Can we identify patients at high risk of harm under a generally safe intervention?

Gerd Rosenkranz

Section for Medical Statistics
Center for Medical Statistics, Informatics, and Intelligent Systems
Medical University of Vienna, A-1090 Vienna, Austria

European Statistical Meeting on Analysis of Safety Data in Clinical Trials
Leiden, NL, June 23, 2017

This work has been funded by the FP7-HEALTH-2013-INNOVATION-1 project Advances in Small Trials Design for Regulatory Innovation and Excellence (ASTERIX). Grant Agreement No. 603160. Website: http://www.asterix-fp7.eu/

Table of Contents

Table of Contents

The current status of personalized medicine

Indication	Marker	Compound
Breast cancer	HER2+	trastuzumab
		pertuzumab
	HER2-/ER+	everolimus
Colorectal cancer	KRAS	cetuximab
		panatumumab
Cystic fibrosis	G551D	ivacaftor
Melanoma	BRAF V600E	vemurafenib
		dabrafenib
	BRAF V600E or V600K	trametinib
NSCLC	ALK	crizotinib
Solid tumors	MSI-H or dMMR	pembrolizumab

Table 1: Drug approvals with biomarkers (more in [1])

The natalizumab case

- Tysabri® (natalizumab) is approved for relapsing multiple sclerosis
- It increases the risk of a rare brain infection called progressive multifocal leukoencephalopathy (PML) that usually leads to death or severe disability
- A precautionary measure is to test for a John Cunningham Virus (JCV) infection which is also increasing the PML risk
- Other than that there is no biomarker to predict an increased risk for PML
- Treatment with natalizumab has to be stopped after 2 years the latest
- Because of the PML risk, the medication is available only through a restricted distribution program

Personalized medicine from a safety point of view

- Currently personalized medicine focuses primarily on **efficacy** to identify patients that **respond better** to treatment than others or patients that do not respond sufficiently well
- However there are situations where a small proportion of patients is facing a high risk of **substantial harm** under a treatment which is safe and efficacious for the majority of patients

Goal of personalized drug safety

To identify **patients not to be treated** with an otherwise safe and efficacious intervention because of **high risk of harm**

Personalized safety versus efficacy

- Risk-balancing differs between personalized efficacy and safety:

	falsely non-personalized	falsely personalized
Efficacy	TX inefficacious for subgroup	subgroup not treated with efficacious TX
Safety	TX harmful for subgroup	

- If an intervention can be really harmful in a subgroup one may wish to identify practically all subjects of this group at the expense of identifying too many
- So-called ‘conservative’ procedures may not be appropriate

Table of Contents

Example 1: Renal safety of contrast media [2]

Background

Contrast–induced nephropathy is a serious complication of diagnostic and interventional procedures

Objectives

To compare nephrotoxicity of two contrast media and to identify predictors of contrast induced nephropathy

Methods

Individual patient data ($n = 2727$) meta-analysis from 16 double-blind, randomized, controlled trials with stratification according to chronic kidney disease (CKD) and diabetes mellitus (DM)

Endpoints

Increase in serum creatinine (Cr) and incidence of post-procedural contrast-induced nephropathy (CIN)

