A Bayesian approach to the 3-parameter Emax model for the assessment of dose response and dose comparison

Toby Batten

Cmed Clinical Services

Introduction

The Emax model is now a well-established technique for assessing the dose response relationship for a new drug during early phase clinical trials. Through Emax modelling it is possible to estimate the maximal treatment effect, the dose which produces 50% of the maximal effect and the placebo effect. The 3-Parameter Emax model has the following equation:

$$f(d,\theta) = E0 - \frac{Emax \times dose}{dose + ED50} + additional\ effects$$

Where E0 is the effect in the absence of treatment, ED50 is the dose which gives 50% of the maximum effect and Emax is the maximum effect. These are displayed graphically in figure 1.

Bayesian analysis enables us to incorporate historical data into our statistical models. The availability of historical data gives justification for a reduced sample size. In early phase clinical studies this is often not applicable to active treatment groups, however it can be applied to the control group, especially in disease areas where a number of clinical trials have already been performed. This is crucial to the Emax model, and specifically the E0 term.

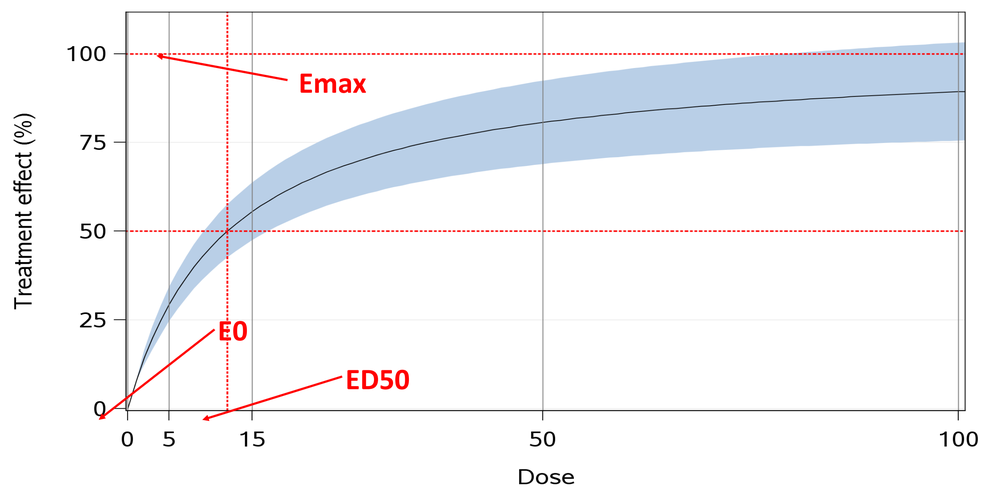


Figure 1: The 3-Parameter Emax curve

Analysis

Before starting the analysis the desired model should be fully parameterised. E0 is not the only parameter which should be assigned a prior distribution. Emax should be given a normal prior, precision (1/variance) a half-normal prior and ED50 a Beta prior. The coefficients for any additional effects should also be assigned relevant priors. Unless validated prior information is available these prior distributions should be non-informative. Therefore high standard deviations for normal distributions and a and b values of ≤1 for Beta distributions.

To ensure the model coefficients are being drawn from a converged distribution it is recommended to discard an initial proportion of the simulations. The number of simulations to be run should be specified as well as any thinning rate. Starting values and a seed number will fix the point of the initial simulation and allow replicable results. The seed number can be randomly generated independently. In general starting values should be logical, for example the minimum response for E0, the maximum response for Emax.

Model suitability can be assessed by examining diagnostic plots. Trace plots of the simulation means should be a random scatter, representative of white noise. Low autocorrelations indicate a smooth distribution. High autocorrelations are symptomatic of 'clumps' within your simulated samples and should be removed through thinning. The shape of the posterior density curve should be representative of the expected distribution.

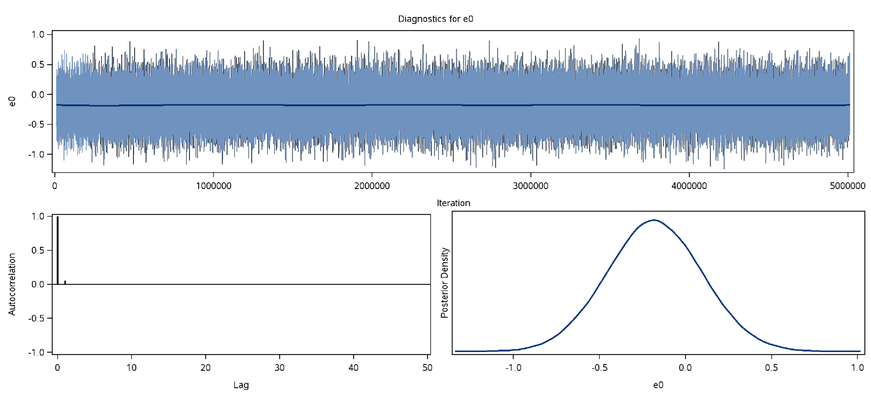


Figure 2: Example of ideal diagnostic plots

Programming

Programming can be performed in WinBUGS or R, however we have used the MCMC procedure in SAS. The following is example code from an Emax model with two additional continuous covariates.

```
proc mcmc data=data1 nbi=10000 nmc=5000000 thin=50 seed=10524
          monitor=(e0 ed50 emax sigma coeff1 coeff2
                   dose0 dose5 dose15 dose50 dose100
                   diff5 diff15 diff50 diff100)
          stats(alpha=(0.1)) plots(smooth)=all ;

 parms precision x1;
 parms e0        x2;
 parms ued50     x3;
 parms emax      x4;
 parms coeff1    x5;
 parms coeff2    x6;

 prior precision ~ normal(0, sd=100,lower=0);
 prior e0        ~ normal(0, sd=0.5);
 prior ued50     ~ beta(a=0.5, b=0.5);
 prior emax      ~ normal(0, sd=100);
 prior coeff1    ~ normal(0, sd=100);
 prior coeff2    ~ normal(0, sd=100);

 beginnodata;
  diff5=((e0 + ((emax*5) / (5+ed50)))) -(e0);
  diff15=((e0 + ((emax*15) / (15+ed50)))) - (e0);
  diff50=((e0 + ((emax*50) / (50+ed50)))) -(e0);
  diff100=((e0 + ((emax*100) / (100+ed50)))) - (e0);

  sigma = sqrt(1/precision);
  ed50 = ued50*300;
 endnodata;

 mu = e0 + (emax*dose) / (dose+ed50) + coeff1*parm1 coeff2*parm2;

 model response ~ normal(mu,sd=sigma);
run;
```

Simulation specifications

Parameters of interest

Starting values

Prior distributions

Calculation of treatment comparison estimates

Convert ED50 from proportion to dose

Model equation

Discussion and conclusions

Introducing Bayesian inference to the Emax model gives us the opportunity to achieve more reliable responses from our dose-response studies with less data. Through SAS proc MCMC, not only can we calculate the terms of the Emax model but we can also adjust for any additional effects and compare the posterior samples to establish individual dose effects and differences. Therefore this approach can give comprehensive insight into drug efficacy as well as identifying the most effective dose. Detailed specifications for model parameterisation is key, as with any Bayesian analysis there is more than one possible result. Therefore early thinking and anticipating the potential pitfalls will lead to a more efficient analysis.

In this poster we have only discussed the 3-parameter Emax model, however adaptation of this approach to the 4-parameter is possible. The 4-parameter Emax model introduces a slope factor which allows for greater flexibility over the shape of the Emax curve.

PSI annual conference 2015, London, UK

www.cmedresearch.com

Pattern Mixture Modelling Approaches for Time-To-Event Data

Dr. James Bell[1] and Simon Fink[2]

[1] Boehringer Ingelheim Ltd, Ellesfield Avenue, Bracknell, RG12 8YS, UK. Email: james.bell@boehringer-ingelheim.com
[2] Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorferstraße 65, 88397 Biberach an der Riß, Germany

Introduction

Censoring At Random (CAR) is a necessary assumption in most common time-to-event (TTE) analysis techniques used for clinical trials, including Kaplan Meier analysis, Cox regression and the log-rank test. However, it is also a strong assumption that is not likely to be that realistic in many cases.

In this poster, we present a set of three related sensitivity analyses around the assumption of CAR in TTE data. These approaches are based around Kaplan Meier (KM) imputation, a method analogous to that of multiple imputation in longitudinal data. In each method, adjustments are made to the imputation procedure to reflect assumptions about the likely or assumed behaviour of patients after censoring.

Throughout, a set of Progression Free Survival (PFS) oncology data is used to demonstrate the real-world application of these methods. This comprises 345 patients with 99 censorings (28.7%) across two treatments; active Treatment A and reference Treatment R. Analysis is presented based upon both the Cox Proportional Hazards model and the Log Rank Test. Results are given for an unstratified analysis. In all cases, 100 imputations are performed.

Kaplan Meier Imputation

Kaplan Meier imputation is a multiple imputation technique that uses KM curves to define the imputation distribution. Bootstrapping is used to derive a separate data set for each imputation. This provides a more accurate estimation of the variance after imputation. KM curves are then created for each treatment within each bootstrapped data. For each censored observation, the appropriate KM curve is then rebased to a probability of 1 at the time of censoring. A random draw is then taken from a standard uniform distribution and used to impute an event time using the following rules: The time corresponding to that probability from the KM curve is taken as the imputed event time. If the probability is lower than any in the KM curve, then the imputation is a censoring at the time of the last event. For censored observations after the last event in the KM curve, no imputation is performed.

This method has been previously shown to reproduce the KM estimator (1), To test our implementation of the method, we compared the results from the imputed and original data (see Table 1 for results) using the Cox Proportional Hazards model and the Log Rank Test.

The statistics can be seen to be very similar between the two data sets. The KM curves for the imputed data (not shown) also closely follow those from the original data. The one difference is the log rank statistic (although the corresponding p-values are comparable). This is due to both the statistic itself and its variance being directly dependent upon the number of events. Further work has demonstrated that if a complete data set is randomly censored and then imputed, the log rank statistics for the complete data and the imputed data are very similar.

Delta Adjustment

To test the robustness of conclusions to the CAR assumption, delta adjustment may be used (2,3). This is the use of a fixed penalty, δ, to reduce the expected time to an event after censoring for the active treatment arm. To implement this with KM imputation, all probabilities for the active treatment curve are adjusted to the power of δ. The CAR case corresponds to δ=1, while δ>1 is penalising and δ<1 is beneficial. Figure 1 shows the impact of applying different values of delta to a KM curve.

For the example data, a sensitivity analysis was performed with a value of delta of 3 for treatment A, corresponding to considerably higher probabilities of events occurring soon after censoring in the active arm. The resulting KM curves may be found in Figure 2, with summary statistics in Table 1.

Despite this high degree of penalisation of censoring in the active treatment, the resulting analysis still shows statistical significance.

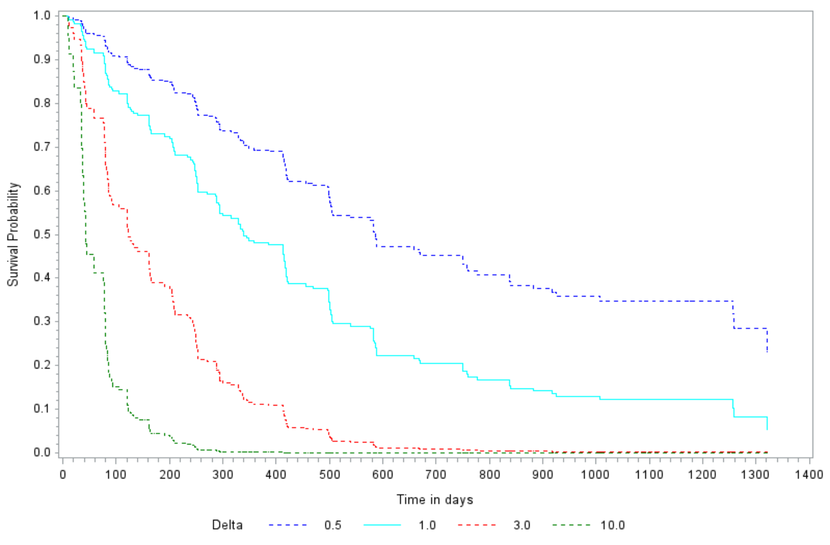


Figure 1 Kaplan Meier curves derived from the active treatment arm (A) of the example data that are used for imputation of censored data using the delta adjustment method. Curves are shown for δ = 0.5, 1, 3 and 10.

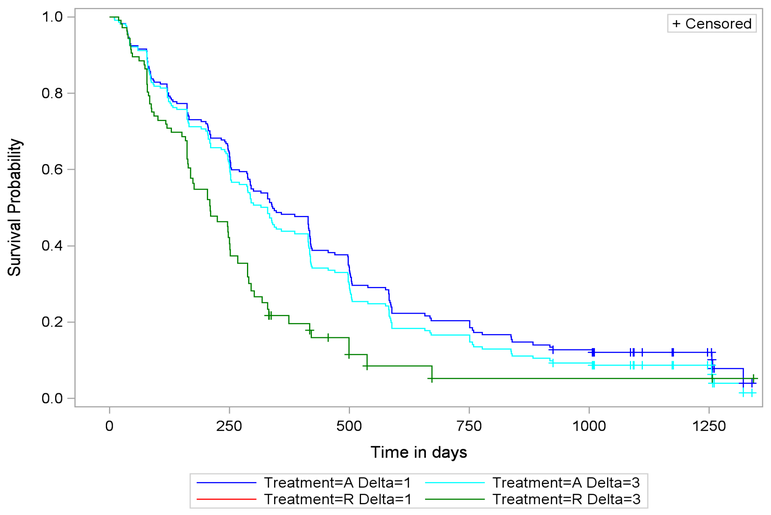


Figure 2 Kaplan Meier curves for the example data imputed using unadjusted and delta adjustment (δ=3) Kaplan Meier Imputation methods. The two curves for treatment R are identical and overlay.

Reference-based Imputation

A novel method of testing the robustness of the CAR assumption is reference-based imputation whereby patients who discontinue from the active treatment are assumed to behave like the control upon dropout. Where the active treatment is more effective than the reference, the impact of this is usually to increase the event rate for the active treatment arm.

To implement this by KM imputation, the KM curve of the reference arm is used to impute censored observations from both the active and reference arms. The reference arm is therefore imputed normally. As an example, Figure 3 shows the effective survival function used to impute the active treatment censoring at day 207. A similar curve could also be constructed for each other active treatment censoring.

KM curves for imputed data can be found in Figure 4, and the summary statistics for this approach are shown in Table 1.

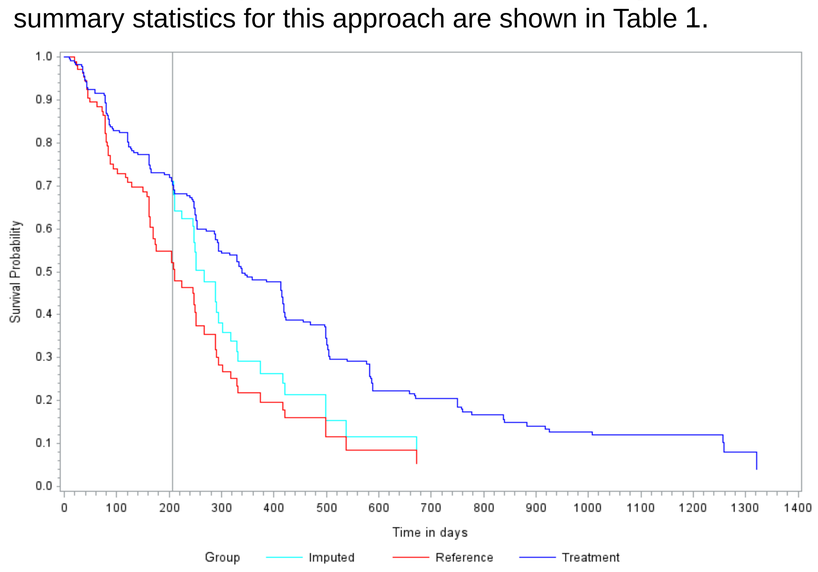
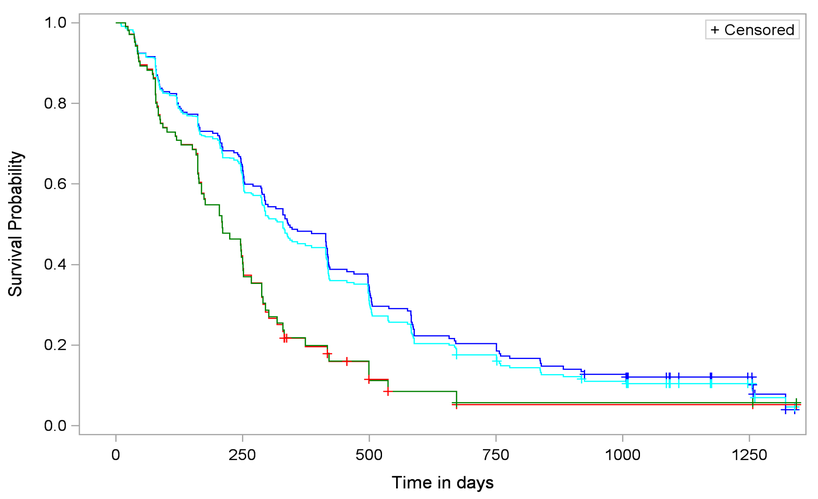


Figure 3 The Kaplan Meier imputation curve for a censoring in the active treatment arm at day 207 (vertical line). The reference and active treatment KM curves are shown for comparison.



