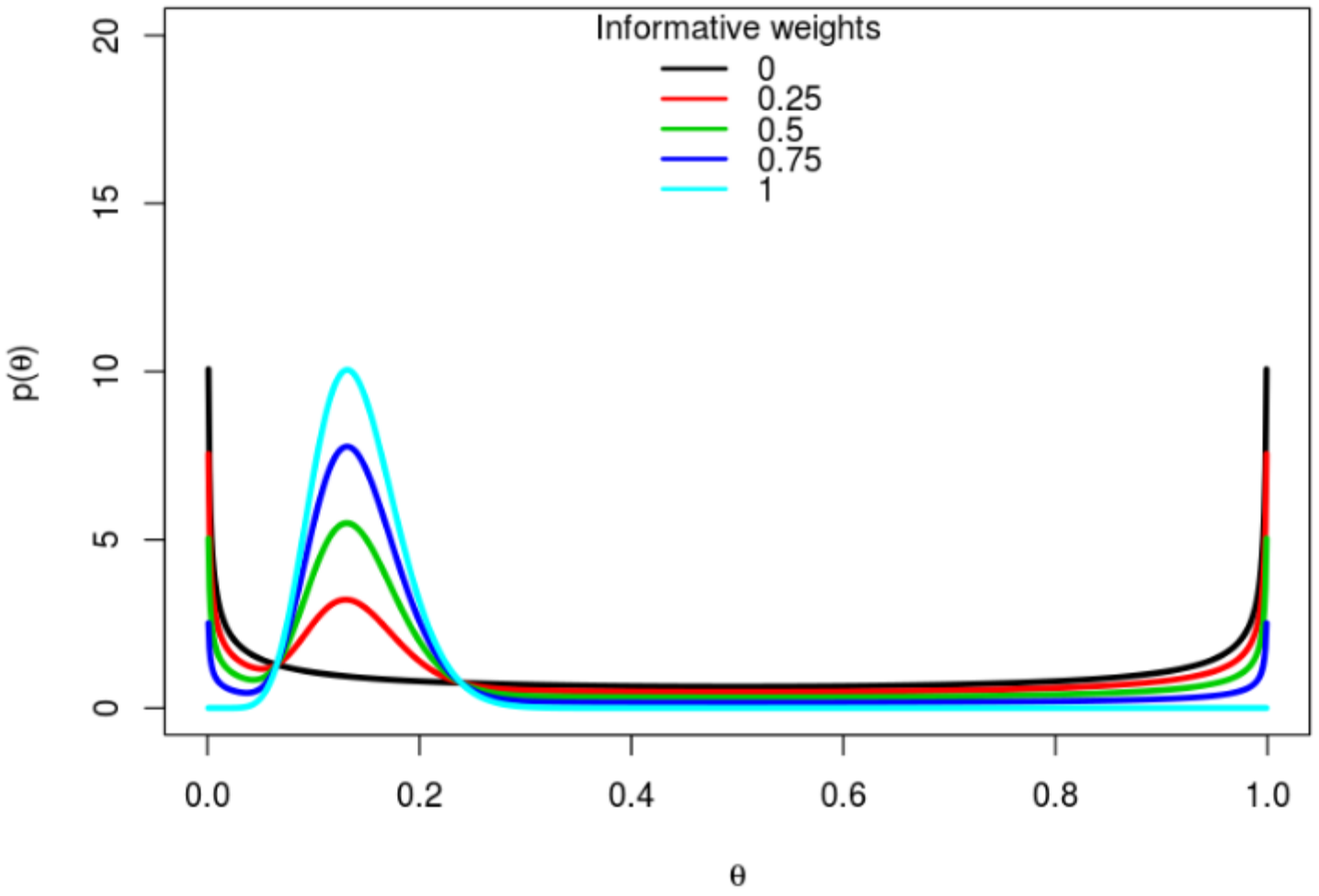
We end up, after a few R steps, with a dataframe showing the available historical Schizophrenia studies, their phase of development, and the numbers of subjects (with AEs, and in the entire study.)

