PROTECT

imi efpia

Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium

Statistical signal detection in Clinical Trial data

Andreas Brueckner

Christiane Ahlers, Anngret Mallick, Nils Opitz, Vlasta Pinkston, Bruno Tran, Janet Scott, Harry Southworth, Bruno Tran, Lionel Van Holle, Nicola Wallis

PROTECT Symposium February 19-20 2015

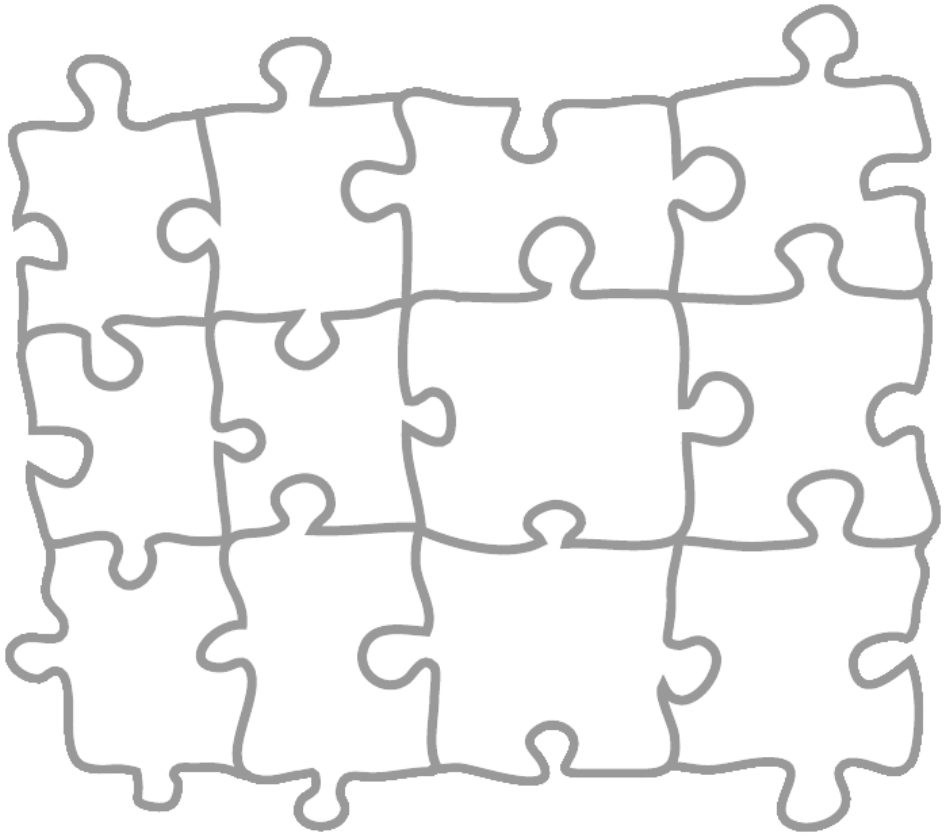
PROTECT

Disclaimer

The views expressed in this talk are those of the author.

They do not necessarily represent those of Novartis.

IMI PROTECT WP 3 – Signal Detection



PROTECT

Why Clinical Trials

- Available early during development
- Exposure Data
- Lab Data
- Longitudinal data
- Randomization

WP 3.9 Improving Signal Detection in Clinical Trials

Focus on evaluation and comparison of statistical methods for signal detection in different databases using:

- Adverse Event Data Screening
- Laboratory Data Modelling

Extreme value modelling of clinical laboratory data

Harry Southworth

PROTECT

The problem

- Typically, it is outlying values that suggest a safety issue.
  - E.g. large values of ALT, creatinine; small values of LVEF
- Most statistical methods aim to characterize expected values, not *unexpected* values.
- Such statistical methods *cannot* help.
  - This is not a criticism. The methods are not designed to characterize the extremes.
- To make any progress, we need new (to us) statistical methods.