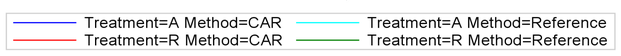


Figure 4 Kaplan Meier curves for the example data imputed using unadjusted (Censoring At Random, CAR) and reference-based Kaplan Meier Imputation methods.

Data	Original Data	Unadjusted KM Imputed Data	Delta = 3 Adjusted Imputation	Reference-Based Imputation	Pattern Mixture Imputation
Treatment A Median	340 days	340 days	330 days	330 days	318 days
Treatment R Median	210 days	210 days	210 days	210 days	209 days
Hazard Ratio - Cox (95% CI)	0.571 (0.428, 0.762)	0.559 (0.414, 0.755)	0.628 (0.467, 0.844)	0.599 (0.451, 0.797)	0.601 (0.453, 0.795)
t-score – Cox Log(HR) (p-value)	-3.81 (0.0001)	-3.80 (0.0002)	-3.09 (0.0021)	-3.53 (0.0004)	-3.57 (0.0004)
Log Rank Statistic (95% CI)	-23.1 (-34.9, -11.4)	-35.3 (-52.9,-17.7)	-29.5 (-47.8, -11.3)	-31.8 (-49.1, -14.5)	-32.0 (-49.2, -14.8)
t-score – log-rank test (p-value)	-3.86 (0.0001)	-3.95 (<0.0001)	-3.17 (0.0016)	-3.61 (0.0003)	-3.65 (0.0003)

Table 1 Summary statistics for the example data set using all methods described.

Pattern Mixture Modelling

If censoring at random is assumed not to hold, then the reasons for censoring may provide information about the likely time to event. Consequently, it may be desirable to reflect this in a 'realistic' sensitivity analysis. This is achievable using KM imputation by borrowing the concept of pattern mixture modelling from longitudinal data analysis.

Patients may be assigned to patterns according to e.g. reasons for discontinuation. Each pattern may then be imputed using different rules to reflect likely post-discontinuation behaviour. These rules are implemented by allowing each treatment within a pattern to be imputed using the KM curve of either treatment, and with a specified delta adjustment. For our example data, we provide a complicated analysis based upon 4 patterns, with patterns and rules defined in Table 2.

For pattern 1, patients starting a new therapy, the active treatment is reference-imputed with an additional penalty as censoring may reflect their deteriorating health and they may start an inferior treatment. The reference treatment is imputed normally. Pattern 2 is imputed assuming censoring at random as there is no indication of worsening health being associated with their censoring. Patterns 3 and 4 are imputed with penalties to reflect that these patients were known to progress/die but were censored as it was not possible to obtain an accurate time of progression/death.

Pattern Number (patients)	Censoring Description	Treatment A Imputation Curve	Treatment A Imputation Delta	Treatment R Imputation Curve	Treatment R Imputation Delta
1 (61)	New therapy	Treatment R	2	Treatment R	1
2 (30)	No progression & censored	Treatment A	1	Treatment R	1
3 (6)	No data before progression	Treatment A	5	Treatment R	5
4 (2)	Missing data then death	Treatment A	3	Treatment R	2

Table 2 Summary of reason for dropout patterns and the associated imputation rules for an example scheme. Curve columns give the treatment curve that is to be used for imputation. Delta columns give the value of delta to be applied to each curve.

KM curves for the imputed data may be seen in Figure 5 and summary statistics in Table 1. Unlike the previous method, both curves show systematic deviation from the unadjusted imputation curves towards worsening PFS.

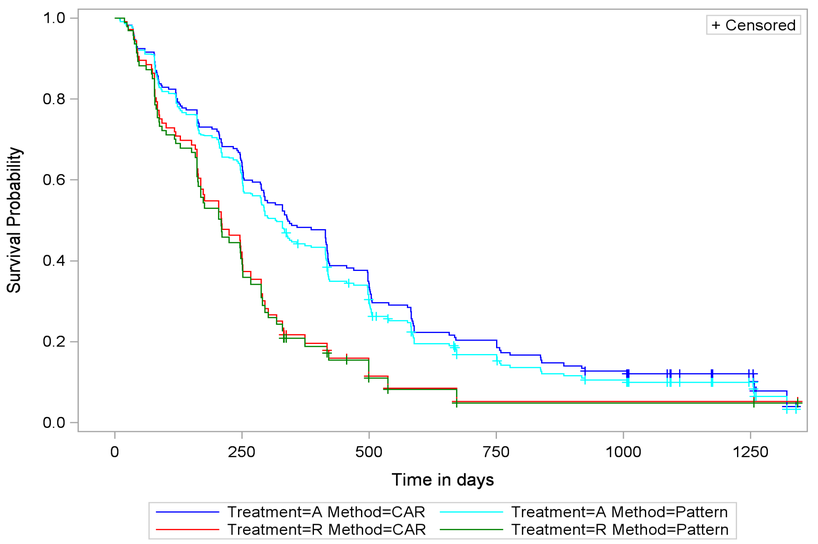


Figure 5 Kaplan Meier curves for the example data imputed using unadjusted and pattern mixture Kaplan Meier Imputation methods.

Discussion and Conclusions

This poster has demonstrated methods for performing sensitivity analyses for the censoring at random (CAR) assumption in time-to-event analysis. These methods may also be used to address estimands that account for treatment discontinuation or treatment switching.

A key strength of these methods is they only require simple and transparent assumptions that can be easily translated to and from clinical understanding. They also allow for a more realistic middle-ground between accepting CAR and imputing censorings as events on the day of censoring.

For the example data, all sensitivity analyses produced 2-sided p-values considerably below 0.05, and consequently the original conclusion of significance can be considered robust to deviations from the assumption of CAR. This robustness is caused by the extremely significant p-value from the original analysis and is despite the large deviations observed in the sensitivity analyses.

All methods have been implemented using standard SAS code and it is hoped to make the programs available in the near future.

References

1) Taylor J M G., Murray S, Hsu C; Statistics and Probability Letters **2002**, *58* 221-232: "Survival Estimation and Testing via Multiple Imputation".
2) O' Kelly M, Lipkovich I; 2014 PSI Conference presentation: "Using Multiple Imputation and Delta Adjustment to Implement Sensitivity Analyses for Time-to-Event Data".
3) Zhao Y, Herring A H, Zhou H, Ali M W, Koch G W: J Biopharm. Stat., 24(2):229–253, 2014. "A multiple imputation method for sensitivity analyses of time-to-event data with possibly informative censoring."

Predicting the date when the n^{th} event will occur

Sandrine Cayez

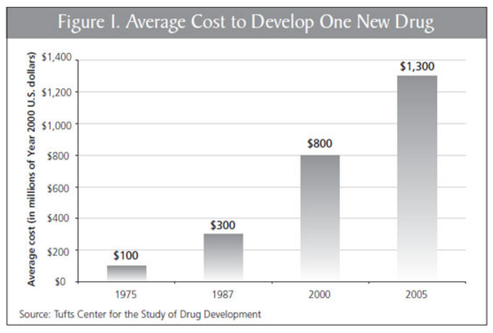
Cmed Clinical Services

Introduction

NEWSFLASH: STATISTICS INVOLVED IN PLANNING OF END OF RECRUITMENT AND INTERIM ANALYSES!

The cost of drug development for pharmaceutical companies continues to increase. In a world where it is harder to find compounds and get them marketed, it is increasingly important to decrease the cost of drug development but also to reduce the duration between when a compound is discovered to the time it is marketed. The aim is to bring better and safer treatment to patients. What are pharmaceutical companies doing to better manage their clinical trials?

Some very innovative companies involved the use of statistics to try and predict when the study recruitment will be completed or when the Interim Analysis will occur.



Using simulation in SAS and known distributions for recruitment (Poisson distribution), for survival data types (exponential, log normal or Weibull distribution) with maybe just a sprinkling of Bayesian can enable us statisticians to provide the team (and the stakeholders – never forget them) with the estimated date of interest and the most important part - confidence intervals!

Are Statisticians also Wizards?

MAGIC STUDY TO DATE:
- 158 RANDOMIZED
- 67 DEATHS
- IA: AFTER 200 DEATHS

Recruitment: Poisson Distribution

Death: Exponential, Log-normal or Weibull distribution

Forecasting the Randomization

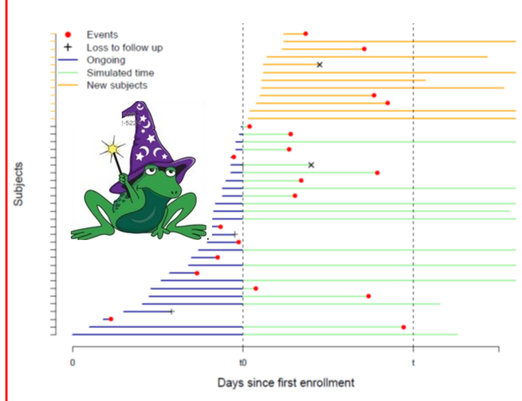
Step 1: Using current data (date of randomization), determine the randomization rate.

Step 2: Using the randomization rate and the Poisson distribution simulate the number of patients randomized each day until all patients are randomized (sample size).

Step 3: Repeat step 2 at least 1 000 times.

Step 4: The estimated date of when the last patient will be randomized is the median date of all the simulations. Associated 90% confidence intervals correspond to the 5% and 95% percentiles.

Forecasting deaths



Step 1: Using the actual data determine the time to death & posterior parameters of the selected distribution (e.g. Exponential).

Step 2: Repeat step 1 but for the time to lost-to-follow-up.

Step 3: Using the estimated posterior parameters of the distribution;

- For patients already randomized and still ongoing at the cut-off date, simulate the time of death and the time to lost-to-follow-up until the simulated times are greater than the censored times
- For patients not yet randomized at the cut-off date, simulate the date of randomization for each patient left to be randomized and then simulate the time of death and the time to lost-to-follow-up

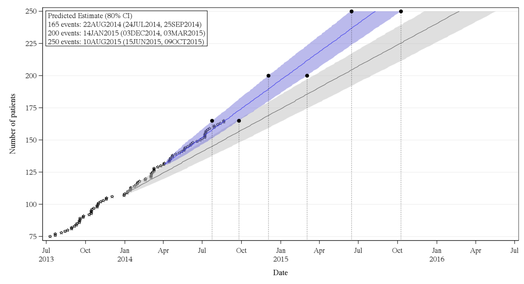
Step 4: Censor patients for which the simulated time to lost-to follow-up is greater that the simulated time to death.

Step 5: Repeat step 3 and step 4, 1 000 times.

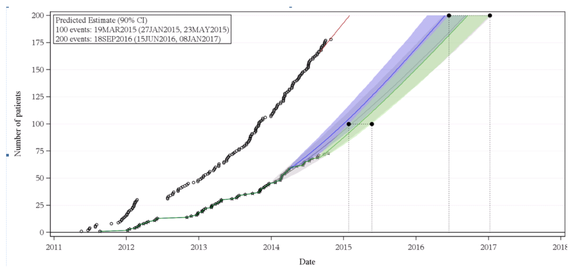
Step 6: The estimated date of when the n^{th} patient will die is the median date of all the simulations. Associated 90% confidence intervals correspond to the 5% and 95% percentiles.

01JAN2014 Prediction
01APR2014 Prediction
28AUG2014 Prediction
Forecasting death graphs
Actual randomization
Predicted randomization

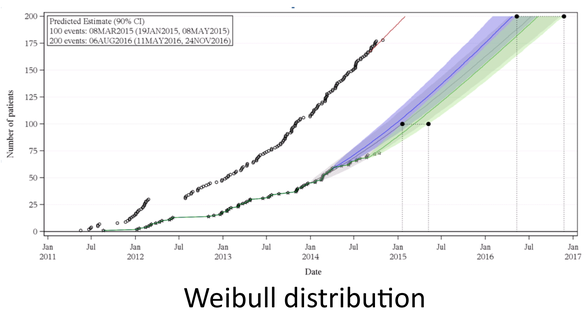
Event	Method	Cut-off	Estimate	CI90
Randomized	Poisson	250	10AUG2015	(15JUN2015; 09OCT2015)
Death	Exponential	200	18SEP2016	(15JUN2016; 08JAN2017)
Death	Weibull	200	06AUG2016	(11MAY2016; 24NOV2016)
Death	Log Normal	200	05OCT2016	(27JUN2016; 08FEB2017)



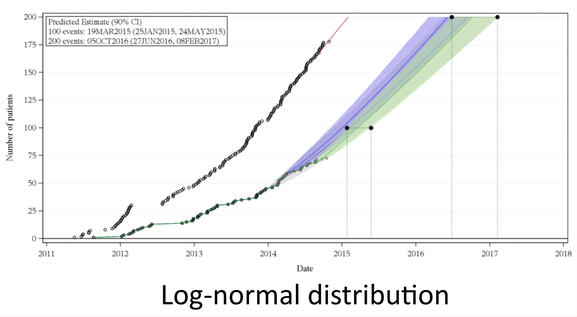
Randomization



Exponential distribution



Weibull distribution



Log-normal distribution

Conclusion

The statistical methods we statisticians apply to this very important problem do not have to be complicated nor require a large set of assumptions or lengthy computations. They rely on statistical assumptions (distribution of the data).

We can always complicate the model with covariates, known bank holidays and complex distributions. But is it really worth it as no clinical trial goes according to plan and predictions must be adjusted several times during the study based on the current data or additional input (e.g. new sites selected and opened)?

PSI annual conference 2015, London, UK

www.cmedresearch.com

An Event Driven Respiratory Trial

Nick Cowans, Abigail Fuller, Andrew Holmes
Statistics, Veramed Limited

1. Introduction

Chronic obstructive pulmonary disease (COPD) often coexists with other chronic diseases and comorbidities that can markedly influence patients' health status and prognosis. This is particularly true for cardiovascular disease (CVD).

This has led to assessment of inhaled COPD medications for overall survival benefits, often using event driven designs.

Such designs present numerous statistical issues, including:

- Predicting the common end date at which the target number of events will have accrued.
- The heterogeneity of the study population over time.
- Variable follow up due to the common end date.
- Missing data for secondary/tertiary endpoints following withdrawal from treatment.

In this poster we consider these issues as they arose in planning the analysis of The Study to Understand Mortality and Morbidity in COPD (SUMMIT) [1].

2. SUMMIT

The **S**tudy to **U**nderstand **M**ortality and **M**orbid**IT**y in COPD

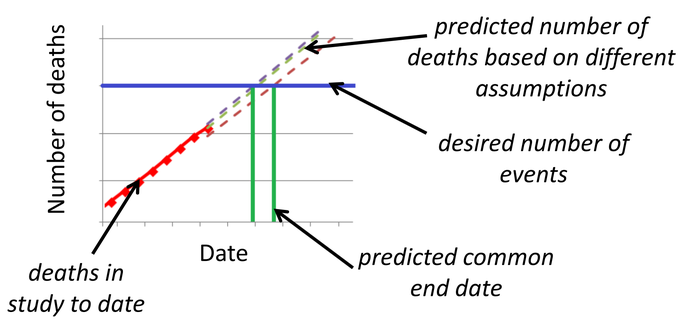
- An event driven, placebo controlled, long term, global, randomised clinical trial to investigate the impact of fluticasone furoate/vilanterol combination and the individual components on the survival of patients with moderate COPD and either a history of CVD or at increased risk for CVD.
- Secondary endpoints are rate of decline in FEV_1 and time to first cardiovascular event and there are various tertiary endpoints.
- FEV_1 is a spirometry lung function test that measures the volume of air that can be expelled in the first second from a maximum inspiration.
- In patients with obstructive diseases like COPD, instigating certain treatment regimes can cause an increase in FEV_1 at the initial visits following baseline, following this, FEV_1 will decline with time.

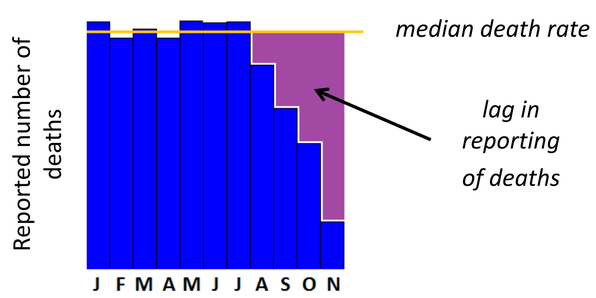
Contrast with the usual respiratory trials for symptomatic endpoints:

- Primary endpoint of time to death.
- Long follow up (1-4) years, with Investigational product (IP) treatment for the duration of follow up.
- Event driven: follow up is until a specified number of events have accumulated, with analysis at a common end date for all subjects.
- Complete follow up for vital status up to the common end date, even if the subject is withdrawn from IP, with negligible loss to follow up or withdrawal from the study.
- Large and multi-regional, with rolling start.
- Secondary & tertiary symptomatic respiratory endpoints (e.g. lung function, exacerbations, quality of life) are only collected whilst subjects are on-treatment.
- Due to the symptomatic relief afforded by these treatments, withdrawal from treatment may be related to (perceived) lack of efficacy, and may therefore be more prevalent in the placebo arm.

3. Common End Date

- The **common end date** was defined as the date at which the number of events that the study was powered for will have occurred. All subjects will be followed up to this date.
- To allow preparation for final visits, the common end date had to be decided in advance.
- In order to do this, it was necessary to predict at what time the desired number of events (deaths) would occur.