studyid	aeterm	nbAdverseEvents	domain	etl_therapeuticarea	Actual Number of Subjects	Trial Phase Classification
M10503	NA	NA	TS	Schizophrenia	72	PHASE II TRIAL
M10854	TREMOR	1	TS	Schizophrenia	68	PHASE II TRIAL
M10855	TREMOR	1	TS	Schizophrenia	144	PHASE II TRIAL
M13608	TREMOR	1	TS	Schizophrenia	51	PHASE II TRIAL
R076477BIM3001	TREMOR	1	TS	Schizophrenia	122	PHASE III TRIAL
R076477SCA3001	TREMOR	8	TS	Schizophrenia	107	PHASE III TRIAL
R076477SCA3002	TREMOR	5	TS	Schizophrenia	95	PHASE III TRIAL
R076477SCH301	TREMOR	25	TS	Schizophrenia	102	PHASE III TRIAL
R076477SCH3015	TREMOR	27	TS	Schizophrenia	238	PHASE IIIB TRIAL
R076477SCH303	TREMOR	4	TS	Schizophrenia	127	PHASE III TRIAL
R076477SCH304	TREMOR	3	TS	Schizophrenia	110	PHASE III TRIAL
R076477SCH305	TREMOR	9	TS	Schizophrenia	123	PHASE III TRIAL
R092670PSY3001	TREMOR	8	TS	Schizophrenia	204	PHASE III TRIAL
R092670PSY3003	TREMOR	2	TS	Schizophrenia	96	PHASE III TRIAL
R092670SCH201	TREMOR	2	TS	Schizophrenia	84	PHASE II TRIAL

Participants were then asked to consider how they would narrow down their selection

Allows us to introduce the systematic-review component of any data re-use exercise

Each participant then selected out the studies they wanted to use as the historical component of their meta-analytic prior (so each participant would have a slightly different experience.)


Informative weights
0
0.25
0.5
0.75
1
p(θ)
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0
θ

