PROTECT

Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium

Statistical signal detection in Clinical Trial data

Christiane Ahlers, Andreas Brueckner, 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

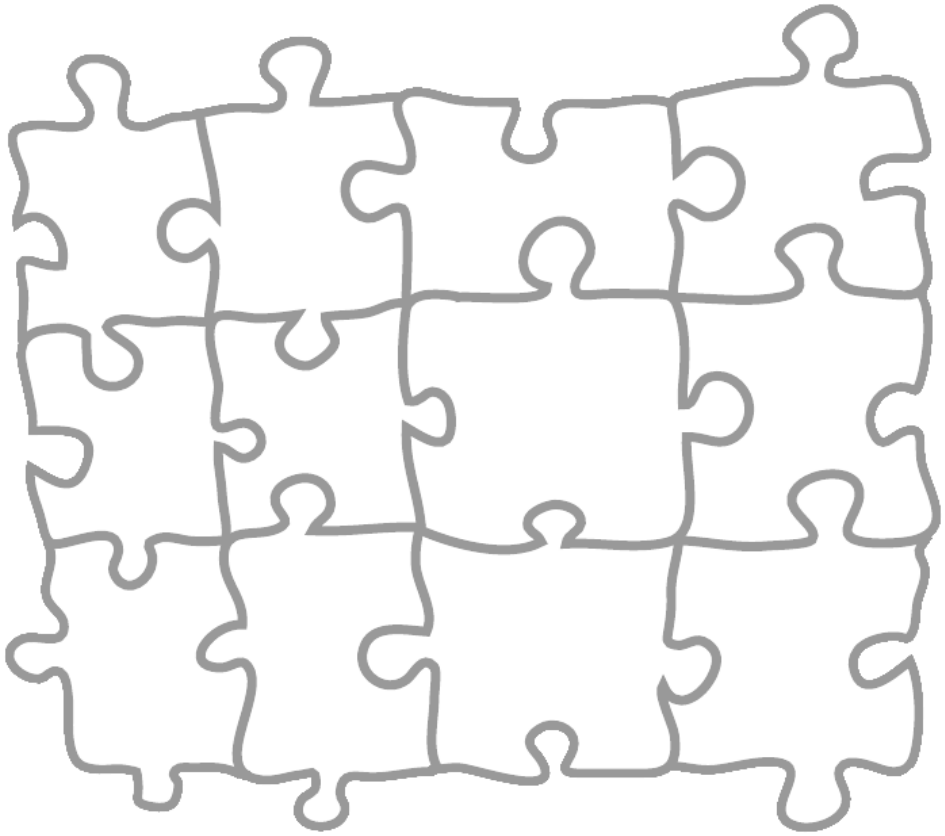
The views expressed in this talk are those of the author.

They do not necessarily represent those of Novartis.

Acknowledgements

- The research leading to these results was conducted as part of the PROTECT consortium (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium, www.imi-protect.eu) which is a public-private partnership coordinated by the European Medicines Agency.
- The PROTECT project has received support from the Innovative Medicine Initiative Joint Undertaking (www.imi.europa.eu) under Grant Agreement n° 115004, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution.

IMI PROTECT WP 3 – Signal Detection



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

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.
- Challenging issue whether multiplicity adjustment should be applied or not.

Real Life Example

In an vaccination clinical trial (1993), in a comparison of the relative frequencies for each AE, one of the 92 p-values attained 'statistical significance' (unadjusted p = 0.016). The AE in question was unusual high-pitched crying (UHPC), with an incidence of 6.7% and 2.3% for treatments A and B, respectively.

No medical rationale or biological plausibility seemed apparent for this unexpected finding. Accordingly, the researchers posited that the finding was a spurious chance event, a statistical artifact of drawing inferences from multiple p-values.

Nevertheless, because of concerns about the potential signal being real due to an unobservable cause, regulators mandated a follow-up study to prospectively compare the incidence of UHPC between the two treatments.

In the resulting trial, the incidence of UHPC was similar for the two groups (A: 5.1%, versus B: 4.3%; p = 0.532), indicating that the UHPC finding in the initial study was a false discovery.