- Weekly death tracking was carried out in order to determine how many deaths were being reported. However, deaths occurring between scheduled contact, and other factors, meant that there was a lag in reporting of events (see above).
- Probability of death so far in the study (excluding the lag time) was used to predict the common end date (see left).
- An interim survival sweep where all subjects were followed up for vital status at a scheduled visit or over the telephone was carried out in order to get a more accurate sense of the true number of deaths.

4. Heterogeneity of patient population

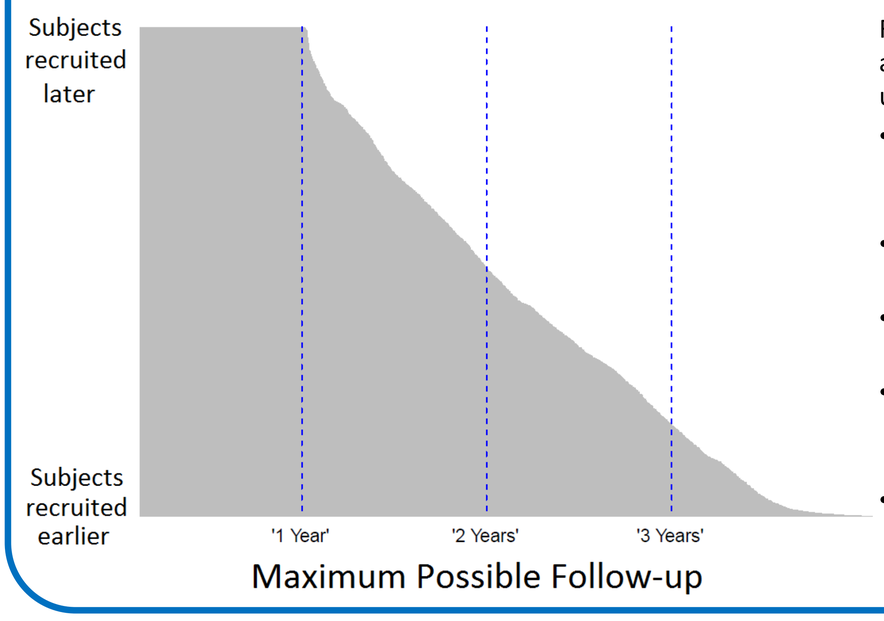
Patient population is heterogeneous over (follow up) time as some regions began recruitment before others but the study ends for all on the common end date.

- Regional differences in prognostic demographic and baseline characteristics.
- Contrasts with many respiratory studies with fixed duration follow up for all subjects.
- Danger of temporal plots being interpreted longitudinally.

To mitigate:

- Guard against longitudinal interpretation of plots, especially Kaplan-Meier curves [2].
- Primary analyses of repeated measures endpoints at 1 year.
- Include known prognostic factors in models and present adjusted plots (e.g. LSMeans for baseline OBSMARGINs for change in FEV_1).

5. Variable follow up



Recruitment occurred over three years, but the study ended for all subjects on the common end date, resulting in variable follow up amongst subjects:

- Subjects recruited later are unable to be in the trial long enough to accumulate long term data so have less potential follow up than those recruited earlier.
- Such censoring at the common end date unlikely to be informative. Assume Missing at Random (MAR).
- For time to event analysis, for example for the primary endpoint, this does not present additional issues.
- For events (e.g. Exacerbations), analyse time to first exacerbation in preference to rate (with the latter as supportive).
- For endpoints measured regularly (e.g. change in FEV_1), use Mixed Model Repeated Measures analyses assuming MAR, with primary analysis of effects at 1 year.

6. On-treatment only follow up

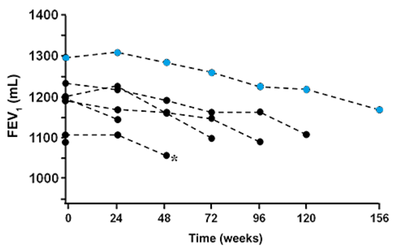
For the primary end point of all cause mortality, survival status will be collected until the common end date, even for subjects who withdraw from IP, with (almost) complete follow up expected.

However some secondary & tertiary endpoints (e.g. COPD exacerbations, FEV_1) are only followed up whist subjects are on-treatment and undergoing regular visits:

- To assess whether withdrawal from IP is related to outcome, and what the missing data post we considered “withdrawal cohorts” (right).
- Missing data post withdrawal from IP unlikely to be MAR.
- In a previous study it was shown that subjects who withdrew were more likely to be older and have baseline characteristics indicating poorer respiratory health (more prone to exacerbations and lower baseline FEV_1).
- Analyse rate of FEV_1 decline using random coefficient models, and assess sensitivity using imputation.

Withdrawal cohorts:

- Group of subjects who withdraw from IP at the same follow up time.
- Plots of outcome against follow up time by withdrawal cohort indicate that the degree of deterioration of respiratory health throughout the study contributes to the likelihood of withdrawal, shown in this previous, similar study:



Adapted from [3]: Subjects who prematurely withdrew from IP tended to do so once their FEV_1 had deteriorated to a certain point (1100ml).

*Key: each group of black dots connected with dotted lines represents the mean FEV_1 at each visit for subjects who withdrew before the next scheduled visit. For example, subjects in the * group withdrew between 48-72 weeks. The blue dots represent the mean FEV_1 at each visit for subjects who completed the study without withdrawing from IP.*

- This leads to an increased likelihood that subjects who make it to study completion are in better health than those recruited.

7. Summary

- In contrast to conventional fixed duration respiratory trials, event driven trials can reduce the length of the study for the primary endpoint.
- However, they add further complexity, especially when dealing with secondary and tertiary continuous endpoints such as FEV_1 over time or rate of exacerbations, which are only collected on treatment.
- It is difficult to predict the analysis date (common end date) accurately.
- Heterogeneity of patient population over time must be considered when making longitudinal inferences.
- Censoring due to the common end date can be considered missing at random but censoring due to withdrawal from treatment can not.
- Withdrawal cohorts are useful for assessing likely patterns of missing data.
- Explanation of event driven trial difficult in setting where fixed duration is considered the norm.

References

[1] Vestbo *et* al. (2013) “The Study to Understand Mortality and Morbidity in COPD (SUMMIT) study protocol” *Eur.Respir.J.* 41:1017-1022 DOI:10.1183/09031936.00087312

[2] Pocock *et al.* (2002) “Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls” *Lancet* 359:1686-89

[3] Vestbo *et al.* (2011) “Bias due to withdrawal in long-term randomised trials in COPD: Evidence from the TORCH study” *The Clinical Respiratory Journal* 5:44-49 DOI: 10.1111/j.1752-699X.2010.00198

5th Floor Regal House, 70 London Road, Twickenham TW1 3QS
+44 (0)20 3696 7240 | www.veramed.co.uk

Ensuring the quality of your data in Respiratory trials: Data management from a statistical standpoint

1. Abigail Fuller, Statistician, Veramed Limited
2. Nick Cowans, Statistician, Veramed Limited

1. Introduction

Large, global late phase studies inevitably involve huge amounts of data of varying quality. Data frequently needs cleaning up prior to locking the database, a responsibility typically lying with data management. The ability to look at multiple extracts of data while the study is ongoing and blinded has enabled us to develop novel methods and processes for increasing the confidence in data quality.

Forced expiratory volume in one second (FEV_1) is the volume of air expelled from the lungs in one second, this is measured in millilitres or litres depending on the equipment. Outliers in FEV_1 can commonly be caused by a decrease in subject effort, illness or equipment failure. These values would not be considered valid and including many subjects with these values in the analysis can cause variations that may not represent the true treatment differences. Looking at these in a visual way emphasises the importance of ensuring that these outliers are genuine data points.

A respiratory exacerbation is an event, such as pneumonia or a COPD exacerbation that effects the airways and hence the patients ability to breathe. Respiratory tract exacerbations may present over a period of a few days with symptoms progressing. Due to this, sometimes these events can be recorded as two separate exacerbations when they are actually the same event progressing over time. When rates of respiratory tract exacerbations are an endpoint, recording duplicate or overlapping events will alter results.

The need for identifying these two different situations is apparent. Previously, clinicians would spend time looking through vast amounts of data, however, these Patient Profile review tools have made clinical review a quick and easy process stressing the importance of these data on our endpoints.

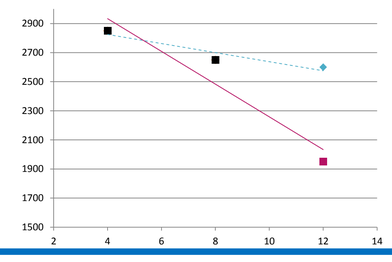
2. Measures of Interest

FEV_1 Rate of Decline

Rate in decline FEV_1 is an endpoint commonly used in respiratory trials. Rate of decline is measured in millilitres per year and a raw rate of decline can be calculated simply using linear regression.

When calculating the slope extreme values of FEV_1 caused by respiratory events or equipment malfunctions can lead to differences in the rate of decline that don't represent the true rate of decline of the subject. This is illustrated here in this example plot of FEV_1(mL) against Time (weeks). As illustrated in the figure, changing the final data point to be an outlier (red) changes the raw rate of decline quite substantially.

Subjects with outliers can be found easily, but the influence on the results is more difficult to explain without a visual/graphical option.



Rate of exacerbations

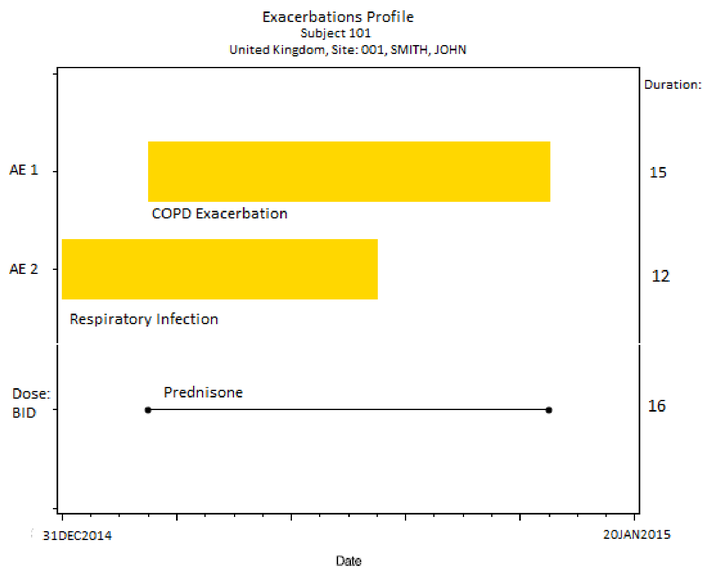
The rate of exacerbation events can be reduced by taking certain medications and hence this rate is also often used as an endpoint.

This can be easily calculated as the number of exacerbations divided by the time of treatment. For example, if one subject has one exacerbation in 6 months, the rate would be 2 exacerbations a year. This means that if an exacerbation event is recorded as two events instead of one, the subjects exacerbation rate per year can be as much as doubled. If this happens with a lot of subjects this can influence the results.

This data would traditionally be reviewed at a subject level, subjects with multiple respiratory events could be programmatically identified but data would be reviewed on a case by case basis, an inefficient and time consuming process.

3. Rate of Exacerbations

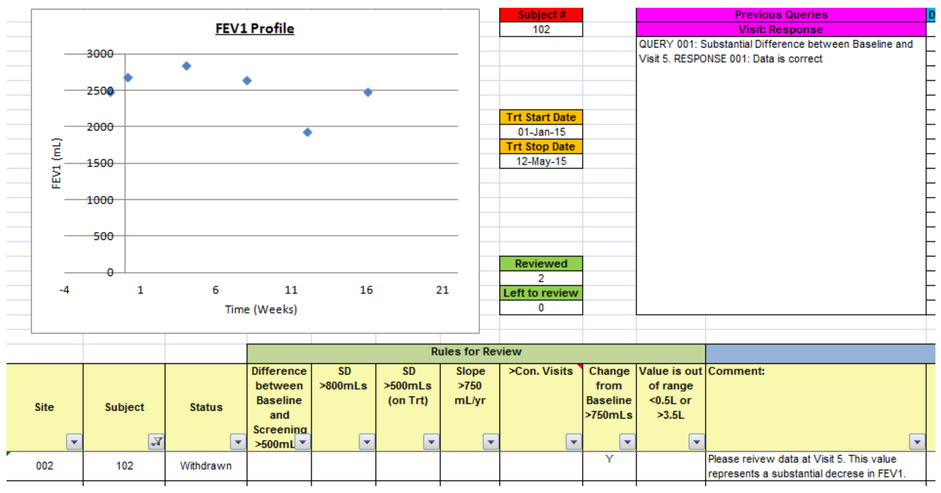
For Rate of Exacerbations, the data was presented in a patient profile bar plot, with one plot per overlapping event. Information on medications taken at the time of the respiratory exacerbation was included as well as duration of the event in days.
Presenting this information in a visual way, it is apparent that the 2nd Adverse event is not a second event but actually the same exacerbation event which probably started on 31st December. This event would increase the rate of exacerbations for this subject. In this example, the case is easy to deal with, however, when events are separated by a few days or even a week, the case is not so simple, and having the medications taken by the subject is more important.
These profiles were reviewed by the clinical team and further action was taken of querying the data or sending the exacerbation profile to the site and asking them to clarify the data. When the site can visually look at the profile, it is immediately clear if they have made any misrepresentation of what happened in the recorded data, and if not they can explain the differences for the clinical team to review.

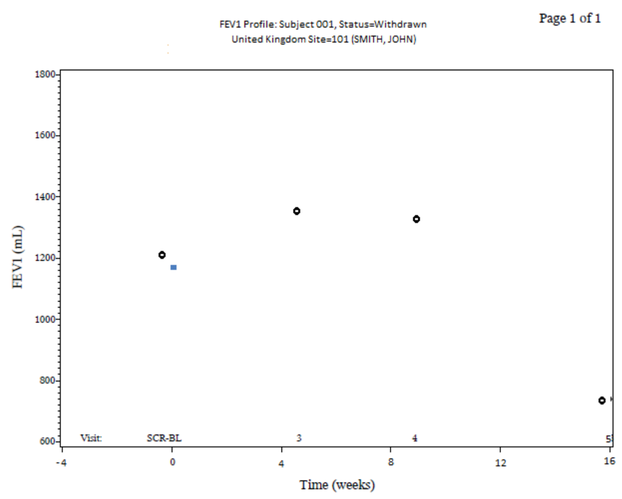


4. FEV_1 Rate of Decline

The need to have a visual tool to ease the clinical review was identified, with an emphasis on user ease and content needed to reach a clinical decision on further action.

Using a combination of excel and SAS, the following tool was developed. Some clinical review rules for FEV_1 rate of decline were identified including large differences between screening and baseline values and calculated individual rate of declines greater than 750mL/year. Utilising these rules, a review spreadsheet was created identifying the subjects programmatically in SAS. Filtering on subject number gives the individual subject information on one page, including treatment information, previous queries, responses and any previous review information.





Upon reviewing the spreadsheet, decisions can be made as to whether to take any further information, if a query has been answered previously with an appropriate clinical reason for the discrepancies in FEV_1 between visits then the decision may be made to not take any further action with the data, as it represents the true values. Previously, upon reviewing the data, query history for subjects would not have been available to view as easily.

Once all subjects have been reviewed and decisions whether to issue a profile made the spreadsheet can be read into SAS and a PDF version of the patient profile is created and issued to the site for review. This document contains the clinical review comment, the graphical representation of the FEV_1 data and space for the site investigator to comment to explain any large variations. Placing all of the information into one document makes the anomalies in the FEV_1 data visually clear to the investigator at the site, whilst providing anymore information they need to investigate.

The PDF profiles can be issued to sites for their review. If they see any immediate data entry issues with the FEV_1 data, it is assumed these will be corrected and the subject will no longer be picked up by the clinical rules set initially. Alternatively, the Principal investigator can return and comment on the discrepancies in the box supplied. These comments can also be entered into an electronic data capture system to track. The comments from the site should explain the discrepancies seen in the data and can be reviewed by a clinician to make sure they make sense medically.

5. Summary

- Reviewing study data in the traditional way can be a time consuming process.
- Data can be difficult to review when displayed in typical dataset standards such as listings. Having unique tools can help our clinical colleagues keep track of their review more easily.
- Visual tools can assist with understanding of impact on endpoints for both clinical reviewers and investigators at sites.
- Seeing an outlying on an FEV_1 plot, or seeing clinical events overlapping in front of you in a diagram highlight the effect these would have on analysis .
- Using more advanced programming approaches to identify and display individual subjects can have benefits for all departments, data management, clinical and statistics and programming.
- With increased communication between departments and providing these visual tools created from a statistical standpoint to data management, we can help increase data quality and hence contribution towards endpoints.