Ximelagatran

Ximelagatran is an anticoagulant, developed by AstraZeneca (London, UK).

Although ximelagatran was granted marketing approval in several countries, the US Food and Drug Administration did not grant approval because, in part, of concerns over potential hepatotoxic effects of the drug suggested by elevations of alanine aminotransferase (ALT). In 2006, development of the compound was halted, and it was withdrawn from those markets in which it had been approved after reports of hepatotoxicity.

The Data

Available phase 2 data are the SPORTIF II study, a randomized, controlled clinical trial studying the prevention of stroke and transient ischaemic attacks in patients with atrial fibrilation.

The trial safety data provide baseline and post-baseline ALT data on a total of 246 patients, approximately 60 per arm, randomized to warfarin or to one of 20, 40 or 60 mg of ximelagatran twice per day.

The Data continued

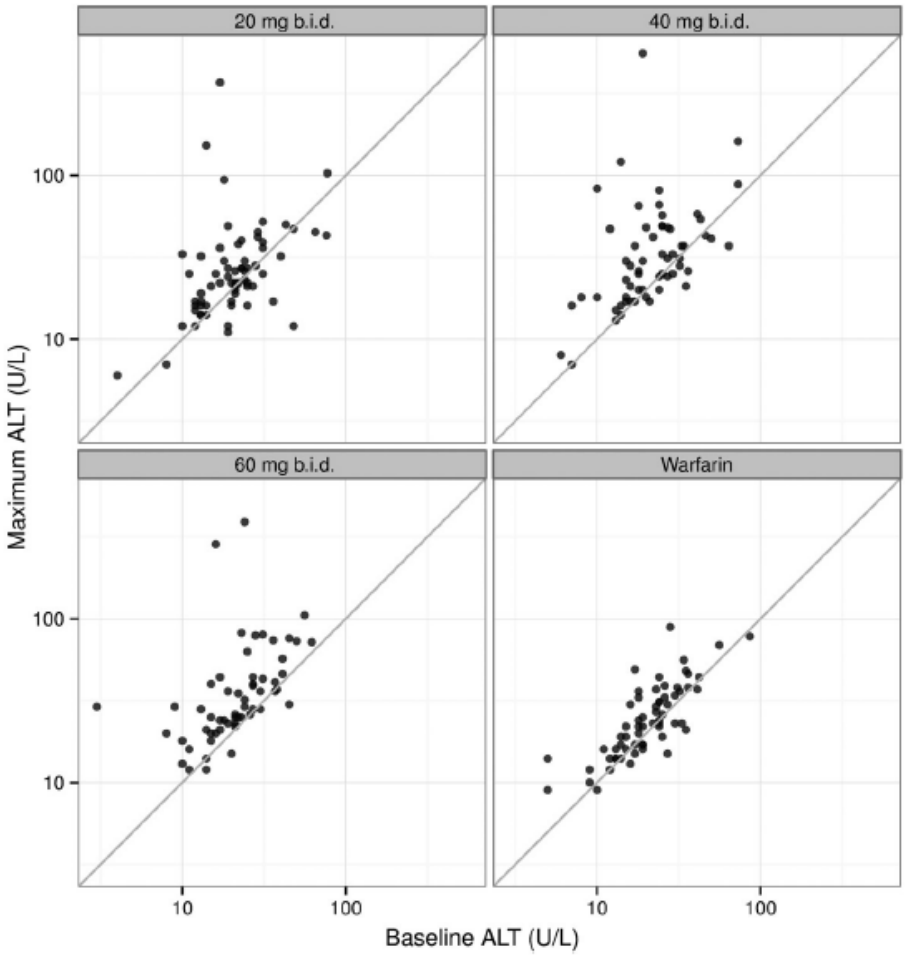


Figure 1. Shift plots of alanine aminotransferase (ALT) in the SPORTIF II study.