Another Real Life Example

A safety and immunogenicity trial of a candidate quadrivalent vaccine against measles, mumps, rubella and varicella (MMRV) conducted in 296 healthy toddlers, 12–18 months of age.

Participants were randomly assigned to receive the quadrivalent MMRV on day 0 (Group 1) or the trivalent MMR on day 0 followed by varicella (V) on day 42 (Group 2).

Mehrotra DV and Heyse JF. (2004). Use of the false discovery rate for evaluating clinical safety data.

MMRV vaccination data

Table 1 Clinical AE counts ('Tier 2' AEs) for the MMRV study

No.	BS	Adverse experience	Group 1 ($n_1 = 148$) X_1	Group 2 ($n_2 = 132$) X_2	Diff (%)	*P* value
1	01	Asthenia/fatigue	57	40	8.2	0.1673
2	01	Fever	34	26	3.3	0.5606
3	01	Infection, fungal	2	0	1.4	0.4998
4	01	Infection, viral	3	1	1.3	0.6248
5	01	Malaise	27	20	3.1	0.5248
6	03	Anorexia	7	2	3.2	0.1791
7	03	Candidiasis, oral	2	0	1.4	0.4998
8	03	Constipation	2	0	1.4	0.4998
9	03	Diarrhea	24	10	8.6	0.0289*
10	03	Gastroenteritis, infectious	3	1	1.3	0.6248
11	03	Nausea	2	7	−4.0	0.0889
12	03	Vomiting	19	19	−1.6	0.7295
13	05	Lymphadenopathy	3	2	0.5	1.0000
14	06	Dehydration	0	2	−1.5	0.2214
15	08	Crying	2	0	1.4	0.4998
16	08	Insomnia	2	2	−0.2	1.0000
17	08	Irritability	75	43	18.1	0.0025*
18	09	Bronchitis	4	1	1.9	0.3746
19	09	Congestion, nasal	4	1	1.9	0.3746
20	09	Congestion, respiratory	1	2	−0.8	0.6033
21	09	Cough	13	8	2.7	0.4969
22	09	Infection, respiratory, upper	28	20	3.8	0.4308
23	09	Laryngotracheobronchitis	2	1	0.6	1.0000
24	09	Pharyngitis	13	8	2.7	0.4969
25	09	Rhinorrhea	15	14	−0.5	1.0000
26	09	Sinusitis	3	1	1.3	0.6248
27	09	Tonsillitis	2	1	0.6	1.0000
28	09	Wheezing	3	1	1.3	0.6248
29	10	Bite/sting, non-venomous	4	0	2.7	0.1248
30	10	Eczema	2	0	1.4	0.4998
31	10	Pruritus	2	1	0.6	1.0000
32	10	Rash	13	3	6.5	0.0209*
33	10	Rash, diaper	6	2	2.5	0.2885
34	10	Rash, measles/rubella-like	8	1	4.6	0.0388*
35	10	Rash, varicella-like	4	2	1.2	0.6872
36	10	Urticaria	0	2	−1.5	0.2214
37	10	Viral exanthema	1	2	−0.8	0.5033
38	11	Conjunctivitis	0	2	−1.5	0.2214
39	11	Otitis media	18	14	1.6	0.7109
40	11	Otorrhea	2	1	0.6	1.0000

BS: Body system.
* *P* value < 0.05

Mehrotra DV and Heyse JF. (2004). Use of the false discovery rate for evaluating clinical safety data.

PROTECT

Signals?

	Group 1 (n = 148)	Group 2 (n=132)	Diff (%)	P- Value
Diarrhea	24	10	8.6	0.0289
Irritability	75	43	18.1	0.0025
Rash	13	3	6.5	0.0209
Rash, measles/ rubella-like	8	1	4.6	0.0388

Guidance on Multiplicity Adjustments

Several guidance documents discuss multiplicity but there is no clear recommendation on whether to adjust or not

- ICH E9 - STATISTICAL PRINCIPLES FOR CLINICAL TRIALS (1999)
- EMA - Points to Consider on Multiplicity Issues in Clinical Trials (2002)
- CIOMS VI – Management of Safety Information from Clinical Trials (2005)

ICH E9

The calculation of p-values is sometimes useful ... as a 'flagging' device applied to a large number of safety variables.