5th Floor Regal House, 70 London Road, Twickenham TW1 3QS
+44 (0)20 3696 7240 | www.veramed.co.uk

THE ASSAY CAPABILITY TOOL (ACT): DRIVING THE ROBUSTNESS OF PRECLINICAL ASSAYS

Katrina Gore[1], Jason A. Miranda[2], Phil Stanley[1], Jamie Turner[2], Rebecca Dias[2], Huw Rees[2]

[1] Research Statistics, PharmaTherapeutics Clinical Research, Pfizer WRD; [2] Neuroscience & Pain Research Unit, Pfizer, Cambridge UK

Neuroscience & Pain
A Pfizer Research Unit

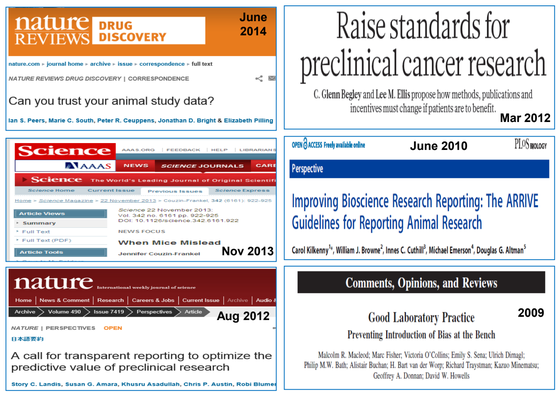
ABSTRACT

It is hard to pick up a recent copy of Nature, Science or many preclinical biomedical research journals without seeing an article on the issue of non-reproducible research. The pharmaceutical industry is not immune to these issues. Replication of published research findings is a key component of drug target identification and provides confidence to progress internal drug projects. Additionally, we use data from internally developed in vitro and in vivo assays to assess the biological and pharmacokinetic activity, selectivity and safety of novel compounds and make decisions which impact their progression towards nomination for clinical development.

This poster outlines steps Pfizer is already taking to improve the scientific rigour of experiments through the use of the Assay Capability Tool. The ACT promotes surprisingly basic but absolutely essential experimental design strategies and represents the distilled experience of the provision of over three decades of statistical support to laboratory scientists. It addresses the age old issue of statistical design, the more recently highlighted issue of bias and the hitherto overlooked issue of whether the assay actually meets the needs of a drug project team.

THE ASSAY CAPABILITY TOOL (ACT) – RATIONALE

- The Pharma Industry relies on externally published and internally generated data to provide confidence to initiate and progress internal drug projects.
- Many literature articles during the past decade have highlighted the need for improved preclinical research and the pace of publication is growing:



- "Sometimes the fundamentals get pushed aside – the basics of experimental design, the basics of statistics" Lawrence Tabak, Principal Deputy Director of the National Institutes of Health (US).
- Pfizer Research Statistics has worked for many years with scientists to increase the robustness of our preclinical research and the ACT is the result of that partnership.

THE ACT – WHAT IS IT

- A tool that promotes surprisingly basic but absolutely essential experimental design strategies; documents the strengths and weaknesses of an assay; and encourages the definition of what a successful assay outcome will look like.
- 13 item checklist assisting the scientist and statistician in designing fit for purpose preclinical assays / experiments.
- "Quality mark" facilitating informed use of assay results by decision makers, e.g. drug project teams, governance bodies.

ADDRESSING 3 KEY ASPECTS OF ASSAY DEVELOPMENT

1. Aligning Assay Capability with Project Objectives

Aligning Assay Capability with Project Objectives
(Does the assay enable a crisp decision?)

Key Considerations	Current Status / Recommendations to address gaps
Are the project team's scientific objectives for running the assay recorded in a protocol/SOP?	
Has the project team adequately pre-defined what a successful assay outcome looks like in order to guide decision making ?	
Is the experimental design described in the protocol/SOP and aligned closely with the objectives?	

- Have we defined a successful outcome in quantitative terms rather than just stating success is a statistically significant p-value?
- Is the study design tuned to the objectives, i.e. can it deliver what the project needs to make crisps decisions?

2. Enabling Assay Capability by Managing Variation

Enabling Assay Capability by Managing Variation
(Are we achieving required precision and using resources efficiently?)

Key Considerations	Current Status / Recommendations to address gaps
Are the assay's development and validation fully documented?	
Have the sources of variability present in the assay been explored?	
Is the proposed sample size/level of replication fit for purpose?	
Is there a comprehensive protocol /SOP detailing *key assay characteristics*?	
How is assay performance monitored over time? What is the plan for reacting to signs of instability?	

- Was the assay soundly developed and does it deliver consistent results?
- Have we identified sources of variability and removed/controlled them?
- What is the impact of variation on sample size and precision of results?
- Are the critical features of the assay defined in a comprehensive SOP/protocol?

3. Objectivity in Assay Conduct

Objectivity in Assay Conduct
(Are results likely to be reproducible?)

Key Considerations	Current Status / Recommendations to address gaps
Are inclusion/exclusion criteria for the assay specified in the protocol/SOP?	
Is the management of subjectivity in data collection and reporting defined in the assay protocol/SOP?	
If the raw data are processed (e.g. by summarization or normalization) prior to analysis, is the method for doing this specified in the study protocol/SOP?	
Are rules for treating data as outliers in the analysis specified in the protocol/SOP	
Is the analysis specified in the study protocol /SOP? Is it fit for purpose?	

- Has the potential for subjectivity in assay conduct, data handling and analysis been considered?
- Have techniques of randomization, blocking and blinding been used, where required, to prevent unintentional biases?

SUMMARISING THE THREE DOMAINS

The three domains are summarised by a low, medium or high grade to indicate the level of confidence in decision making a team can have when using data from the assay.

Assay Name Project:	Aligning Study Capability with Project Objectives	Enabling Assay Capability by Managing Variation	Objectivity in Assay Conduct
Confidence in Decision Making using data from this assay (Low/Medium/High)			
ACT Summary			

	Technical Specification
Target Value	
Required Precision	
Required Replication	

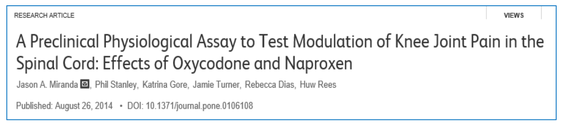
ACT IMPLEMENTATION & AWARENESS

- Involves a partnership between statisticians and scientists, with the aim that the tool is "owned" by scientists.
- Implementation has been tackled on many fronts:
  - Guidance documents and other supporting materials
  - Incorporation of the ACT into existing statistical training
  - Awareness presentations to scientists and project leaders
- 2014/2015 Research Statistics goals require the ACT in place for assays providing data to support project progression from the early stage of lead development through to drug candidate nomination.
  - Goal is also becoming part of Research Unit annual quality goals.

ACT CASE STUDY: MODULATION OF KNEE JOINT PAIN IN SPINAL CORD

- This is a novel *in vivo* spinal cord neurophysiological assay developed to test the efficacy of treatments for pain.
- Electrophysiological data is collected from single neurons in the spinal cord of deeply anaesthetised rats pre-sensitised with monoiodoacetate (MIA) while performing non-noxious and noxious knee joint rotation.
- **The Assay Capability Tool was used to guide experimental design, leading to a high quality and robust preclinical assay that won an internal 2014 3Rs award from the Pfizer Animal Care and Welfare Board.**

This poster focuses on the use of the ACT in the development of the assay. Full details of the assay methodology and the ACT can be found here:


RESEARCH ARTICLE VIEWS

A Preclinical Physiological Assay to Test Modulation of Knee Joint Pain in the Spinal Cord: Effects of Oxycodone and Naproxen

Jason A. Miranda, Phil Stanley, Katrina Gore, Jamie Turner, Rebecca Dias, Huw Rees

Published: August 26, 2014 • DOI: 10.1371/journal.pone.0106108

1. Aligning Assay Capability with Project Objectives

- **Project objectives**: the overall objective was to develop a high quality *in vivo* electrophysiology assay to confidently test novel compounds for efficacy against pain.
- **Defining success**: a structured approach to assay development was performed using known agents to validate the methodology and define target values for the effect size and precision.

2. Enabling Assay Capability by Managing Variation

- **Structured assay development:** a series of pilot and confirmatory experiments were run:
  - A small **pilot oxycodone experiment** was used to assess viability of the experimental approach
  - An **exploratory oxycodone experiment** aimed to identify the primary endpoint and understand the interplay between spike reduction and assay variation
  - A **follow-up naproxen experiment** was designed based on learnings from the previous experiments to test the prediction that COX-1/COX-2 inhibition reduces the primary endpoint of tonic spiking activity in response to noxious joint rotation
- **Identifying variability:** appropriate design of the individual experiments allowed sources of variability to be identified and their impact quantified.
- **Minimising variability:** detailed protocols describing the experimental timelines and procedures were developed to minimise and control future experimental variation.
- **Sample sizing:** sample size calculations were performed after the oxycodone and naproxen experiments to ensure the appropriate number of animals were used to meet each study objective and to guide the sample size for future drug studies**.**
- **Quality Control (QC) Charts:** oxycodone or naproxen will be used in future experiments as a positive control and QC charts created to assess stability of response over time.

3. Objectivity in Assay Conduct

- **Inclusion/exclusion criteria:** these were developed for individual cells based on recording quality, animal health and joint rotation and were documented in the experimental protocol.
- **Randomisation & Blinding:** to ensure the scientist remained unaware of the treatment an animal received and prevent unintentional biases, random allocation to experiment groups, allocation concealment and blinded outcome assessment were implemented and documented.
- **Blocking:** each individual study was run in smaller separate blocks to prevent the introduction of bias from changing conditions over time. A futility analysis was performed halfway through the study to prevent unnecessary subsequent animal usage.
- **Data processing & statistical analysis:** any data processing methods were documented so they could be reproduced and a statistical analysis appropriate for the design, e.g. including baseline and block information, was used.

ASSAY CAPABILITY TOOL – THE BOTTOM LINE

The ACT was developed to **guide the early development of assays** and to **assess their capability to generate reliable data**.

- It ensures good statistical design and analysis is embedded into the already established good scientific practices.
- It explicitly addresses the question that is often implicitly taken for granted: "Does the assay actually meet the needs of the project?".

We believe the ACT is a practical step forward in improving the reproducibility of preclinical research and is central to Pfizer's continued drive to embed statistical design and analysis into all of our research.

ACKNOWLEDGEMENTS

We'd like to thank Phil Woodward, Global VP PTx Research Statistics, for his development contributions and his ongoing support for the ACT.

Assessment of various continual reassessment method models for dose-escalation phase 1 oncology clinical trials: AZD3514 data and simulation studies

Gareth James[1], Stefan Symeonides[2], Jayne Marshall[2], Julia Young[2], Simon Smith[2], Glen Clack[2]

[1] PHASTAR House, 2 Heathfield Terrace, London W4 4JE, UK [2] Alderley Park, Macclesfield, Cheshire, SK10 4TF, UK [3]Edinburgh Cancer Centre, Western General Hospital, Edinburgh EH4 2XU

Introduction

The continual reassessment method[1] (CRM) is considered more efficient and ethical than standard methods for dose-escalation trials in oncology, but requires an underlying estimate of the dose-toxicity relationship ("skeleton"). Previously we conducted post-hoc dose-escalation analyses on real-life clinical trial data from an early oncology drug (AZD3514) using the 3+3 method and CRM using six different skeletons; we found each CRM model outperformed the 3+3 method by reducing the number of patients allocated to suboptimal and toxic doses. The CRM models with conservative and sigmoidal skeletons were the most successful.

Aim

To compare the CRM with different skeletons and the 3+3 method in their ability to determine the true maximum tolerated dose (MTD) of various "true" dose-toxicity relationships. .

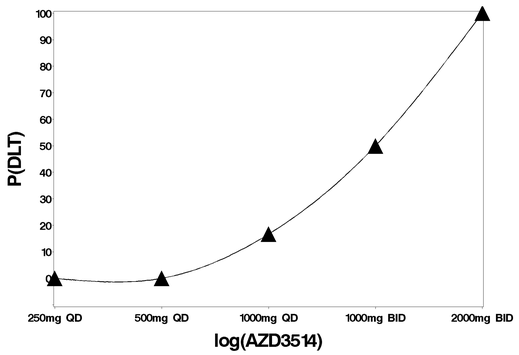
Methods

We considered seven true dose-toxicity relationships, one based on AZD3514 data and six theoretical with the true MTDs identified as the highest dose where the probability of suffering a DLT is below 33%. For each true dose-toxicity relationship 1000 simulations were conducted and the MTD was identified using the 3+3 method, one-parameter logistic CRM (CRM 1PL) with six skeletons and two parameter logistic model (CRM 2PL) as proposed by Neuenschwander[2]. The CRM will begin after the first patient has experienced a DLT, as proposed by Goodman[3].

AZD3514 data

- Patients had metastatic castrate resistant prostate cancer
- For each dose the P(DLT) was deduced using the first six patients assigned a dose below 2000mg QD who received this dose for at least 28 days and all 4 patients who received 2000mg BID.

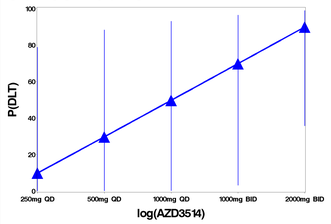
DLT curve



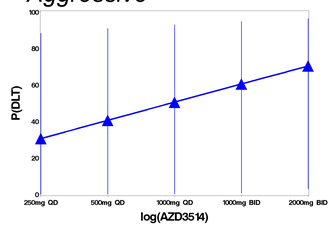
Skeletons and 95% prediction intervals

The skeletons were used in the CRM 1PL models and were used as theoretical true dose-toxicity curves.

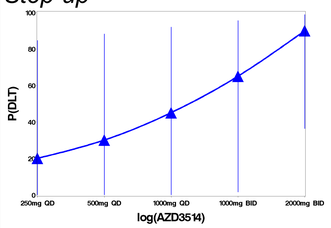
Conservative



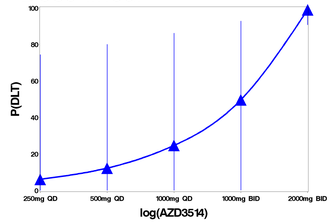
Aggressive



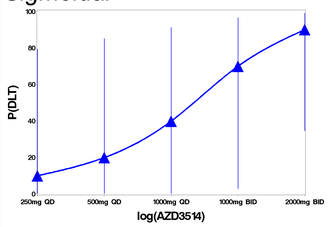
Step-up



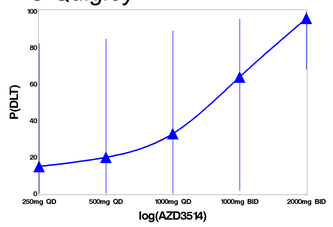
Dose-linear



Sigmoidal



O'Quigley



Results

Percentage of simulations which identified the true MTD

Dose-escalation method		True dose-toxicity curve: Cons	Aggr	Step	Dose	Sigm	O'Qu	AZD3
CRM logistic one-parameter	Conservative	63	44	44	57	50	32	74
	Aggressive	45	49	31	53	34	38	38
	Step-up	45	49	30	64	33	43	58
	Dose-Linear	40	45	24	63	23	51	36
	Sigmoidal	52	45	33	68	33	46	73
	O'Quigley	40	49	27	66	29	44	58
CRM logistic two-parameter		66	27	45	28	77	10	71
3+3	-	33	32	23	37	47	18	65

Key: Green: >70%, Blue: 50-70%, Red: <25%, Pink: skeleton=true curve

- All CRM 1PL models outperformed the 3+3 for most true dose-toxicity curves. The ability of the CRM 2PL model varied considerably depending on true dose-toxicity curve.
- For the CRM 1PL model, the skeleton has considerable influence on the model's ability to identify the true MTD.
- Starting with an accurate estimate of the dose-toxicity curve does not guarantee the best results.

Dose escalation method summary

Dose escalation method		MTD selected by dose-escalation method (%): True selected	Under-estimated	Over-estimated	Could not determine
CRM logistic one-parameter	Conservative	52	24	21	7
	Sigmoidal	50	20	26	7
	Step-up	46	22	28	7
	O'Quigley	45	22	30	7
	Aggressive	40	22	34	7
	Dose-Linear	40	18	38	6
CRM logistic two-parameter		46	39	4	16
3+3	-	36	43	7	20

Key: Green: best performing method, Red: worst performing method

- CRM 1PL and 2PL models had similar ability to identify the true MTD, and both outperformed the 3+3 method.
- Compared to CRM 1PL, the 3+3 and CRM 2PL models were twice as likely to underestimate the MTD, but considerably less likely to overestimate the MTD.
- For the CRM 1PL, the conservative and sigmoidal skeletons were the most successful at identifying the true MTD and they overestimated the MTD in a lower proportion of simulations.

Discussion

- The ability of the CRM model to identify the true MTD would be increased by increasing the number of patients.
- Constraints could be added to the CRM 1PL model to minimize overestimating of the MTD, but this may influence its ability to identify the true MTD.

Conclusion

- The CRM generally outperformed the 3 + 3 method for the clinical and simulated data.
- The conservative and sigmoidal were the optimal skeletons for the CRM 1PL model on real clinical and simulated data.
- Clinical opinion should also be used in decisions to recommend the next dose.
- Further work is needed to determine the optimum combination of dose-toxicity model and skeleton.

For more information please email: **gareth,james@phastar.co.uk**

[1] O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. Biometrics. 1990;46:33–48.
[2] Neuenschwander, Beat et al. (2008). \Critical aspects of the Bayesian approach to phase I cancer trials". In: Statistics in Medicine 27.13, pp. 2420{2439. issn: 1097-0258.
[3] Study design principles in the clinical evaluation of new drugs as developed by the chemotherapy programme of the National Cancer Institute. Carter SK. In The Design of Clinical Trials in Cancer Therapy, Staquet MJ (ed) Editions Scientifique Brussels 1973: 242-289.

Comparing methods for handling missing glycated haemoglobin (HbA1c) values in clinical trials on patients with Type II diabetes.