- The phase 3 trials with which we will compare the predicted values of ALT are SPORTIF III and SPORTIF V. In these trials, patients were randomized to warfarin or to 36 mg bid of ximelagatran.
- SPORTIF III collected data on approximately 1700 patients per group, and SPORTIF V collected data on approximately 1960 patients per group.
- The phase 3 studies were of 12–26 months duration, we should expect our predictions to be a little on the low side because in a longer p3 trial, there is clearly more opportunity for any patient to have an extreme ALT elevation.

Prediction

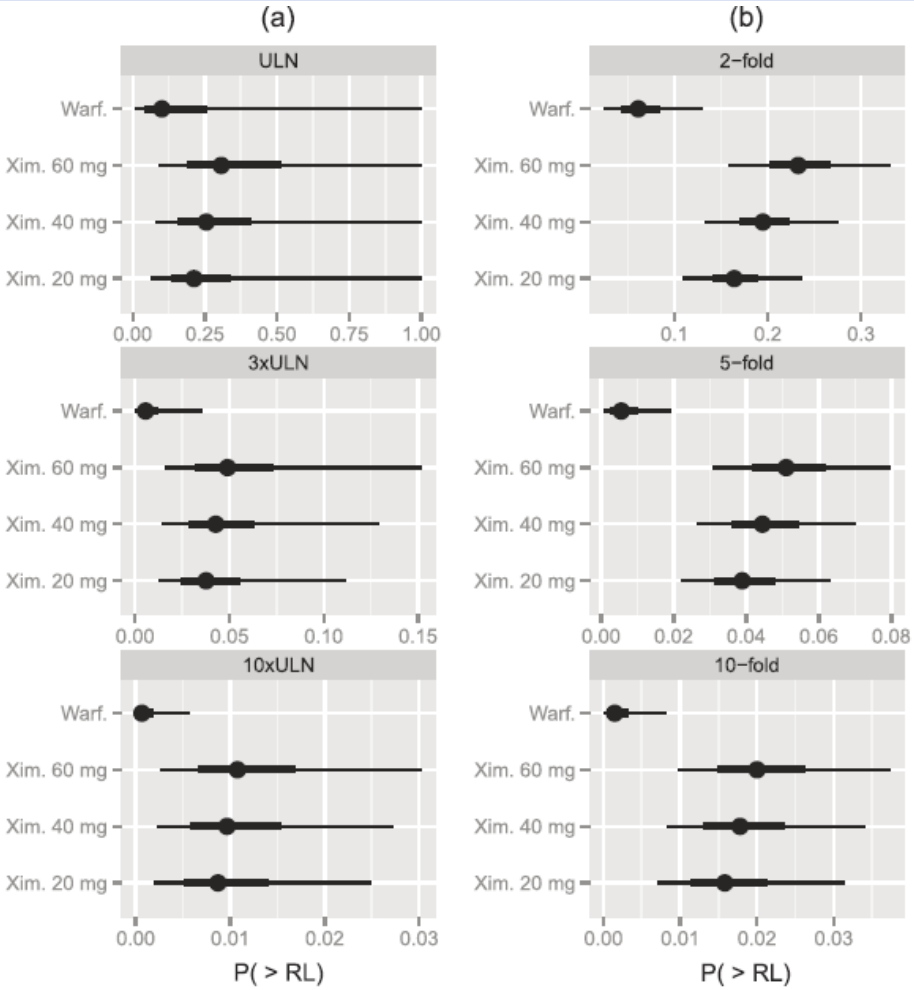


Figure 3. Predicted probabilities of exceeding (a) 1, 3 and 10 times upper limit of normal (ULN) and (b) 2-fold, 5-fold and 10-fold increase from baseline. Ximelagatran and warfarin are abbreviated as *Xim.* and *Warf.*, respectively. Note that the horizontal scales vary from panel to panel.

Observed vs. Predicted

Table V. Observed and predicted probabilities of exceeding specified multiples of upper limit of normal.