Example 1: Renal safety of contrast media

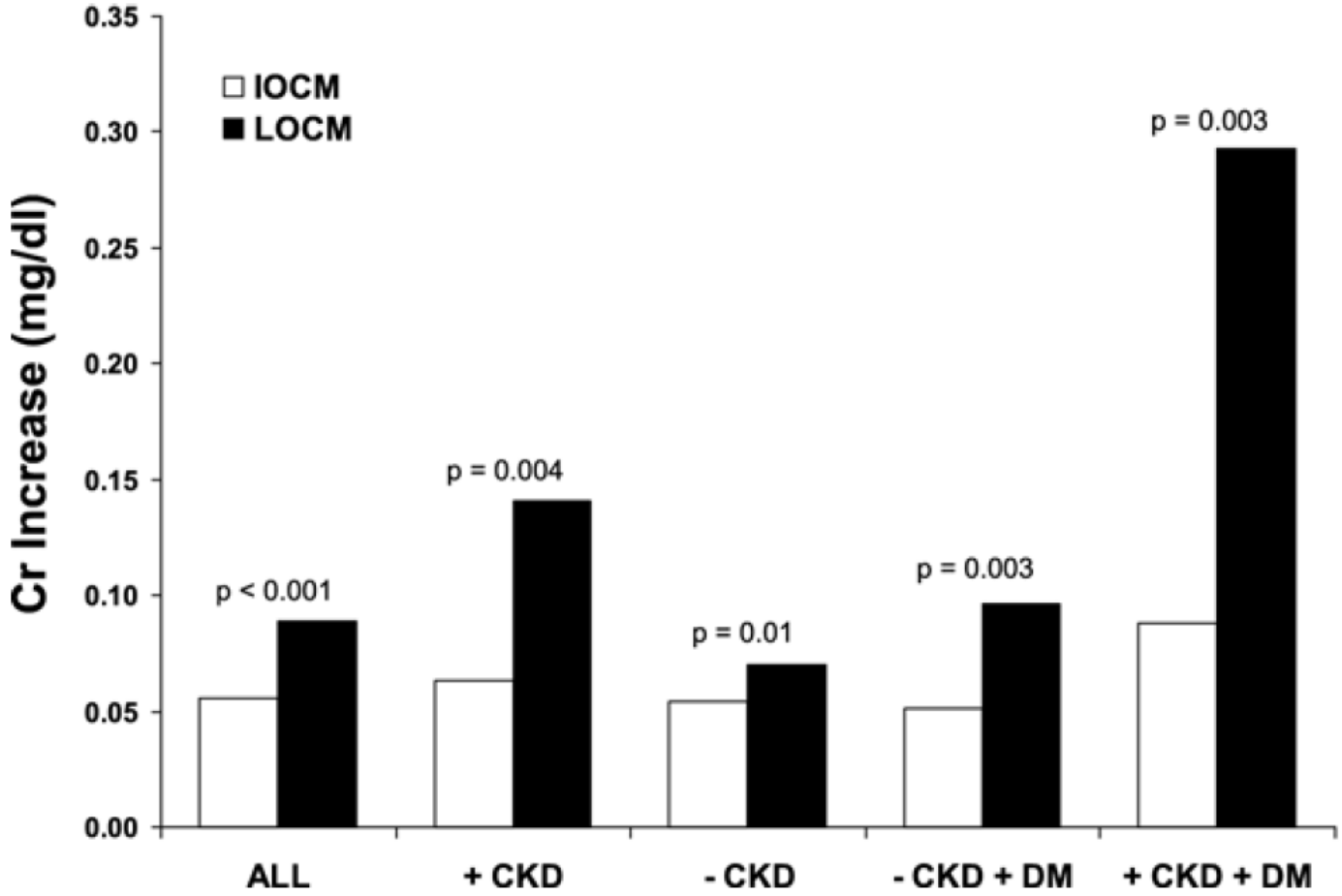


Figure 1: Maximum increase of creatinine concentration from baseline to day 3, after contrast administration (CKD = chronic kidney disease, DM = diabetes mellitus)

Example 2: Risk of bleedings under Alteplase [3]

Background

Although recombinant tissue plasminogen activator has been approved for acute ischemic stroke, concerns linger regarding its safety

Objectives

To analyze whether special subgroups of patients (i.e., age > 70 years, NIHSS score > 30, diabetes, CHF, hispanic origin) have a higher risk of symptomatic intracerebral hemorrhage (SICH)

Methods

Four prospective observational studies of acute stroke patients treated within 3h with Alteplase ($n = 1966$) were analyzed