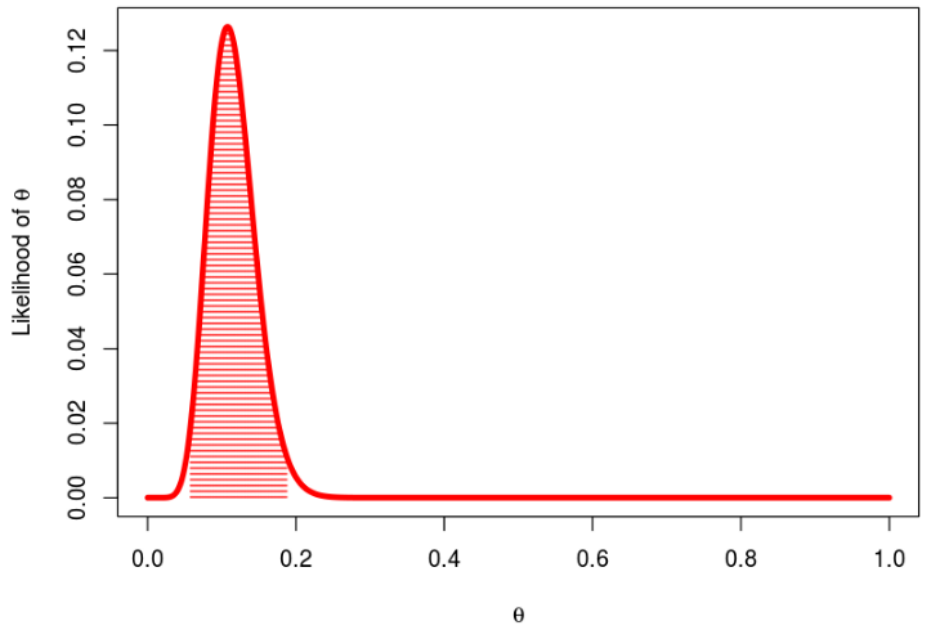
Generating new data

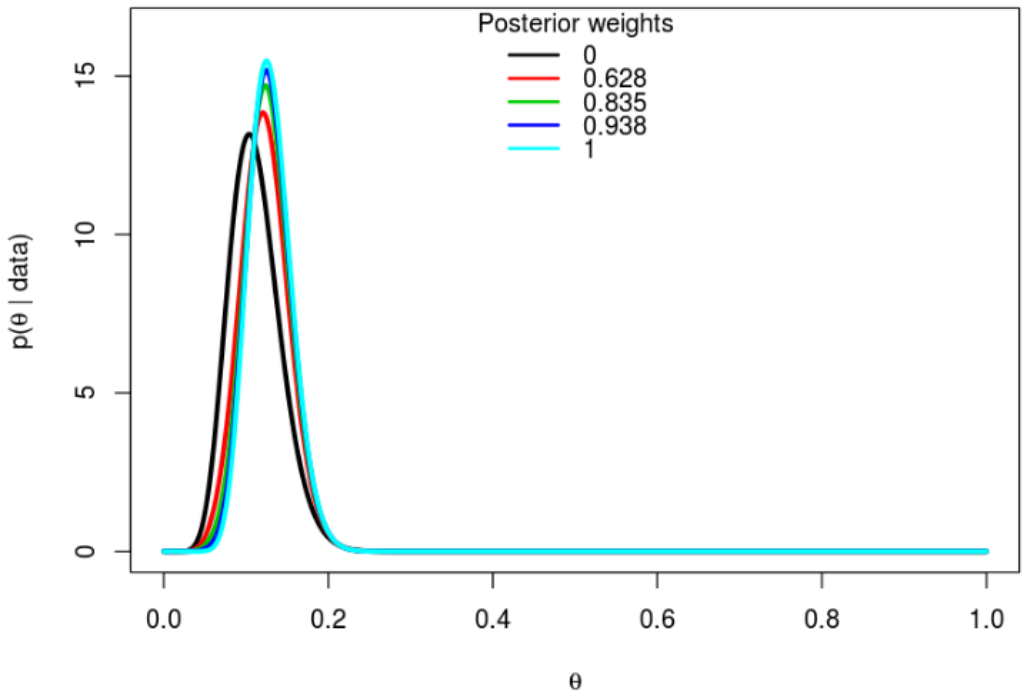
OK: we've specified a prior (assuming you've chosen a value for ω.) Before we can do posterior inference about θ, we'll need some "new" study data.

Let's frame this new data up as the first section of a new study: that is, we're going to make posterior inference about θ at an interim analysis, using the mixture informative prior.

We'll pick a "true" (study level) value for θ, and the sample size of the new study's interim data set, at random. We'll calculate the MLE based on the interim data and draw its likelihood, shading in a 95% confidence interval (boo! hiss!).

Teams were then shown how to simulate new data to take the place of their new study