Sophie Lee[1], Tina Rupnik[1], Gareth James[1]
[1] Unit 2a 2a Bollo Lane, London, W4 5LE

Introduction

Context
Glycated haemoglobin (HbA1c) is the most commonly used measure of severity in Type II diabetes. In clinical trials, HbA1c is usually measured longitudinally, but levels of HbA1c can vary considerably between patients and within patients across time.

Issues
Due to the variability of the change from baseline HbA1c measure, simple missing data methods such as ignoring missing data and carrying data forward from a previous measurement may increase the bias and reduce precision of estimates of treatment effects.

Aim
Compare four missing data methods to investigate their ability to estimate the ethnicity effects in terms of bias and precision.

Background

Data
We used primary care longitudinal data measures of HbA1c for 6104 patients with Type II diabetes on a yearly basis from two inner London boroughs between 2007 and 2009. This data contained a number of healthcare measurements such as BMI and serum cholesterol. Only patients with complete healthcare data were included in the analysis, leaving 5264 patients.

Motivation
Previous analysis on this data found that, of the three ethnic groups investigated, south Asian people had less improvement in HbA1c than white or black African/Caribbean people[1]. We chose to investigate the ethnicity effect rather than treatment as this is what the publication did. In particular, we were interested in how well the two-fold fully conditional specification algorithm (FCS) imputed values as this method has performed well in epidemiology but has yet to be applied in clinical trials[2].

Methods

➢Using patients with complete data we reproduced results from recently published manuscript. We fitted a linear multilevel model to estimate change from baseline in HbA1c in 2008 and 2009.
➢The model was adjusted for available healthcare measurements.
➢Three levels were used instead of the four in the manuscript: year of measurement, patient and practice.

➢Approximately 30% of the data for HbA1c measures was set to missing using a missing completely at random mechanism.
➢We used four methods to handle the missing HbA1c measures and generated a multilevel model for each of the datasets:

Complete case analysis (CC) – only analyse patients for which all measurements are recorded.

Last observation carried forward (LOCF)– replace missing measures with the last observed value from that patient.

Multiple imputation (MI) – impute each missing measure with a set of plausible values drawn from a conditional distribution based on other available variables at any time point, analyse these sets separately and combine the results.

Simplified two-fold FCS algorithm – impute missing values using multiple imputation based on other variables from the same time point or immediately adjacent time points only.

➢Each method's ability to estimate the true ethnicity effect based on the complete data was compared by calculating the bias and precision.

Results

Table 1: Comparison of ethnicity effects (white vs black African/Caribbean)

Model	Ethnicity effect	Bias	Precision
True model (full)	-0.099	-----	288.6
Complete case	-0.133	0.034	165.8
Last observation carried forward	-0.155	0.056	221.9
Multiple imputation	-0.110	-0.011	194.0
Simplified two-fold FCS algorithm	-0.049	-0.050	198.9

➢MI outperformed other methods based on bias.
➢ All methods underestimated precision.

Table 2: Comparison of ethnicity effects (white vs south Asian)

Model	Ethnicity effect	Bias	Precision
True model (full)	0.126	-----	397.5
Complete case	0.163	-0.037	222.7
Last observation carried forward	0.138	-0.012	304.5
Multiple imputation	0.139	-0.013	291.2
Simplified two-fold FCS algorithm	0.122	0.004	184.8

➢Two-fold method performed best in terms of bias.
➢All methods outperformed CC in terms of bias. LOCF and MI produced a similar bias.
➢All methods underestimated precision.

Table 3: Differences between true change from baseline HbA1c values and imputed values

	LOCF	MI 1	MI 2	MI 3	MI 4	MI 5	Two-fold 1	Two-fold 2	Two-fold 3
Mean difference	-0.394	-0.009	-0.009	0.042	-0.108	-0.053	0.034	-0.021	-0.041
Standard deviation	1.546	2.043	2.055	2.041	2.025	2.047	2.081	2.091	2.064

➢LOCF has biggest difference between true and imputed values and lowest standard deviation.
➢MI and simplified two-fold methods produce small mean differences between true and imputed values.

Discussion

➢Further work is needed to determine the two-fold FCS algorithm's ability to estimate treatment effects in a clinical setting using data with more time points. The data used in our research only contained 3 time points but the algorithm was developed to reduce problems due to overfitting when numerous time points and covariates are included in a model.

➢This research could be extended by running the full two-fold FCS algorithm on the data and comparing results with the simplified method we used. Other missingness mechanisms could also be investigated.

➢ Table 3 shows that the mean difference between true and estimated values of change from baseline HbA1c measures calculated using the LOCF method is larger than MI and the simplified two-fold method. This method also produced a lower SD, reducing the variability of estimates. This implies that, although the models found using the LOCF method had a low bias, this method may not be as robust as MI and the simplified two-fold approaches.

[1] James GD, Baker P, Badrick E, Rohini M, Hull S, Robson J. Type 2 diabetes: a cohort study of treatment, ethnic and social group influences on glycated haemoglobin. *BMJ open* 2.5 (2012): e001477.. 1990;46:33–48.
[2] Welch C, Barlett J, Peterson I. Application of multiple imputation using the two-fold fully conditional specification algorithm in longitudinal clinical data. Statistics in medicine 33.21 (2014): 3725.

Bolstering with Bayes:

A framework for interpreting the risk of rare adverse events in the presence of limited clinical trial data

MedImmune
A member of the AstraZeneca Group

11 May 2015

Introduction

When a new drug completes phase II trials, drug development teams are primarily interested in assessing the risk-benefit profile of a new drug, in accordance with the Target Product Profile (TPP). There may be interest in evaluating the risk of an adverse event of special interest (AESI). For rare adverse events, limited data are available, and making inferences about relative increase in risk using traditional statistical methods becomes challenging, especially if zero events are reported in one or more treatment groups. As part of a Safety GO/NO-GO project, methods were explored for quantifying the risk of rare AESIs that could be used in decision making. In this poster, we present a Bayesian approach utilising informative conjugate priors as described by Kerman[1]. A case study will illustrate the application of the method.

Methodology

Kerman Priors for the Poisson-Gamma model

Consider a Poisson-Gamma model for the AESI data. Let y = number of reported AESI and X=follow up time. Then

$$y \sim \text{Poisson}(\lambda X) \quad \text{where } \lambda \sim \text{Gamma}(c,d)$$

Kerman[1] defines a set of priors for λ of the form

$$\text{Gamma}(1/3 + ky,\ kX)$$

Under these priors, the implied prior rate is y/X. k is a scale factor (taking a value between 0 and 1) which can be used to down-weight the influence of the prior, e.g. if the prior information comes from 1000 yrs of follow up, using k=0.1 would down-weight this to be worth 100 yrs. If the assumed prior follow up time is the same as that in the observed data, k= 1 gives the prior and the observed data equal weight and k=0.5 gives the prior half the weight of the data. The prior Gamma(1/3,0) is considered a neutral prior in the sense that it gives posterior median distributions that are approximately equal to the observed y/X, as long as y is not 0.

Using Kerman priors in the analysis of reported AESI data from a drug development program can enable posterior probability statements to be made about the likely incidence of the AESI for the new treatment. Further, a decision framework can be constructed based on the posterior probability for incidence rate ratio (IRR) (active/ placebo) of the AESI being greater or less than a pre-specified value. Where relevant and reliable historical data for the AESI exist, these can be used as values for y and X in setting the Kerman prior. Alternatively, the prior may be elicited.

Simulation Study – Evaluating the method

A simulation study was run to evaluate different Kerman priors for a range of true underlying data scenarios. Some possible decision rules were also included to demonstrate the utility of the method. In both the prior scenarios and the underlying true data scenarios, it is assumed that the AESI is rare. Use of so-called 'non-informative' priors would not be appropriate in such cases.

The following simulation parameters were investigated:

- True data scenarios: 10 different scenarios considered, with underlying placebo and active rates varying from 0.5 to 5 per 100 patient years
- Priors: 14 different priors considered with rates varying from 0.5 to 5 per 100 patient years and weighting with k=1 (equal weight to data), k=0.5 (half the weight of the data) and k=0.1 (a tenth the weight of the data).
- Decision thresholds: IRR<1, IRR<1.5, IRR >2, IRR >3.

For each scenario, the posterior summaries of the distributions for for the active and placebo treatments and for the incidence rate ratio were obtained and averaged over 1000 simulations. Full details of the simulations were pre-defined in a simulation plan.

Simulation Study – Results

Review of the posterior summaries for the 1000 simulations for each scenario showed that whilst the analyses gave consistent results in the majority of cases, some analyses produced posterior means for the IRR that were extreme. The posterior medians were more stable, although still with some extreme results. The influence of the prior was apparent, with many cases where the posterior medians were somewhat different from the true underlying data scenario from which the data were simulated.

Nonetheless, looking at the posterior summaries averaged over the 1000 simulations enabled some assessment of the operating characteristics of various decision rules for different scenarios. One such scenario is presented in Table 1. Here, rules for stopping development are considered. A good rule will lead to a decision to stop development if the relative incidence of the AESI is high on active compared to placebo, and will not lead to a stop decision if the incidence is the same in both groups.

Table 1: Operating characteristics for one simulation scenario.

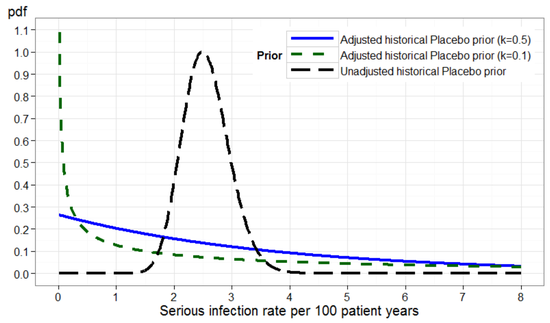
True Underlying Scenario (events per 100 patient years) Placebo, Active	Desired Outcome	Probability of a STOP decision (STOP if P(IRR< **1.5**) is less than **0.2**)	
		Prior based on 2 AESIs per 100 patient years weighted with	
		k=0.5	k=0.1
0.5, 0.5	GO	0.01	0.07
2,2	GO	0.06	0.11
5,5	GO	0.08	0.11
0.5, 1	PROBABLE GO	0.04	0.15
1,2	PROBABLE GO	0.13	0.26
2,5	BORDERLINE	0.35	0.44
0.5, 2	STOP	0.23	0.42
1,5	STOP	0.56	0.67
0.5, 5	STOP	0.70	0.82

We see that for a prior based on 2 AESI per 100 patient years, for a study with 100 patient years exposure in each group, then using the decision rule 'STOP if the P(IRR<1.5) <0.2 has reasonable operating characteristics. In the case where the true incidence is 10 fold higher on active than placebo, a STOP decision would be made 82% of the time. When the true incidence is the same in both groups, a STOP decision is made less than 15% of the time.

Case Study – Mavrilimumab

A phase II program in rheumatoid arthritis recently completed for mavrilimumab, a fully-human monoclonal antibody targeting the alpha subunit of GM-CSFR. The occurrence of serious infections was of interest since they are known to be associated with use of biologics[2]. Across two placebo-controlled phase 2 studies, 2 serious infections were reported in 439 patients treated with mavrilimumab, with a total exposure of 173 patient-years, an incidence of 1.16 cases per 100 patient-years. No serious infections were reported in 177 patients treated with placebo, with a total exposure of 53 patient-years[3,4].

Figure 1: Kerman Priors for serious infection rate based on historical data



Data from programs of similar competitor products in the same indication are available, with incidence of serious infections in the range of 3 to 5.3 cases per 100 patient-years exposure. A meta-analysis of approximately 1600 patient-years of placebo data was performed. The resulting placebo serious infections rate was 2.5 cases for 100 patient-years exposure. This placebo rate was used to construct a Kerman prior for analysis of the mavrilimumab data.

After appropriate adjustments to account for differing exposure, the mavrilimumab data were analysed using the Kerman priors, with k=0.5 (prior carries half the weight of the data) and k=0.1 (prior carries a tenth the weight of the data). The priors are illustrated in Figure 1, and the results are presented in Table 2.

The mavrilimumab case study is retrospective, no decision rules were agreed a priori for the incidence of serious infections. However, it is unlikely that a decision rule to stop development would have been met; the posterior median for mavrilimumab is lower than the range from competitor products, and the posterior distribution for the IRR is widely spread, as we might anticipate with a small amount of available data.

Table 2: Posterior summaries for mavrilimumab data under different priors

Prior	Node	Posterior Percentiles			P(IRR<x) where		
		10%	50%	90%	x=1	x=1.5	x=3
Historical placebo prior, k=0.5	Mavrilimumab	0.80	1.61	2.79	n/a		
	Placebo	0.14	0.88	2.99			
	IRR	0.44	1.81	12.02	0.31	0.44	0.65
Historical placebo prior, k=0.1	Mavrilimumab	0.51	1.30	2.65	n/a		
	Placebo	0.01	0.33	2.30			
	IRR	0.45	3.99	247.11	0.23	0.31	0.45

Concluding Remarks

- Kerman priors, based on historical data or elicited opinion, can be used in a Bayesian analysis of a rare AESI, from which interpretations can be made about the probability of the true incidence of the AESI.
- The method has limitations. An appropriate weighting should be given to the prior distribution. Too great a weight and the posterior will essentially just reflect the prior data. Too little weight and the posterior will be very spread and interpretations will not be very informative.
- The method will not add value in situations where there is little safety follow-up data available, or where there are no events in any treatment group.
- The methods described do, however, permit a decision making framework for a rare AESI to be established prior to the conduct of the study, and the posterior distributions can aid in interpreting the data. When setting decision rules before the study, consideration should be given to the amount of follow up and simulation should be used to assess the appropriateness of decision rules based on the study design and the TPP.

References

[1]Kerman, J (2011) 'Neutral Non-Informative and informative conjugate beta and gamma prior distributions', *Electronic Journal of Statistics.*
[2]Singh, J et al (2011) 'Adverse effects of biologics: A network meta-analysis and Cochrane overview' *The Cochrane Collaboration* Wiley
[3]Takeuchi, T et al (2015) 'Efficacy and safety of mavrilimumab in Japanese subjects with rheumatoid arthritis:Findings from a phase IIa study' *Modern Rheumatology* Vol 25(1) pp21-30.
[4]Burmester et al (2014) 'Efficacy and Safety/Tolerability of mavrilimumab, a human GM-CSFRα monoclonal antibody in patients with rheumatoid arthritis' *Meeting abstract, American college of rheumatology*

Authors

Rachel Moate[1], Alex Godwood[1], Jianliang Zhang[2]

Contact email: moater@medimmune.com

[1]Clinical Biostatistics and Data Management, MedImmune, Cambridge, UK.
[2]Clinical Biostatistics and Data Management, MedImmune, Gaithersburg, US.

‘Hot deck’ Imputation: Determining a nonparametric statistical model for the distribution of missing data and its application in a Rate of Decline Analysis.

Amy Newlands[1], Abigail Fuller[2]

1. GlaxoSmithKline, Clinical Statistics, Respiratory, Stockley Park
2. Veramed Ltd, Statistics, Twickenham

[1]GlaxoSmithKline, Clinical Statistics, Respiratory, Stockley Park; [2]Veramed Ltd, Statistics, Twickenham;

Introduction

- Missing data is a serious problem in clinical trials and may compromise the validity of treatment comparisons because “missingness“ may be related to the drug’s effectiveness, safety or patient prognosis.
- There are three types of missing data to consider: Missing Completely at Random (MCAR); Missing at Random (MAR) and Missing Not at Random (MNAR). For the mechanism of ‘missingness’ to be considered MCAR, the probability of the data being missing is unrelated to both observed and unobserved data. An example would be a researcher removing a randomly selected sample from the data. For the mechanism to be MAR, the behaviour of the ‘missingness’ is related to values of observed data but not unobserved data and for the mechanism to be MNAR the data is missing for a specific reason and is related to unobserved data.
- Although best efforts are taken to minimise missing data, missing values are inevitable and how this is accounted for in the analysis is an important consideration. There are several existing methods for handling missing data, however, all methods rely on assumptions that cannot be verified.
- ‘Hot-deck’, or non-parametric imputation is one approach. Non-parametric imputation is appealing when dealing with a large dataset of subjects since parametric imputation (i.e. choosing, estimating and imputing from a parametric statistical model for the distribution of the missing data given the observed data) can prove challenging.

Background

- SUMMIT is an on-going large clinical outcomes study comparing the effect of the once daily ICS/LABA combination Flucticasone Furoate/Vilanterol Inhalation powder 100/25mcg with placebo on the survival in subjects with moderate COPD (≥50 and ≤70 % predicted FEV1) and a history of or at risk for cardiovascular disease.
- One of the secondary endpoints is the rate of decline of on-treatment post-bronchodilator FEV1. When considering an on-treatment analysis, there are two missing data situations, data is missing due to a subject missing a visit or data is missing due to a subject withdrawing from treatment. The latter of these is considered here. As the ’missingness’ of data in this endpoint is related to withdrawal, a MAR mechanism is assumed, and covariates relating to withdrawal are considered for the ‘hot-decking’ method.
- For the main analysis of this endpoint, a random coefficients model is used. However, this approach gives more weight to subjects with more data; hence a sensitivity analysis of individual regression slopes is performed. To perform this individual slopes analysis, every subject with at least two on-treatment FEV1 values has their on-treatment slope calculated using linear regression. (See Table 1). These slopes are then analysed using an analysis of covariance.