	$P(ALT > ULN)$	$P(ALT > 3\,ULN)$	$P(ALT > 10\,ULN)$
SPORTIF III	41.8%	9%	2%
SPORTIF V	38.6%	8.9%	1.5%
Predicted	25%	4.3%	0.97%

ALT, alanine aminotransferase; ULN, upper limit of normal.

Conclusions

If prior knowledge suggests data from a particular organ system should be monitored, consider extreme value modeling on data arising from each trial for the compound of interest. For example, if preclinical data suggested a potential liver issue, prepare to model ALT; if another compound in the class showed kidney signals, prepare to model creatinine.

PROTECT

Signal Detection in Clinical Trial AE data

PROTECT

Multiple Testing Problem

- Potential for a multiplicity issue in the monitoring of clinical trial safety events.
- Multiple tests for several hundred, several thousand events.
- Multiplicity issue for multiple tests over time
- Challenging issue whether multiplicity adjustment should be applied or not.

Multiple Testing Problem - Methods

Global Error Rate Control

Probability of making one or more false discoveries (type I error) among all the hypotheses when performing multiple hypotheses tests.

(Double) False Discovey Rate

FDR procedures are designed to control the expected proportion of incorrectly rejected null hypotheses ("false positives").

Bayesian Hierarchical Models

Approach that models the complete AE dataset.

Including hierarchical relationships

Do nothing – Unadjusted Analysis

Selected Signal Detection Methods

Method	Threshold
No Adjustment	$\alpha = 0.05$ $\alpha = 0.025$
FDR	$\alpha = 0.025$ $\alpha = 0.05$
Double-FDR method	$\alpha1 = 0.025$, $\alpha2 = 0.05$ $\alpha1 = 0.05$, $\alpha2 = 0.10$
Bayesian hierarchical 3-stage	$\alpha = 0.025$

Note: FDR – False Discovery Rate , MH – Mantel Haenszel, OR – Odds Ratio

The Diabetes Database

- 72 placebo controlled studies between 1988 and 2007 with >50 pats per trial
- Population:
  - 10300 patients on active treatment
  - 7800 patients on placebo
- Predominantly Caucasian (75%) , Black (8%), Asian(4%).
- Mean age 56 years (range 18-99).
- Sex distribition: Female patients (44%).
- Main Countries: USA (20%), Great Britain (20%), Germany (17%), Canada (11%).
- Indications: mostly Type 2 (NIDDM 80%), IGT (14%), Type 1 (IDDM 7%),

„Gold Standard“

- To evaluate the performance of different Signal detection methods requires the definition of a gold standard, which is a set of known and unknown safety topics. The performance of the individual methods will be evaluated by comparing the signal detection results (signal, no signal) versus the gold standard (ADR, no ADR).
- An event is considered to be an ADR, if it is currently listed in the ADR section of the corresponding CDS with a frequency of rare or higher.
- Rare events are included based on the theoretical chance to create a signal in an unadjusted analysis given the size of the database (i.e., assume frequency to by 1/1000.

Method Ranking

Signal detection methods will be ranked according to their performance with regard to