Endpoints

Occurrence of SICH

Example 2: Risk of bleedings under Alteplase

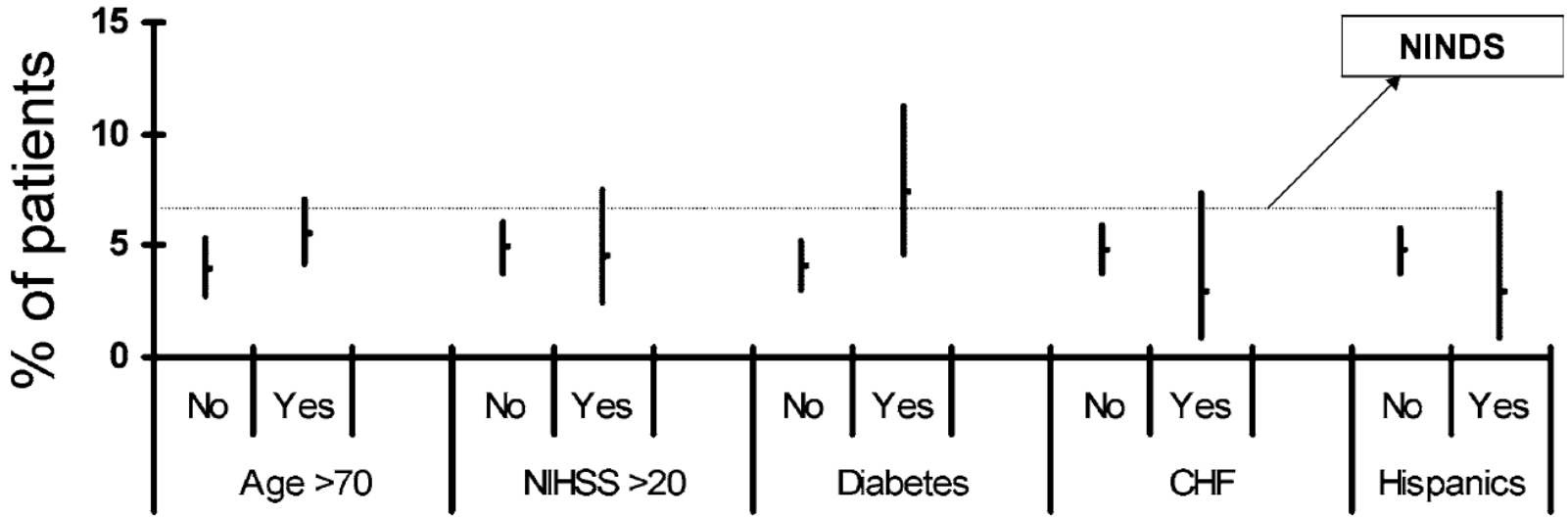


Figure 2: Incidence of SICH with 95% confidence intervals for subgroups among alteplase treated patients. Overall SICH rate was 4.7% (not shown).

Reference line (6.4%) from randomized, placebo controlled National Institute of Neurological Disorders and Stroke (NINDS) trial [4]

Example 3: Lumiracoxib-related liver injury [6]

Background

Concerns over hepatotoxicity have contributed to the withdrawal or non-approval of lumiracoxib, efficacious in osteoarthritis and acute pain

Objectives

To identify genetic markers able to identify individuals at risk for developing drug-induced liver injury (DILI)

Methods

Case-control genome-wide association study on 41 treated patients with ALT/AST $>$ 5 times ULN and 176 patients without liver injury using DNA samples collected from the TARGET study [5]