Table 1. Example of Missing Data for FEV1

Subject No.	Column heading 2	Column heading 3	Column heading 4		
1	1540	1550	1540	X	→ Slope not missing
2	1700	1750	X	X	→ Slope missing
3	2150	2300	X	2180	→ Slope not missing

- To provide an alternative assessment of treatment efficacy assuming a MAR mechanism, a supporting analysis using a ‘hot-deck’ imputation approach is proposed.
- As SUMMIT is an on-going study, and unblinding has not yet occurred, a dummy randomisation has been applied.
- One of the secondary endpoints is the rate of decline of on-treatment post-bronchodilator FEV1. When considering an on-treatment analysis, there are two missing data situations, data is missing due to a subject missing a visit or data is missing due to a subject withdrawing from treatment. The latter of these is considered here. As the ’missingness’ of data in this endpoint is related to withdrawal, a MAR mechanism is assumed, and covariates relating to withdrawal are considered for the ‘hot-decking’ method.

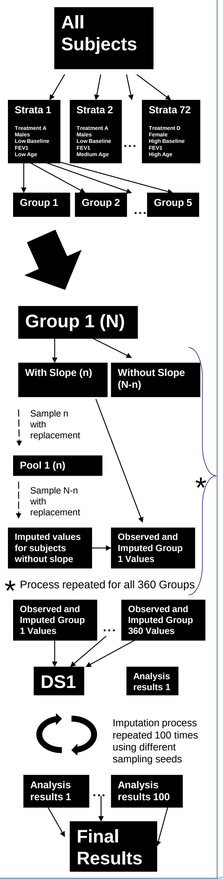
Hot-Deck Imputation Methods

A slope was calculated for all those subjects in the ITT population by regressing FEV1 on time. For those subjects who do not have at least 2 post-baseline measurements, ‘Hot-deck’ imputation has been carried out in the following steps:

1. Subjects were stratified by treatment, gender, age (3 levels) and baseline FEV1(3 levels) (i.e. 72 different strata, e.g. Treatment A, Males, Low Age and Low Baseline FEV1 or Treatment D, Females, Medium Age, High Baseline FEV1). The 3 categories for age and baseline FEV1 were defined by splitting the data into tertiles. As some of the strata did not have enough non-missing data when all data was considered, the tertiles were defined within gender groups. The average of the individual regression slopes will vary between the defined stratum.
2. A propensity score was calculated by fitting a logistic regression model to all subjects including covariates that were expected to be related to withdrawal from treatment. The presence or absence of a slope was fitted as the response with covariates expected to affect withdrawal like BMI, previous exacerbation history, etc

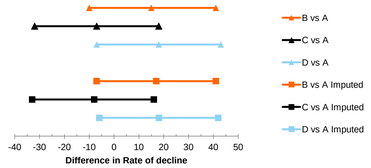
Figure 1. Hot-Deck Imputation Methods

3. To illustrate how the propensity score is used, let us take an example of Strata 1 with say 300 subjects. Each subject in that strata has had a propensity score calculated; the subjects are then ordered and split into 5 groups of N subjects (20% in each). This is done for each of the other 71 strata, giving 360 group-strata combinations, each with approx N = 60 subjects (strata may vary in size, because there are fewer females than males).
4. Within each group-strata combination, subjects with a slope were selected. Again using the example of group-strata 1, a donor pool, Pool(1) of size n (subjects with a slope, say 40) is drawn at random with replacement from the n subjects in this group-stratum with slopes. To determine the slopes of those subjects without a slope (N-n, say 20 subjects), the number of subjects missing a slope is sampled at random with replacement from Pool(1). This process is repeated for each of the remaining 359 group-stratum combinations.
5. The observed and imputed slope data from all 360 group-strata combinations are combined to form an imputed dataset (DS(1)).This procedure is then repeated 100 times to produce a series of 100 imputed datasets DS(1), DS(2),...., DS(100). The double re-sampling (i.e. sampling with replacement from the subjects with slopes to form a donor pool and then for each subject that has a missing slope, drawing a donor from the donor pool with replacement) approximates the Bayesian bootstrap Rubin (1981).
6. The analysis of individual slopes is performed on each of the 100 imputed datasets to give 100 sets of results. These are then pooled and one set of results obtained using PROC MIANALYZE in SAS..



Results

Figure 2. Plot of Treatment differences with 95% CI for two methods – No imputation & Imputation applied



Using dummy treatment groups we compared the analysis using imputed data to that using observed data. The results were very similar. However, the imputation shows marginally narrower confidence intervals as the variation has been slightly reduced as we would expect given the number of subjects contributing to the analysis has increased.

Conclusions

- In this run, the sensitivity method is robust, the results were very similar between the two methods. However, due to dummy randomisation being applied we do not know the full impact of missing slopes data being imputed.
- The hot-deck imputation method is easy to apply to the SUMMIT database and provides an alternative method when using other approaches in large datasets become difficult when a MAR mechanism is assumed.

References

1. James Carpenter, Jonathan Bartlett, and Mike Kenward, London School of Hygiene and Tropical Medicine www.missingdata.org.uk.
2. James R Carpenter and Michael G. Kenward, WILEY, ISBN: 978-0-470-74052-1 Multiple Imputation and its Application
3. Rubin (1981) The Bayesian Bootstrap *The Annals of Statistics* **9** 130 – 134

Acknowledgements

- This study was funded by GSK (HZC113782)
- James Roger for providing references to the method and invaluable advice when implementing the method
- Colleagues in GSK clinical statistics

Count data: should we bother with Poisson or zero-inflated models?

Anna Rigazio, Audrone Aksomaityte, Michael Fitzgerald, Pinal Patel

[1] PHASTAR, Unit 2, 2A Bollo Lane Chiswick (London) United Kingdom W4 5LE

Introduction

In asthma clinical trials, Poisson regression is frequently used to analyse exacerbation rates, assuming that the mean occurrence rate of the event is equal to its variance. However, asthma exacerbation data are often characterised by over-dispersion and frequent zero-count observations. Thus, a Poisson regression might fit these data poorly and other generalised linear models could perform better. When the variance is higher than the mean event rate, a negative binomial (NB) regression model should be preferable. Zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models are also used to avoid the underestimation of rates of excess zero-count events.

Aim

The aim of this work was to investigate the performance of a Poisson, NB, ZIP and ZINB regression models on data characterised by over-dispersion and zero-inflation.

Methods

Data

We simulated exacerbations data from Poisson, NB, ZIP and ZINB distributions using different parameter values (Table 1) with ranges that are relevant in asthma (e.g., 0-2 expected exacerbations per year).

Table 1: Parameter selection

	Distribution			
	Poisson (λ)	NB (k, $\frac{\lambda}{\lambda+k}$)	ZIP (λ, p_{zero})	ZINB(k, $\frac{\lambda}{\lambda+k}$, p_{zero})
λ	0.5, 1, 2	0.5, 1, 2	0.5, 1, 2	0.5, 1, 2
k	NA	1, 2	NA	1, 2
p_{zero}	NA	NA	0.2, 0.5, 0.8	0.2, 0.5, 0.8

λ - expected number of events per year, k – shape or over-dispersion parameter, p_{zero}– probability used for generating zeros in ZIP and ZINB.

We used the common example of asthma exacerbations to inform the parameter space for the count data. However, this can be extended to any count data.

For each distribution we simulated 100 samples with 100 subjects in each sample. Subjects were randomly allocated to treatment A or B.

Statistical analysis

For every type of simulated data we applied four regression models: Poisson, NB, ZIP and ZINB (proc genmod in SAS). We compared the fit of the models in terms of Akaike Information Criterion (AIC – the lowest the best) and by plotting the average probability distribution and simulated frequency of exacerbations.

Results

Our findings are displayed in the Table 2 and Figure 1.

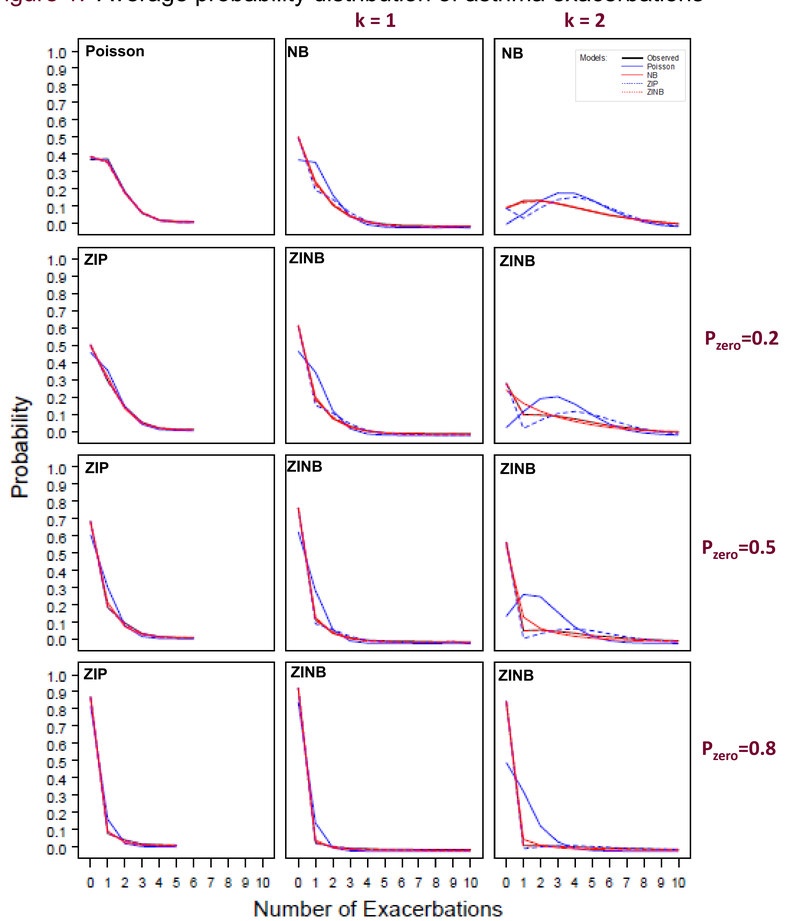
- All four models approximated Poisson and ZIP data well.
- Poisson regression model performed best only with Poisson data, whilst its performance on all other data (mainly in presence of high variances such as the NB and ZINB distributions) was the poorest.
- On the contrary, NB model provided the best or second best fit to all of the data.
- ZINB performance was similar to NB for data with higher variances. However, the model did not approximate well Poisson and ZIP data. In such instances the ZIP model provided a better fit.

Table 2: Summary statistics and model AIC for various parameters

	Parameters			Summary statistics			Model AIC			
Data	λ	k	p_{zero}	Mean	Var	% of zeros	Poisson	NB	ZIP	ZINB
Poisson	0.5	NA	NA	0.5	0.51	60	189.16	191	190.9	193
	1	NA	NA	1	0.96	40	260.37	262.2	262.1	264.2
	2	NA	NA	2.0	1.98	10	342.66	344.4	344.4	346.4
ZIP	0.5	NA	0.2	0.4	0.44	70	168.45	169.5	169.4	171.4
	0.5	NA	0.5	0.3	0.32	80	130.42	130.1	129.8	131.8
	0.5	NA	0.8	0.1	0.13	90	68.67	67.5	66.93	69.08
	1	NA	0.2	0.8	0.96	50	245.42	245	244.3	246.3
	1	NA	0.5	0.5	0.77	70	203.1	195.9	194.1	196.1
	1	NA	0.8	0.2	0.37	90	117.7	107.5	106.1	108.1
	2	NA	0.2	1.6	2.23	30	344.91	339.9	335.1	337.1
	2	NA	0.5	1.0	1.98	60	307.91	277.9	268.8	270.7
	2	NA	0.8	0.4	1.03	80	193.09	150.5	145	146.9
NB	0.5	1	NA	2.0	5.79	30	449.04	386.3	410.2	387.6
	0.5	2	NA	8.0	40.71	0	825.14	617	795.3	618.6
	1	1	NA	1.0	2.02	50	302.31	281	286.1	282.3
	1	2	NA	1.0	2.02	50	302.31	281	286.1	282.3
	2	1	NA	0.5	0.79	70	202.67	196.5	197.3	198.2
	2	2	NA	2.0	4.06	30	404.61	378.8	391.1	380.4
ZINB	0.5	1	0.2	1.6	5.56	50	433.96	345.1	365.6	345.7
	0.5	1	0.5	1.0	4.01	70	363	259	268.4	257.9
	0.5	1	0.8	0.4	2.07	90	222.83	134.2	137.1	133.3
	0.5	2	0.2	6.5	42.39	20	942.12	591	715.8	578.8
	0.5	2	0.5	4.0	35.37	50	948.97	454.4	517.6	435.9
	0.5	2	0.8	1.5	18.55	80	652.27	219.5	239.3	209.4
	1	1	0.2	0.8	1.72	60	273.25	245.9	250	247
	1	1	0.5	0.5	1.24	70	217.98	185.2	187.2	186
	1	1	0.8	0.2	0.55	90	122.67	95.98	94.71	96.06
	1	2	0.2	3.2	12.19	30	597.88	466.7	499.7	462.5
	1	2	0.5	2.0	9.92	60	556.12	359.1	371.6	350.3
	1	2	0.8	0.8	4.58	80	348.47	181.5	182.8	176.7
	2	1	0.2	0.4	0.64	70	178.28	171.2	171.5	172.7
	2	1	0.5	0.2	0.45	80	133.88	123.8	123.4	125
	2	1	0.8	0.1	0.19	90	70.9	63.87	63.08	65.06
	2	2	0.2	1.6	3.97	40	394.98	348.5	357.9	348.6
	2	2	0.5	1.0	3.05	60	341.05	274.5	277	272.8
	2	2	0.8	0.4	1.45	80	207.74	144.1	143	143

Summary statistics and model AIC for various parameters in all four types of data. Best model fit corresponds to the lowest AIC coloured in green, worst (highest AIC) – in red.

Figure 1: Average probability distribution of asthma exacerbations



Simulations for λ=0.5. Similar results were obtained for λ values 1 and 2 and are not reported here.

Discussion

In asthma clinical trials, Poisson regression is frequently used to analyse exacerbation rates. However, in our study, Poisson model only worked well with strictly Poisson data and in other cases provided the worse fit.

Our data suggest that NB always provides the best or second best fit for both over-dispersed and non-over-dispersed count data, even in the presence of high zero inflation.

If the nature of the experiment strongly suggests that the data come from a specific model then this should be considered first.

In the future we would like to consider additional parameter values in order to identify potential thresholds to use as guidance in the choice of the best fitting model. Furthermore, the correlation between these parameters should be investigated in more detail.

[1]Donald Erdman, *et al.* SAS Institute Inc., Cary, NC. Zero-Inflated Poisson and Zero-Inflated Negative Binomial Models Using the COUNTREG Procedure.
[2]Paul Allison, Do We Really Need Zero-Inflated Models? August 7, 2012 http://statisticalhorizons.com/zero-inflated-models

Statistical models for *de facto* estimands - beyond sensitivity analysis.

James Roger and Michael Kenward

London School of Hygiene and Tropical Medicine.

Premise

1. A well defined *estimand* is critical for handling data from patients who withdraw early.
   - The quantitative value of some types of *estimand* depends directly on the extent and nature of the withdrawal process.
   - A *de facto estimand* will be an average over those who complete treatment up to final visit and also over those who do not fully comply.
2. Any *estimand* must define a treatment strategy for those who do not comply, either because they cannot or will not. [*follow-on regime*]
   - The actual regime used in the trial after treatment withdrawal may or may not match this *follow-on regime*.

There is no agreed methodology for handling such *de facto estimands* in a primary analysis.

The Big Issue

- Regulators are moving the emphasis for estimands in confirmatory studies away from efficacy and towards effectiveness.
  - *De facto* rather than *de jure* estimands may soon be required for primary analyses.
  - If so, what will replace MMRM as the default approach for handling early withdrawal in longitudinal studies?
  - Currently such *de facto* estimands lie in the domain of sensitivity analysis using multiple imputation, often known as reference-based imputation. These serve a different purpose.
  - This is a modelling issue with computational implications. It can be handled in a Bayesian, likelihood frequentist or perhaps even semi-parametric fashion.
- For primary analysis in a confirmatory trial we require any data imputation model and analysis model be congenial.
  - They are not congenial in current *de facto* sensitivity analyses which use Multiple Imputation.
  - To be congenial the analysis model would need to know whether the subject withdrew or not and allow for it, despite having complete data.

Not about "Missing Data". Proposed way forward.

- Two major events can occur to a trial participant before they complete which can both directly effect their future outcome.
  1. End of randomized treatment
  2. Withdrawal from study (end of observation)
- Data collected between these two events may or may not be relevant to the estimand in question.
  - Occurence of both events is likely to be related to previous observations.
  - Both events must be simultaneously modelled together with the outcome of interest *estimand*.
- Then from this estimated joint model derive the required assessment.
  - Average the possible outcomes implied by the *estimand's* scenario to obtain an expected prediction for the *estimand*.
  - This marginalisation step is crucially important and parallels approaches used for causal inference in the epidemiological setting.

Some notation

For simplicity we restrict to only one event (combined withdrawal of treatment and end of observation) and likelihood is that from a single subect.

- $\mathbf{R}$ is a vector indicating whether or not subject is observed at this visit.
- $\mathbf{y}_{obs}$ is vector of observed values whose length depends on $\mathbf{R}$
- $\mathbf{y}_{mis}$ is vector of non-observed potential values whose length also depends on $\mathbf{R}$, often called the "missing data".