If hypothesis tests are used, statistical adjustments for multiplicity to quantify the type I error are appropriate.

ICH E9 - STATISTICAL PRINCIPLES FOR CLINICAL TRIALS (1999)

CPMP - PtC on Multiplicity Issues

where a large number of statistical test procedures is used to serve as a flagging device to signal a potential risk ... adjustment for multiplicity is counterproductive for considerations fo safety.

CPMP Points to Consider on Multiplicity Issues in Clinical Trials (2002)

Multiple Testing Problem - Approaches

- Global Error Rate Control
- False Discovey Rate
- Bayesian Hierarchical Models
- Do nothing – Unadjusted Analysis

Global Type I Error Rate Control

- Probability of making one or more false discoveries (type I error) among all the hypotheses when performing multiple hypotheses tests.
- Penalise P-values - Low sensitivity

To control the global type 1 error rate with the Bonferoni procedure at the 5% level, an analysis of 10 AE would require a comparison of each individual p-values against a treshold of:

$$0.05/10 = 0.005.$$

For 2000 analyses, this treshold would be 5%/2000= 0.000025.

PROTECT

Signals?

	Group 1 (n = 148)	Group 2 (n=132)	Diff (%)	P- Value
Diarrhea	24	10	8.6	0.0289
Irritability	75	43	18.1	0.0025
Rash	13	3	6.5	0.0209
Rash, measles/ rubella-like	8	1	4.6	0.0388

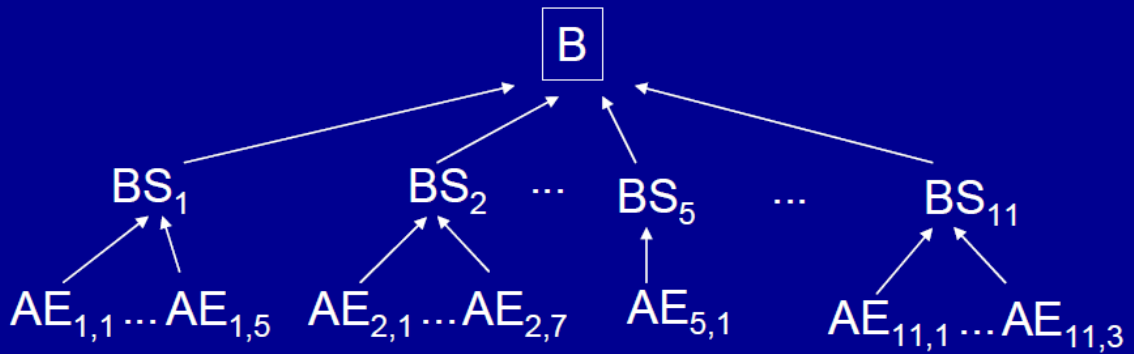
Bonferoni procedure:

0.05/40 = 0.00125.

Bayesian Hierarchical Models

- Approach that models the complete AE dataset
- Including hierarchical relationships

- Clinical trial with the MMRV to assess ADR for varicella component.
- 40 AE were studied (level 1), grouped in 8 "body systems" (BS) (level 2), that were grouped in one final "whole body" group (level 3)



MMRV vaccination data

The three-level hierarchical model results for the example of Table 1. $p(\theta = 0)$ is the probability that the treatment and control have the same AE rates and $p(\theta > 0)$ is the probability that treatment has a higher AE rate. The entries in bold-faced type correspond to the asterisked entries in Table 1.