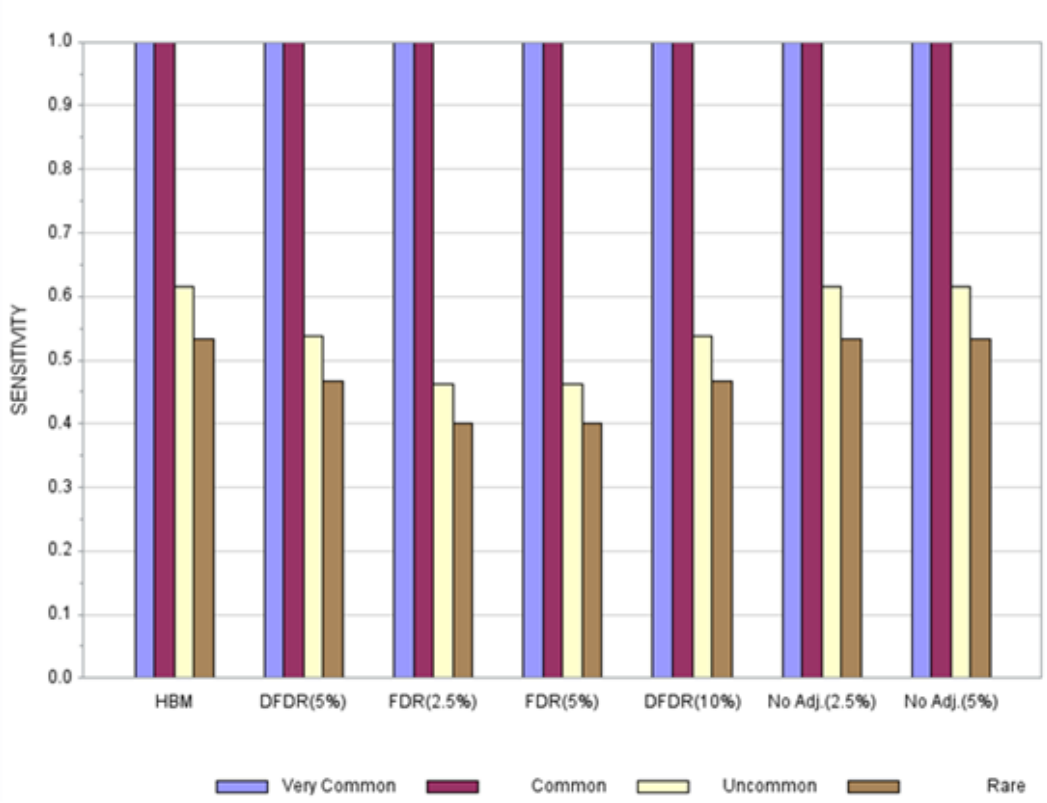
- Positive predictive value (PPV), i.e. the proportion of true signals among all flagged events
- Specificity, i.e. the proportion of not flagged events within all true non-signals
- Sensitivity, i.e. the proportion of flagged events within all true signals
- Negative predictive value (NPV), i.e. the proportion of true non-signals among all not flagged events

PROTECT

Results - Ranking

Model	PPV	Specificity	Sensitivity	NPV
Hierarchical Bayes	0.8889	0.9996	0.5333	0.9974
Double FDR adj., alpha2=5%	0.8750	0.9996	0.4667	0.9970
FDR adj., alpha=2.5%	0.8571	0.9996	0.4000	0.9966
FDR adj., alpha=5%,	0.8571	0.9996	0.4000	0.9966
Double FDR adj., alpha2=10%	0.7778	0.9992	0.4667	0.9970
No adj., alpha=2.5%	0.5333	0.9974	0.5333	0.9974
No adj., alpha=5%	0.2667	0.9917	0.5333	0.9974

Figure 1 Sensitivity by descending frequency with different multiplicity adjustments



Note: Sensitivity is estimated for ADRs of the displayed frequency or higher.

Comparative time in years to first signal

Model	Mean	Minimum	Median	Maximum
Hierarchical Bayes	0.67	0.00	0.00	3.00
Double FDR adj., alpha2=5%	1.67	0.00	0.00	10.00
FDR adj., alpha=2.5%	0.43	0.00	0.00	1.00
FDR adj., alpha=5%,	0.43	0.00	0.00	1.00
Double FDR adj., alpha2=10%	1.22	0.00	0.00	9.00
No adj., alpha=2.5%	0.22	0.00	0.00	1.00
No adj., alpha=5%	0.00	0.00	0.00	0.00

Note: time in years to first signal for each method was compared to the earliest time when any of the statistical methods under consideration generated a flag.