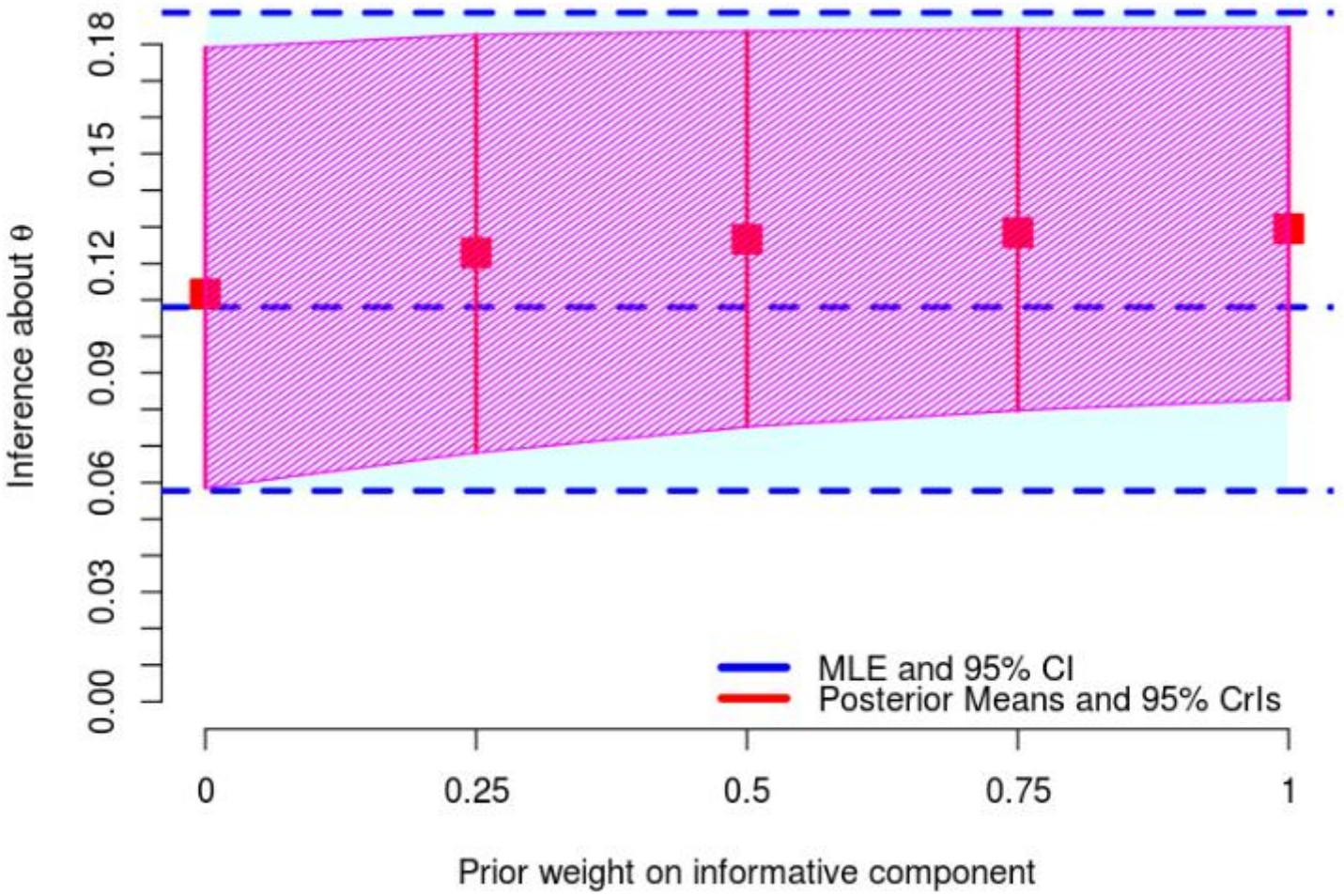


Participants then produced posterior distributions that combined a range of prior weights on their choice of historical data with their new (simulated data).

Key learning: the dynamic borrowing algorithm has updated the weight on the historical data