			Post probability					Post probability	
b	j	Type of AE	$\theta = 0$	$\theta > 0$	b	j	Type of AE	$\theta = 0$	$\theta > 0$
1	1	Asthenia/fatigue	0.762	0.211	9	4	Cough	0.906	0.062
1	2	Fever	0.827	0.122	9	5	Infection, respiratory	0.897	0.083
1	3	Infection, fungal	0.796	0.101	9	6	Bronchitis	0.898	0.047
1	4	Infection, viral	0.813	0.100	9	7	Pharyngitis	0.906	0.061
1	5	Malaise	0.826	0.116	9	8	Rhinorrhea	0.904	0.051
3	1	Anorexia	0.821	0.117	9	9	Sinusitis	0.903	0.051
3	2	Cendisiasis, oral	0.835	0.083	9	10	Tonsillitis	0.905	0.042
3	3	Constipation	0.812	0.101	9	11	Wheezing	0.907	0.050
3	4	**Diarrhea**	0.743	**0.231**	10	1	Bite/sting	0.859	0.087
3	5	Gastroenteritis	0.823	0.093	10	2	Eczema	0.860	0.070
3	6	Nausea	0.805	0.050	10	3	Pruritis	0.868	0.062
3	7	Vomiting	0.849	0.076	10	4	**Rash**	0.784	**0.190**
5	1	Lymphadenopathy	0.717	0.136	10	5	Rash, diaper	0.852	0.099
6	1	Dehydration	0.666	0.087	10	6	**Rash, measles/rub-like**	0.836	**0.126**
8	1	Crying	0.655	0.185	10	7	Rash, varicella-like	0.862	0.076
8	2	Insomnia	0.661	0.153	10	8	Urticaria	0.852	0.048
8	3	**Irritability**	0.214	**0.780**	10	9	Viral exanthema	0.855	0.055
9	1	Bronchitis	0.900	0.059	11	1	Conjunctivitis	0.721	0.079
9	2	Congestion, nasal	0.901	0.058	11	2	Otitis media	0.757	0.102
9	3	Congestion, respiratory	0.896	0.040	11	3	Otorrhea	0.749	0.121

Berry SM, Berry DA. (2001) Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model.

Double False Discovery Rate

- FDR procedures are designed to control the expected proportion of incorrectly rejected null hypotheses ("false positives")
- As example: an analysis controlling the FDR at the 5% level identifies 20 signals. We would expect maximal one of these signals to be false positive.
- Double FDR is a two step adjustment procedure, first on the higher group level (e.g. SOC), then on the analysis level (e.g. PT).

Double False Discovery Rate

Table 4 Illustration of the Double FDR adjustment procedure

Body system ID	Number of AE types	Unadjusted minimum *P* value	FDR adjusted minimum *P* value
First level FDR adjustment			
Nervous system and psychiatric	3	0.0025	0.0200
Skin	9	0.0209	0.0771
Digestive system	7	0.0289	0.0771
Body site unspecified	5	0.1673	0.2952
Special senses	3	0.2214	0.2952
Metabolic / immune	1	0.2214	0.2952
Respiratory	11	0.3746	0.4281
Hematologic and lymphatic	1	1.0000	1.0000
Second level FDR adjustment			
Body system: nervous system and psychiatric			
Irritability		0.0025	0.0075
Crying		0.4998	0.7497
Insomnia		1.0000	1.0000

AE: Adverse experience.
FDR: False discovery rate.

Mehrotra DV and Heyse JF. (2004). Use of the false discovery rate for evaluating clinical safety data.

PROTECT

The Alpha-Glucosidase Inhibitors Bayer Research 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%),

Events and Coding

Used for analysis:

- Each event is assigned to either a MedDRA Preferred Term (PT) or a
- Bayer specific MedDRA Labelling Grouping (MLG).
  - MLGs summarize medically similar MedDRA Preferred Terms to allow consideration of event groupings that are not as specific as MedDRA Preferred Terms.
  - As it will not be distinguished between event terms and MLGs in this report, these groupings will be referred to as event terms in the following

Bayer's MLGS - Examples

MLG: Increase in transaminases

- PT: Alanine aminotransferase increased
- PT: Aspartate aminotransferase increased
- PT: Hypertransaminasaemia
- PT: Transaminases increased

MLG: Jaundice

- PT: Cholestasis
- PT: Jaundice
- PT: Jaundice cholestatic
- PT: Jaundice hepatocellular