Conclusions

- Multiplicity adjustment provides a useful tool to improve the quality in signal detection in clinical trial data by increasing the positive predictive value.
- It is helpful tool to prioritize medical review
- The use of multiplicity adjustment needs to be evaluated against the size of the available clinical trial database.
- Bayesian Hierarchical Models can improve the efficiency of signal detection through borrowing of strength from other relevant events in the clinical trial dataset. This must be weighed against the more complex requirements of Bayesian modelling.

References and Further Reading

- Alvarez Y, Hidalgo A, Maignen F and Slattery J. Validation of Statistical Signal Detection Procedures in EudraVigilance Post-Authorization Data A Retrospective Evaluation of the Potential for Earlier Signalling. Drug Saf 2010; 33 (6): 475-487
- Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B, 57(1):289–300.
- Benjamini Y, Krieger AM, and Yekutieli D (2006). Adaptive linear step-up procedures that control the false discovery rate. Biometrika, 93(3):491–507.
- Berlin. Are all quantitative postmarketing signal detection methods equal? Performance characteristics of logistic regression and Multi-item Gamma Poisson Shrinker. Pharmacoepidemiol. Drug Safety 2012 Jun;21(6):622-30
- Berry SM, Berry DA. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. Biometrics, 60: 418-426).
- Chen W, Zhao N, Qin G and Chen J (2013): A Bayesian Group Sequential Approach to Safety Signal Detection, Journal of Biopharmaceutical Statistics, 23:1, 213-230.
- Council for International Organizations of Medical Sciences (2005). Final Report of CIOMS Working Group VI: Management of Safety Information from Clinical Trials. CIOMS, Geneva, 2005.
- DuMouchel W. Multivariate Bayesian Logistic Regression for Analysis of Clinical Study Safety Issues. Statist. Sci. Volume 27, Number 3 (2012), 319-339.Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. Journal of the Royal Statistical Society, Series B, 66(1):187–205
- EMEA (2002). CPMP Points to Consider on multiplicity issues in clinical trials.
- Hu JX, Zhao H, Zhou HH. False discovery rate control with groups. Journal of the American Statistical Association 2010 105: 1215–1227.

PROTECT

References and Further Reading

- Gould AL. Detecting potential safety issues in clinical trials by Bayesian screening. Biom. J. 2008 Oct;50(5):837-51.
- Mallick, A. (2012, March 27). From AE to ADR: Medical aspects. Copenhagen, Denmark: 24th Annual DIA EuroMeeting.
- Mehrotra DV and Heyse JF. (2004). Use of the false discovery rate for evaluating clinical safety data. Stat Methods Med Res. 2004 Jun;13(3):227-38.
- Mehrotra DV and Adewale AJ (2012), Flagging clinical adverse experiences: reducing false discoveries without materially compromising power for detecting true signals. Statist. Med., 31: 1918–1930.
- Prieto-Merino D. Use of Bayesian Hierarchical Models in Signal Detection. Presentation. PSI 2009.
- Romano JP, Shaikh AM, and Wolf M (2008). Control of the false discovery rate under dependence using the bootstrap and subsampling. (Invited Paper with discussion), TEST 17, 417-442.
- Rosenkranz GK. An Approach to Integrated Safety Analyses from Clinical Studies. Drug Information Journal November 2010 vol. 44 no. 6 649-657.
- Southworth H and O'Connell M. Data mining and statistically guided clinical review of adverse event data in clinical trials. Journal of Biopharmaceutical Statistics, 19:803 - 817, 2009.
- Xia A, Ma H, and Carlin BP. Bayesian hierarchical modeling for detecting safety signals in clinical trials. Journal of Biopharmaceutical Statistics, 21:1006 - 1029, 2011.
- Southworth H & Heffernan A, Extreme value modeling of laboratory safety data from clinical trials, Pharmaceutical Statistics.
- Southworth H. Predicting potential liver toxicity from phase 2 data: a case study with ximelagatran; Statistics in Medicine.