Extrapolation Factorizaton (EF) of the full data model

In contrast to both pattern mixture and selection models ...

$$p(\mathbf{R}, \mathbf{y}_{obs}, \mathbf{y}_{mis}|\omega) = p(\mathbf{y}_{obs}, \mathbf{y}_{mis}|\mathbf{R}, \omega)\, p(\mathbf{R}|\omega)$$

$$p(\mathbf{R}, \mathbf{y}_{obs}, \mathbf{y}_{mis}|\omega) = p(\mathbf{R}|\mathbf{y}_{obs}, \mathbf{y}_{mis}, \omega)\, p(\mathbf{y}_{obs}, \mathbf{y}_{pot}|\omega)$$

... the distribution of the full data is factored into joint distribution of the observed values $\mathbf{y}_{obs}$ and the Response pattern $\mathbf{R}$, and then that for unobserved values $\mathbf{y}_{mis}$ conditional on the previous.

$$p(\mathbf{y}_{obs}, \mathbf{R}, \mathbf{y}_{pot}|\omega) = p(\mathbf{y}_{pot}|\mathbf{y}_{obs}, \mathbf{R}, \omega_E)\, p(\mathbf{y}_{obs}, \mathbf{R}|\omega_O)$$

[The Extrapolation Factorizaton (EF). See Daniels & Hogan. Chapter 9, section 9.1.1.]

In EF the potential part $\mathbf{y}_{pot}$ could be the data after withdrawal, which may or may not be seen. But it could similarly be some *potential*

Fitting the model to observed data

The model can be fitted using maximum likelihood (NLMIXED) or in a Bayesian way (MCMC) based on $p(\mathbf{y}_{obs}, \mathbf{R}|\omega_O)$ often seen as

$$p(y_1) \times p(R_1|y_1) \times p(y_2|y_1, R_1) \times p(R_2|y_2, y_1, R_1) \text{ etc.}$$

Both deliver an approximate "posterior" distribution for the parameters ω_O based on the observed part of the model.

- When ω_E and ω_O are the same set of parameters, we can then immediately predict the outcome.
- If ω_E contains extra parmaters not contained in ω_O then these need to be specified in some other way, either as fixed values or as some form of distribution (a prior).

Predicting the estimand - equivalent of Least-squares Means

- The *estimand* reflects some function $h(\mathbf{y}_{pred}, \mathbf{R}_{Pred})$ of the predicted outcome based on the full EF model.
  - For instance HbA1c at final visit. But might be an average, or even a score involving the time of event $\mathbf{R}$.
- The mean of $h(\ )$ can be evaluated at specific values of covariates $\mathbf{x}$.
- The distribution $D(\mathbf{x})$ across the covariate space defines the scenario required by the *estimand*.
  - For instance half male and half female.
  - It may be redundent if $h(\ )$ defines treatment difference.
  - This is what is usually specified in AT, OM and DIFF options of the SAS LSMEANS statement.

The required value is then the margin over the distribution of parameters (either from MLE or based on true Bayesian posterior).

$$\int P(\omega|\mathbf{y}_{obs}, \mathbf{R}) \left[\int E_{EF}[h(\mathbf{y}_{pred}, \mathbf{R}_{Pred})|\omega, \mathbf{x}] D(\mathbf{x}) d\mathbf{x} \right] d\omega$$

where

- $P(\omega|\mathbf{y}_{obs}, \mathbf{R})$ is the posterior distribution for the parameters ω using.
  - Posterior for ω_O based on observed data.
  - Other external information about additional parameters in ω_E

Computational aspects

- Computation is easy but possibly slow within standard software such as SAS.
  - The outer integral involves summation over a sample from the posterior distribution (output from MCMC). Easy to do.
  - The inner integral either uses an algebraic solution for specific situations or numerical integration. May be costly.
  - Precision of the estimate can be recovered from the distribution of values across outer loop. May require correction for imprecise value from inner loop.

Example: Repeated Measures Multivariate Normal

- Observed part of the EF model
  - $p(y_2|y_1, R_1)$ is simple regression on any previous values.
  - $p(R_2|y_2, y_1, R_1)$ is logistic regression model conditional on not failing previously.
  - Identical to MVNormal $N(\mu, \Sigma)$ with withdrawal imposed dependent on observed values.
- Projected part of EF model under MAR
  - Mean $\mu_{pot} + \Sigma_{pot,obs}\Sigma^{-1}_{obs,obs}(\mathbf{y}_{obs} - \mu_{obs})$ and Var. $\Sigma_{pot,obs}\Sigma^{-1}_{obs,obs}\Sigma^{T}_{pot,obs}$
- Now we can modify mean following J2R, CIR or CR.

The crucial step is to get difference between arms in response at final visit **margined** over the **joint** EF model including withdrawal process.

Results based on MAR, CR, CIR and J2R

Diff. Mean expected value	Bayesian EF model				% of MAR at Week 6
Week	1	2	4	6	
Method: MAR	0.22	-1.28	-2.16	-2.65	
SED	0.68	0.87	0.92	1.03	
CR	0.22	-1.18	-1.91	-2.23	84
SED	0.68	0.88	0.85	0.90	87
CIR	0.22	-1-03	-1.74	-2.01	76
SED	0.68	0.86	0.87	0.87	84
J2R	0.22	-1.05	-1.68	-1.71	65
SED	0.68	0.83	0.81	0.74	72

Table: Both estimate and SED are shrunk, with estimate shrunk slightly more.

5000 MCMC iterations and 1000 simulations within subject-by-simulation based on Observed Margins.
Data set & model from Mallinckrodt et al (Stat. Biopharm. Res., 5:4, 369-382, 2013)

James@livedata.co.uk Mike.Kenward@lshtm.ac.uk

Post-Authorisation Efficacy Studies

Ann Smith (AZ), Chrissie Fletcher (Amgen) on behalf of PSI Working Party

Contact information:
Ann Smith
Biostatistics & Information Sciences
AstraZeneca, da Vinci Building,
Melbourn Science Park,
Royston, Cambs, SG8 6HB

ann.smith2@astrazeneca.com

Introduction

Ongoing development of EMA guidance on scientific principles for post-authorisation efficacy studies is an opportunity to ensure consistent use and common understanding of terminology regarding study types and design. Existing regulation is restricted to studies conditional within a marketing authorisation. Key requirements include clear definitions for interventional and non-interventional studies, and alignment between study design options and adherence to Good Clinical/Pharmacovigilance/Pharmacoepidemiology Practice.

Background

A post-authorisation efficacy study (PAES) is a study that aims to clarify the benefits of a medicine on the market including efficacy in everyday medical practice. Guidance on how these studies should be carried out is under development, by a rapporteur working group set up in July 2014. This can support voluntary and imposed PAES

PAES Not a New Concept

Prior to Delegated Regulation (DR) (EU) 357/2014 separate legal frameworks existed for PAES:
- ❖Conditional Marketing Authorisation (MA)
- ❖MA in exceptional circumstances
- ❖MA for Advanced Therapy Medicinal Products
- ❖Paediatric use of a medicinal product
- ❖Referral procedures

PAES Delegated Regulation (EU) 357/2014 may be required for centrally or nationally authorised products either:

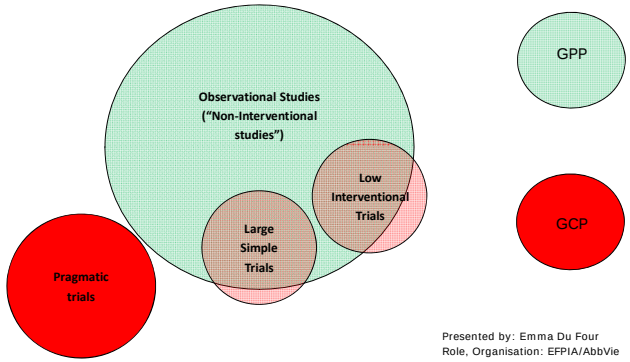
❖**At the time of granting the marketing authorisation**
(concerns relating to some aspects of the efficacy of the medicinal product are identified and can be resolved only after the medicinal product has been marketed)

❖**After granting the marketing authorisation**
(the understanding of the disease or the clinical methodology or the use of the medicinal product under real-life conditions indicate that previous efficacy evaluations might have to be revised significantly)

Considerations for PAES Guidance Content

- ❖ Principles-based approach, not specific methods for specific designs
- ❖ High-level scope – performance inside and outside the delegated regulation (DR)
- ❖ Covering imposed PAES, but mindful of voluntary PAES
- ❖ Aim for consistency of requests

Study Designs and Good Principles



General Methodological Considerations

Agreement recommended between sponsor and regulator covering:
- ❖ Proposed study design
- ❖ Path to interpretable and useable results (sufficiently addressing the uncertainty in question)

Further considerations
- -Early proactive planning versus reactive later in MAA
- -Role of scientific advice
- -Interacting with other stakeholders

Considerations Within PSI Working Group

Clear definitions of key terminology and study designs in the PAES guidance will be essential, for example Pragmatic Trials, Low Interventional Studies

Distinguish between designs for Hypothesis-testing and Estimation

Need acceptable process for how Observational Studies can be conducted and clarification of what regulatory processes need to be followed

Key principles for Design considerations (matching procedures; handling missing data; analysis methods including sensitivity analyses)

Considerations for Prospective versus retrospective study designs, aligned with availability and quality of data sources

Could forthcoming guidance aim for an equivalent of ICH E9 taking into account the different study designs, data sources and analysis and reporting considerations required for PAES?

Next Steps: Draft guidance for public consultation expected Q2/3 2015 for 3-month consultation. PSI (including the SIG for real world studies), and EFSPI, are keen to collaborate to optimise the final document. Do contact the authors, or Hermann Huss [Hermann-josef.huss@bayer.com] if you have experience or interest in supporting the PSI working group.

References

EMA website, top-level info and PASS info
http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/document_listing/document_listing_000377.jsp&mid=WC0b01ac058066e979

Slides Jane Moseley Jan2015
http://www.ema.europa.eu/docs/en_GB/document_library/Presentation/2015/03/WC500184245.pdf

October 2013 PAES Workshop
http://www.ema.europa.eu/docs/en_GB/document_library/Minutes/2013/11/WC500155692.pdf

Presented at the 2015 PSI Conference, London, UK

Increasing the Efficiency of Early Phase Decision Making Studies by Using ACRn within a Bayesian Framework

Foteini Strimenopoulou[1], Emma Jones[2], Ros Walley[1] and Margaret Jones[1]
[1] UCB, [2] Veramed

Introduction

ACR20 is, undoubtedly, the gold standard for assessing efficacy in diseases with an arthritis component such as rheumatoid arthritis (RA) and psoriatic arthritis (PsA). However, despite its considerable discriminant ability for detecting efficacy of treatment and, its ease of interpretation, there are significant drawbacks. The first is that ACR20 is a binary measure that lacks sensitivity to small changes and therefore studies that are powered to detect a difference in ACR20 typically require a large sample size. Secondly, ACR20 is unable to measure large improvements at an individual level and although this can be addressed by using ACR50 or ACR70, these endpoints are often less sensitive than ACR20 and, thus, require many more subjects in the study. In addition the wealth of information from historical studies enables the use of Bayesian methods, which may further reduce sample sizes.

Objective

- Reduce the number of subjects in early RA/ PsA studies by:
- ✓ Using an endpoint that is sensitive to change while keeping the clinical relevance of ACR20/50/70
- ✓ Using prior information from other similar studies

ACRn

ACRn is a continuous endpoint characterizes the percentage of improvement from baseline that a patient has experienced across a number of core measurements and relates directly to ACR20, ACR50, and ACR70 responses.

Relationship between ACRn and ACR20/50/70:

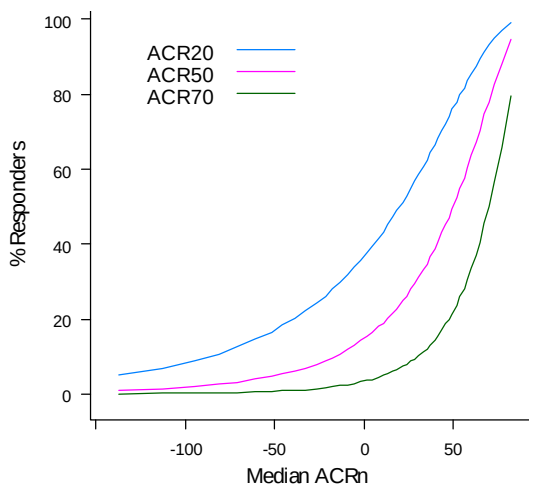
- If ACRn>=20% then the patient is ACR20 responder
- If ACRn>=50% then the patient is ACR50 responder
- If ACRn>=70% then the patient is ACR70 responder

Table 1. Individual examples of the relationship between ACRn and ACR20/50/70

	Patient 1	Patient 2	Patient 3
ACRn	19%	69%	90%
ACR20	Non responder	Responder	Responder
ACR50	Non Responder	Responder	Responder
ACR70	Non responder	Non responder	Responder

Figure 1 illustrates the relationship of the population median ACRn with the ACR20/50/70 responder rate.

Figure 1. Relationship of ACRn with ACR20/50/70



Therefore, ACRn:

- ❑ maintains the clinical relevance
- ❑ can account for both large and small clinical improvement
- ❑ is more sensitive to change

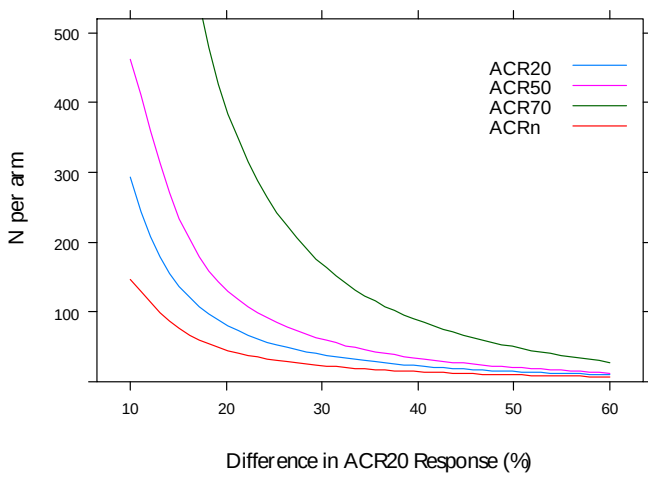
The reasons above make ACRn an appealing alternative endpoint for many situations such as:

- ✓ head-to-head comparison of two products,
- ✓ assessing risk/benefit of a drug,
- ✓ dose optimisation
- ✓ early proof-of-concept studies with limited resources.

Why using ACRn instead of ACR20/50/70 can reduce the sample size?

ACRn as seen previously is a continuous endpoint and by dichotomizing it we can calculate the ACR20/50/70 scores. Cohen *et al* (1983) have shown that in general if we dichotomize a continuous endpoint we lose power and so the number of subjects required in a study is increased. The amount of increase is mainly dependent on the cut-off point that is chosen (i.e. 20%, 50% or 70%). Figure 2 illustrates how the sample size required in a study for detecting a treatment effect increases when we use ACR20/50/70 as compared to ACRn. For illustration purposes, we have set the false negative and false positive rate to 20% and 5%, respectively (for a 2-sided test).

Figure 2. Comparison of the ACRn and ACR20/50/70 endpoints based on the number of subjects required in a study for detecting a range of treatment effects



Please note that since ACRn is not normally distributed, for calculating the sample size and the power of a test, we have performed a transformation to achieve normality. This transformation is $ACRn^t = \log(100-ACRn)$.

Bayesian approach to relevant historical data

Often at the start of a study, literature or in-house summary data are available that are relevant to placebo (or active comparator) response in the new study. These data come from historic studies with a similar study population and protocol to the current study. Traditional statistical approaches ignore this information, arguably leaving to the individual to synthesize information from the current study with the historical studies in an ad hoc manner at the end of the study.

Building a prior

At UCB we have adopted the Bayesian Meta-Analytic Predictive (MAP) approach described in Neuenschwander et al (2010) and Di Scala et al (2013) to summarise historical data and build a prior distribution that will be integrated in the design of Proof of Concept studies, as well as in the analysis of the new study data. Informative priors are constructed for the placebo response or the active comparator (and not for the treatment difference, nor the experimental drug arm) and are used for internal decision making only.

In the case of limited historical information available where the study to study variability is not easy to estimate and therefore not possible to use the MAP approach, we suggest an arbitrary discounting of the prior information to account for study to study variability. Two examples are:

- Normal case: inflate variability by 2*SEM → discounted prior reduces the effective sample size by 75%
- Binary case: Beta(a/4,b/4) → discounted prior reduces the effective sample size by 75%

Bayesian design and decision making

Sample size determination: Choose the sample size large enough to ensure that the trial will provide convincing evidence that treatment is better than control based on a chosen success criterion (see Walley et al.). This is:

- Success criterion S: $Pr(\delta > z|data) > 1-\alpha$
- Sample size such that $Pr(S|\ \delta = \delta^*) > 1-\beta$

Decision making

Once the data from the new study are available, the decision criteria applied are the same with the ones defined at the design stage. The priors remain the same or updated to include new information. Regarding the model, the same or a more complex model can be used than the one used at the decision making stage.