MLG: Flatulence

- PT: Flatulence

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

„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

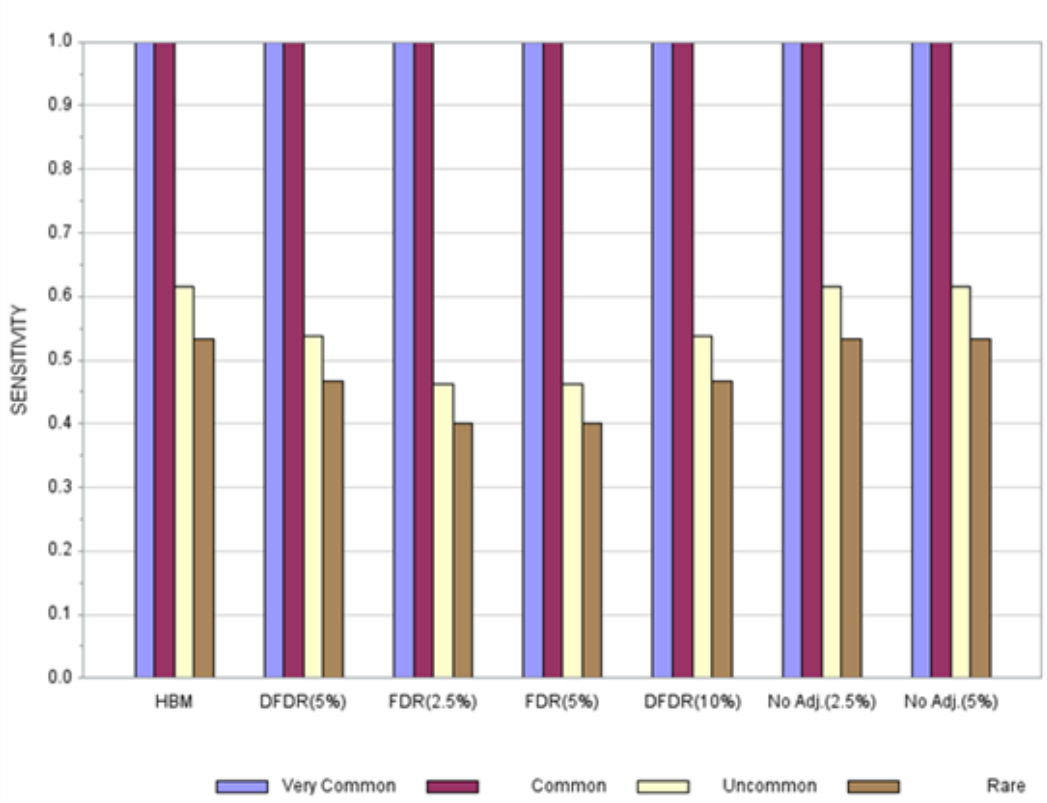
- Positive predictive value (PPV), i.e. the proportion of true signals among all flagged events
- Specificity, i.e. the proportion of not flagged events within all true non-signals
- Sensitivity, i.e. the proportion of flagged events within all true signals
- Negative predictive value (NPV), i.e. the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

AE Data Modelling - Conclusions

- Multiplicity adjustment provides a useful tool to improve the quality in signal detection in clinical trial data by increasing the positive predictive value.
- 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.
- The use of specific MedDRA groupings can further improve signal detection in clinical trial data.

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.

References and Further Reading

- Gould AL. Detecting potential safety issues in clinical trials by Bayesian screening. Biom. J. 2008 Oct;50(5):837-51.
- Grunert J: Signalerkennung in klinischen Studiendaten unter Verwendung der False Discovery Rate. 2010. Diplomarbeit. Technische Universität Dortmund.
- International Conference on Harmonisation (ICH). Statistical principles for clinical trials, ICH topic E9. EMEA: Canary Wharf, London, 1998.
- Lutkewitz S: Signalerkennung in klinischen Studiendaten mit hierarchischen Bayes Modellen. 2012. Masterarbeit . Technische Universität Dortmund.
- 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.