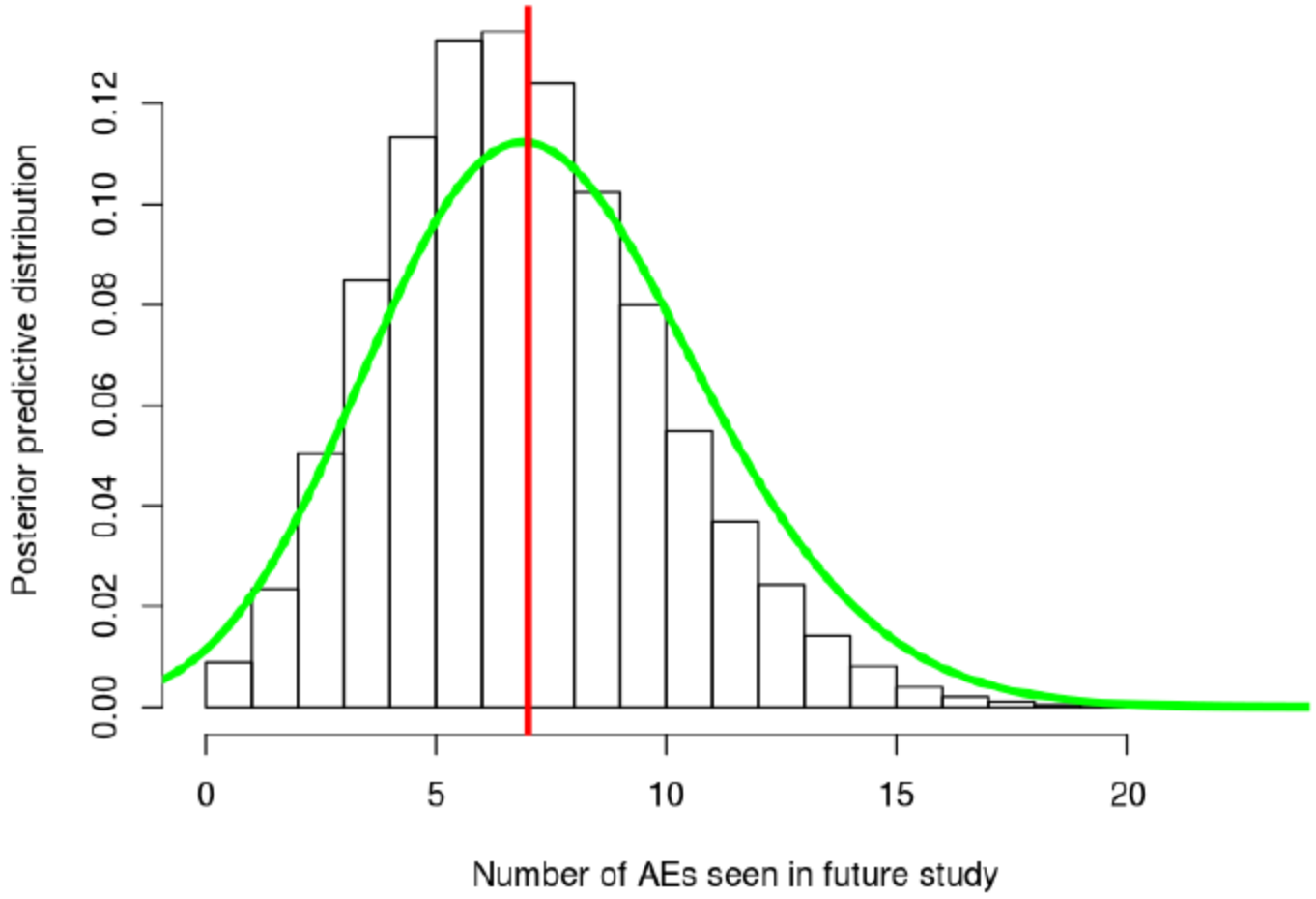
	PriorWeight	PosteriorWeight	FoldChange	Delta
1	0.00	0.000	Can't do it aargh!	0.000
2	0.25	0.628	2.511	0.378
3	0.50	0.835	1.67	0.335
4	0.75	0.938	1.251	0.188
5	1.00	1.000	1	0.000

Posterior (and likelihood) inferences in this example over the range of prior weights



Posterior predictions are now straightforward…

```
The probability to observe more than  7 AEs in a future sample of size 75  patients is  0.452
```



Conclusions

- The TransCelerate initiative is a great source of information for statisticians designing studies in therapy areas with which they might have little prior experience
- The RDIP platform at GSK provides user-friendly access to data *and* computational statistical functionality
- Bayesian dynamic borrowing has the potential to either reduce the sample size of future studies or to increase their decision-making power
- Our experience is that a hands-on workshop is a great tool to enthuse statisticians about (1) new statistical methods and (2) new sources of historical data and (3) new working environments – all at the same time.