Application

Introduction to the study design:

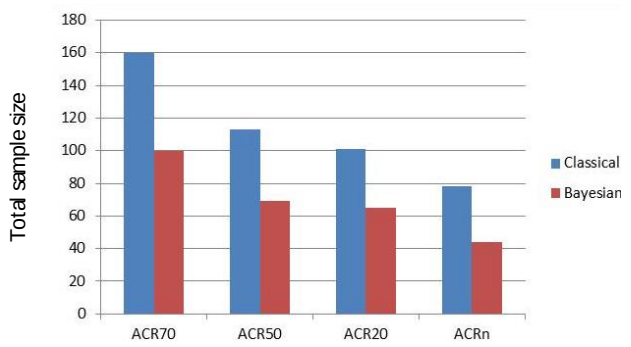
- ❑ Phase 2a multiple dose study in psoriatic arthritis → assess safety, tolerability and efficacy
- ❑ Efficacy analysis based on pooling subjects from placebo and top 3 dose levels.
- ❑ Primary efficacy endpoint: ACRn response at week 8
- ❑ Efficacy study decision criteria
  1. $Pr(\delta > 0|data) > 97.5\%$
  2. $Pr(\delta > 0.31|data) > 70\%$

Optimizing the quality and cost:

- ❑ The target efficacy was based on ACR20 and we have translated this to ACRn to take advantage of the good properties of the continuous endpoint and so→ reduce study size by 23 subjects
- ❑ We have used internal historical placebo data from a similar inhouse study to form the prior distribution for the placebo effect . Doing so we have replaced placebo subjects from the study with 'pseudo' subjects from the prior → further reduce study size by 34 subjects
- ❑ Design operating characteristics

		If effective	If Ineffective
Probability of the decision being:	No-Go	9%	97.5%
	Pause	11%	2%
	Go	80%	0.5%

Figure 4. Total sample size required for different endpoints assuming both a Classical and Bayesian framework



Other applications

Since the application of the methods described here for RA and PsA, the use of a primary continuous endpoint rather than established binary endpoints within a Bayesian framework is now routinely considered for many other therapeutic areas in the early phase group at UCB.

Where it is not possible to use an alternative continuous endpoint, the Bayesian approach may still be applied to the binary endpoint and lead to a reduction in sample size and/or a decrease in error rates.

Conclusion

- Using continuous endpoints (such as ACRn) as the primary endpoint rather than established binary endpoints and incorporating historical information by following the Bayesian paradigm, the sample size is reduced resulting in:
  - Lower study cost
  - Faster recruitment
- The use of continuous endpoints together with Bayesian methods are now routinely considered for early phase studies to increase efficiency

References

1.Cohen J. (1983) The cost of dichotomization. Applied Psychological Measurement 7(3):249–253.
2.Di Scala LK., Kerman J., Neuenschwander B. (2013) Collection, synthesis, and interpretation of evidence: a proof-of-concept study in COPD. Statistics in Medicine 32:1621-1634.
3. Neuenschwander B., Capkun-Niggli G., Branson M., Spiegelhalter D. (2010) Summarizing historical information on controls in clinical trials. Clin Trials 7: 5-18
4. Walley RJ., Birch CL., Gale JD., and Woodward PW. Advantages of a wholly Bayesian approach to assessing efficacy in early drug development: a case study. To appear in Journal of Pharmaceutical Statistics.

First Experiences in Observational Research – A Statistician's Perspective

Chris Toffis

Acknowledgements: Lucy DeCosta (Amgen Ltd)

An observational study is a study in which conditions are not under the control of the researcher. In particular, the exposures or treatments of interest are not assigned at random to experimental units by the investigator. In the pharmaceutical industry observational studies are increasingly performed to characterise and demonstrate the clinical value of drug products in real world populations: to assess comparative effectiveness of two or more medications, to continually evaluate the risk:benefit profile of medications and for post-marketing safety monitoring and evaluation.

There may be instances when an observational study would be more appropriate as opposed to a randomised clinical trial:

Violation of ethical standards

Hypothesis: Smoking causes lung cancer.

- In theory, to test this hypothesis individuals would be randomised to smoke or not to smoke then followed to subsequently determine the effect of smoking on lung cancer. This would be highly unethical experimental practice. An observational study can determine if an association exists between smoking and lung cancer by observing the smoking status of those individuals with the disease.

Rare events

- If we wanted to investigate the link between a certain medication and a very rare group of symptoms the pool of subjects for investigation would be very small. An observational study could identify a group of symptomatic patients and use historical information to ascertain if there is a relationship between the medication and symptom.

In this poster I discuss 3 observational studies I have been leading during the first 18 months of my biostatistical career at Amgen: one fully retrospective study investigating the effects of switching from our IP to a biosimilar, a prospective study examining persistence rates of our IP and a part-prospective/part-retrospective study investigating the effectiveness of a new dosing regimen for our IP. I present some challenges faced in each and the methods employed to overcome them.

Depending on the period of interest, observational studies can be either prospective, retrospective or (partly) both.

Prospective

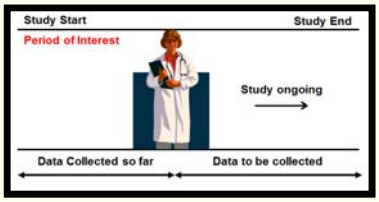
The period of interest aligns with the present and data is collected as it is generated.

Pros

- Able to follow patients in real-time.
- Can define data to be collected and query inaccurate information.

Cons

- Period of interest must finish in order to have the data: expensive and time-consuming.
- Susceptible to observation bias.



Retrospective

The period of interest exists wholly in the past.

Pros

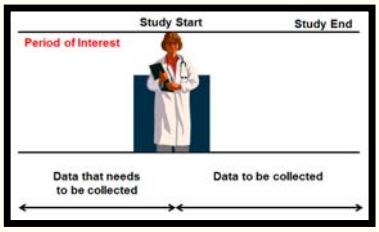
- Not affected by observation bias.
- Data can be procured 'now': cost- and time-efficient.

Cons

- Data can be incomplete – generally have to accept 'as is' even if missing.
- Limited to data collected per routine clinical practice during period of interest.

Part Retrospective/Part Prospective

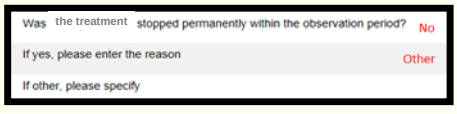
A hybrid of the two scenarios above. This type of observational study could be used if we wanted to monitor the effect of a change which has occurred – we collect data in the 'pre' period and follow up with data in the 'post' period.



Challenges

Data Inconsistencies

Missing data and data inconsistencies are common in observational studies. If there appears to be a systematic flaw in measuring exposure or outcome variables this will lead to information bias. An example of a data inconsistency that could not be resolved through query in my part-prospective/part-retrospective study is exhibited below:



There are two scenarios: a) the treatment was not stopped and the reason of 'Other' was entered erroneously or b) the treatment was stopped but the site entered 'No' instead of 'Yes'. Which scenario is correct? The following possible options were considered:

1. Identify other information that would support treatment continuing, for example dosing data, and if present assume the correct response is 'No'.
2. If no other supportive data is available, assume the worst case scenario relevant to the study outcomes i.e. assume the correct response is 'Yes'.
3. Treat response as 'missing' to avoid the inconsistency contributing to the analysis
4. Use the response given 'as is' but qualify the uncertainty in a footnote.

Fundamentally, whichever approach is taken, some information has been lost and thus the correct measurement of the exposure may not be accurate.

Observation Bias

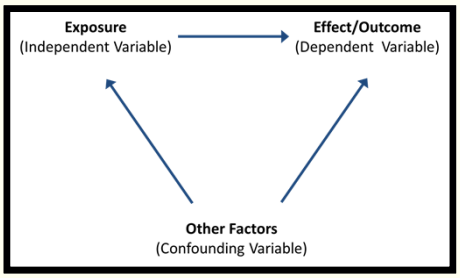
This form of bias is generally unavoidable for prospective studies . One of my studies investigating persistence of subjects using our IP is fully prospective. The study was designed to ensure that clinical practice of enrolled subjects was as close to routine as possible, to minimise any differences in care that may impact the outcomes of the study.

Selection Bias

Selection bias may have been introduced on my fully retrospective study. Certain lab values were 'censored' in the analysis due to the perception that they were artificially inflated following an external process (received by some subjects purely at the investigator's discretion). A sensitivity analysis was conducted including all lab values to assess the impact of the external process on the final results.

Confounding

Patient characteristics are often used to inform treatment decisions which may confound the relationship between exposure and outcome. Failure to account for such variables may bias the estimates of treatment effect. Our analyses were designed to investigate the influence of factors such as age and diabetic status on the outcomes of interest (lab parameter over time/in a specific range or persistence).



Summary

- Observational studies play a significant role in real-world evidence generation.
- Depending on the period of interest, different types of data collection can be considered.
- Many sources of bias are inherent to observational studies.
- Results of observational studies often confirm those reported in randomised controlled trials or generate hypotheses for further research.

Glossary

Confounding Variable – an extraneous variable in a model that correlates with both the dependent variable and the independent variable of interest

Information Bias – distortion in effect estimation that occurs when measurement of either the exposure or the outcome is systematically inaccurate

Observation Bias – Subjects and investigators taking part in a study may alter their behaviour as a result of knowing they are being observed

Selection Bias – refers to the systematic omission of individuals, groups or data for analysis thereby ensuring that the sample obtained is not representative of the population intended to be analysed.

ASSESSING THE CARDIOVASCULAR RISK OF ANTI-DIABETIC THERAPIES IN PATIENTS WITH TYPE 2 DIABETES MELLITUS

Richard C. Zink, JMP Life Sciences, SAS Institute on behalf of the ASA Biopharmaceutical Section Safety Working Group:
Aloka Chakravarty, Christy Chuang-Stein, Qi Jiang, Chunlei Ke, Haijun Ma, Jeff Maca, Olga Marchenko, Estelle Russek-Cohen & Matilde Sanchez-Kam

Introduction

- Type 2 diabetes mellitus (T2DM) accounts for 90-95% of all diabetes patients. These individuals are characterized with insulin resistance and/or insulin deficiency.
- FDA called for assessment of cardiovascular (CV) risk for non-insulin therapeutics for T2DM. Asked that hazard ratio (HR) of treatment compared to control be < 1.8 in pre-market evaluation (122 events, two-sided, α=0.05) [1]. Further, guidance suggested additional data collected post-market to show HR of Major Adverse Cardiovascular Event (MACE: CV death, nonfatal myocardial infarction and nonfatal stroke events) be < 1.3 (611 events, two-sided, α=0.05).
- Figure 1 displays the α-spending functions for a hypothetical sequential trial using OBF-like boundaries.
- Reviewed drugs approved by U.S. FDA to treat T2DM during 2002-2014. Main objective was to understand the impact of FDA guidance on assessment of CV risk in T2DM development programs.

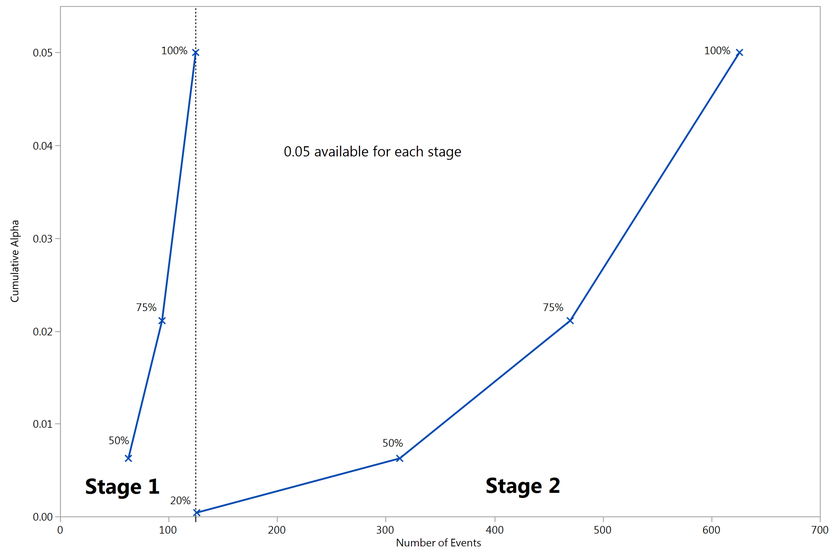


Figure 1. Cumulative α for Hypothetical Sequential Design Across Two Recommended Stages for T2DM

Methods

- Majority of information taken from www.clinicaltrials.gov, advisory committee materials, and Drugs@FDA: FDA Approved Drug Products.
- Review included drug name and class, initial US NDA submission and approval dates, initial MAA submission and approval dates, and strategy to address CV risk.
- Details on pre-marketing meta-analysis such as primary endpoint, whether prospectively adjudicated, study population, size of database, statistical hypotheses and methods, primary outcome, subgroup analyses, and major issues identified in FDA briefing documents.
- Details from CV outcome trial (CVOT) included whether initiated to address post-marketing commitment and/or contributed to pre-marketing CV analysis, study design, population, treatment groups, sample size, duration, primary endpoint and how adjudicated, primary objectives, completion date, and primary outcome.

Results

- CV risk assessment population typically consists of all randomized patients who received at least one dose of double-blind study therapy.
- For NME, CVOT usually required. Many sponsors started studies during late phase 3. Products whose individual components have been or currently evaluated for CV risk exempt from CV requirements.
- At NDA submission, sponsors typically proposed to conduct CV meta-analysis (MA) that included completed phase 2-3 studies. Because MA at submission stage, development programs often included plan to prospectively adjudicate CV events. In some cases (canagliflozin, alogliptin), MA included interim data from CVOT.

Results

- Cox proportional hazard models stratified by study were often the primary analysis method for more recent MA.
- CMH methods treating endpoints as binary often used as sensitivity analyses. MH and CMH methods stratified by study used as primary analysis for earlier submissions. Extensive subgroup analyses, pre-planned and ad hoc, conducted with consistency of results examined.
- Requirements at post-approval stage may include another MA including completed phase 2-3 studies plus CVOT; or analyzing CVOT as stand-alone study if planned with enough events.
- CVOTs designed as large, randomized, double-blinded, placebo-controlled in T2DM patients with high CV risk. T2DM drugs where development program was designed and/or completed but drugs not yet approved prior to FDA guidance held to same standards. FDA proposed post-hoc evaluations of CV events collected during development (liraglutide, saxagliptin hydrochloride, and exenatide XR).
- Steady increase in number of treated subjects included in pre-marketing CV risk assessment since 2008 (Figure 2). Observed two strategies to assess CV risk since the guidance.
- Substantial similarity in CV endpoints, adjudication, population, and statistical methods across recent CVOTs. No substantial delay in time between submission and approval due to addressing CV risk in recent programs. Did not review duration of development programs to determine if increased from first study in human to regulatory submission. All completed CVOTs ruled out HR > 1.3.

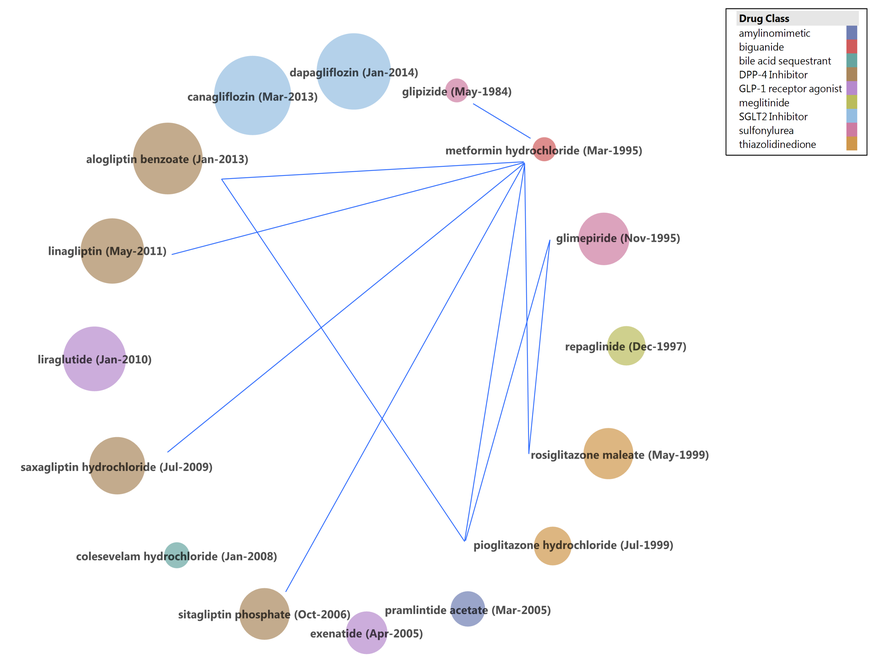


Figure 2. Drugs for Type 2 Diabetes Approved by the FDA up to Jan 2014

Conclusions

- Lot of similarities in approaches taken and subsequent analyses for T2DM programs thus far.
- Balance between evidence on CV safety and excessive delay of novel therapies.
- Access to interim data critical for CV assessment strategies. Releasing interim data when full approval granted (Stage 1) can undermine integrity for ongoing CVOTs. Guidance and buy-in from other regulatory agencies needed.
- Questions yet to answer: Can stop CVOT early? Post-market studies to assess CV risk instead of CVOT? Active-controlled CVOTs? Possible for indirect comparisons for CV risk among T2DM products?
- Additional details on this topic found in [2].

References

1.U.S. Food and Drug Administration. (2008). Guidance to Industry: Diabetes Mellitus — Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes.

2.Chakravarty A, Chuang-Stein C, Jiang Q, Ke C, Ma H, Maca J, Marchenko O, Russek-Cohen E, Sanchez-Kam M & Zink RC. (2015). Evaluation and review of strategies to assess cardiovascular risk in clinical trials in patients with type 2 diabetes mellitus. Submitted to *Statistics in Biopharmaceutical Research*.