Endpoints

(Time to) liver enzyme elevations of AST/ALT $>$ 3 or $>$ 5 times ULN

Example 3: Lumiracoxib-related liver injury

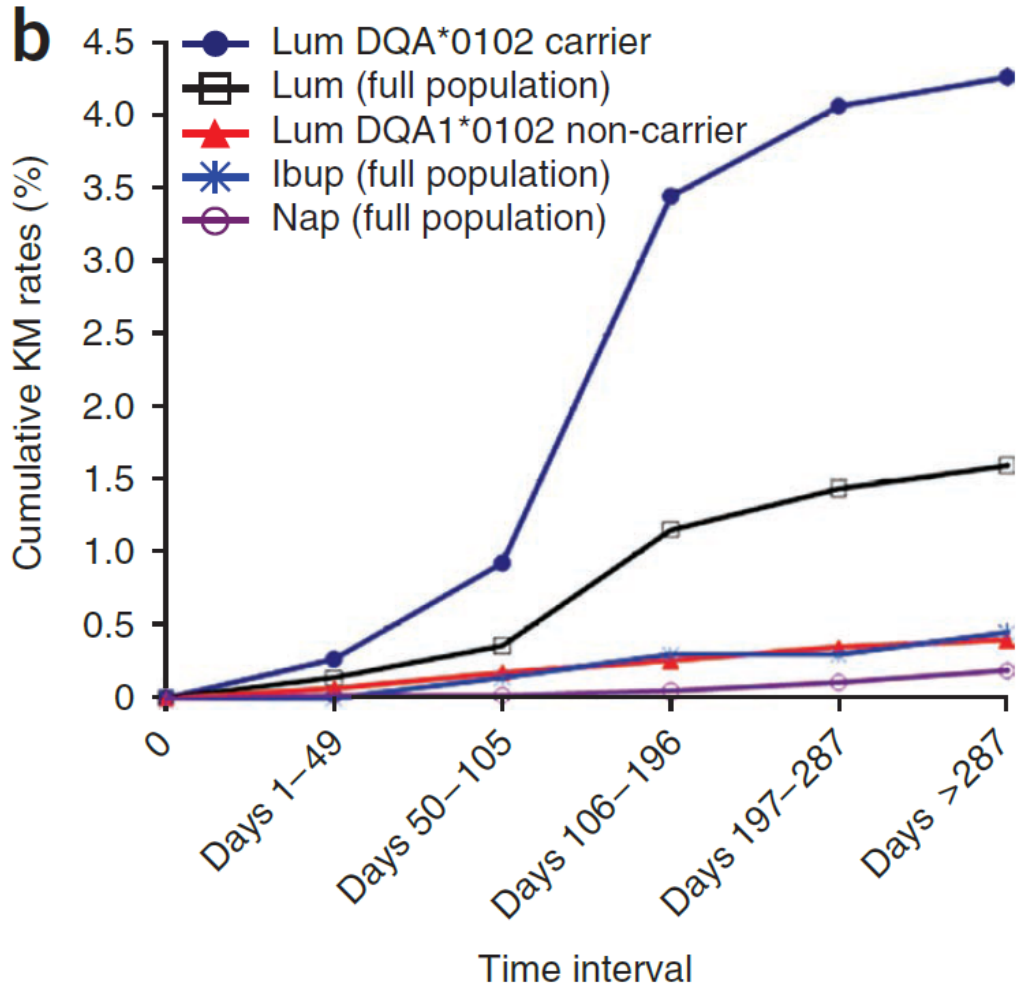


Figure 3: Cumulative Kaplan-Meier estimates for time to liver enzyme elevations $>$ 5 times ULN in the TARGET study (Lum = lumiracoxib, Ibup = ibuprofen, Nap = naproxen)

Table of Contents

Subgroups in terms of subsets of covariates

- Split the range of each covariate in two (or several) subsets
- Form subgroups from the intersections of the subsets
- Identify subgroups based on
  1. treatment effect differences within subgroups
  2. treatment by subgroup interactions (requires modeling)

x_2	$X_1 < x_1, X_2 < x_2$	$X_1 \geq x_1, X_2 < x_2$
	$X_1 < x_1, X_2 \geq x_2$	$X_1 \geq x_1, X_2 \geq x_2$
		x_1

The predicted individual treatment effect

For **controlled** studies, let $Y(z, \mathbf{x})$ be the outcome of a subject with covariates $\mathbf{x}$ and intervention $z \in \{0, 1\}$.

The **predicted individual treatment effect** (PITE) [7, 8] is defined by

$$D(\mathbf{x}) = g\{E[Y(1, \mathbf{x})]\} - g\{E[Y(0, \mathbf{x})]\}$$

for some link function g, e.g.,

- $g(y) = y$ for normal/log-normal outcomes (lab values)
- $g(y) = \text{logit}(y)$ for binary data (adverse events)

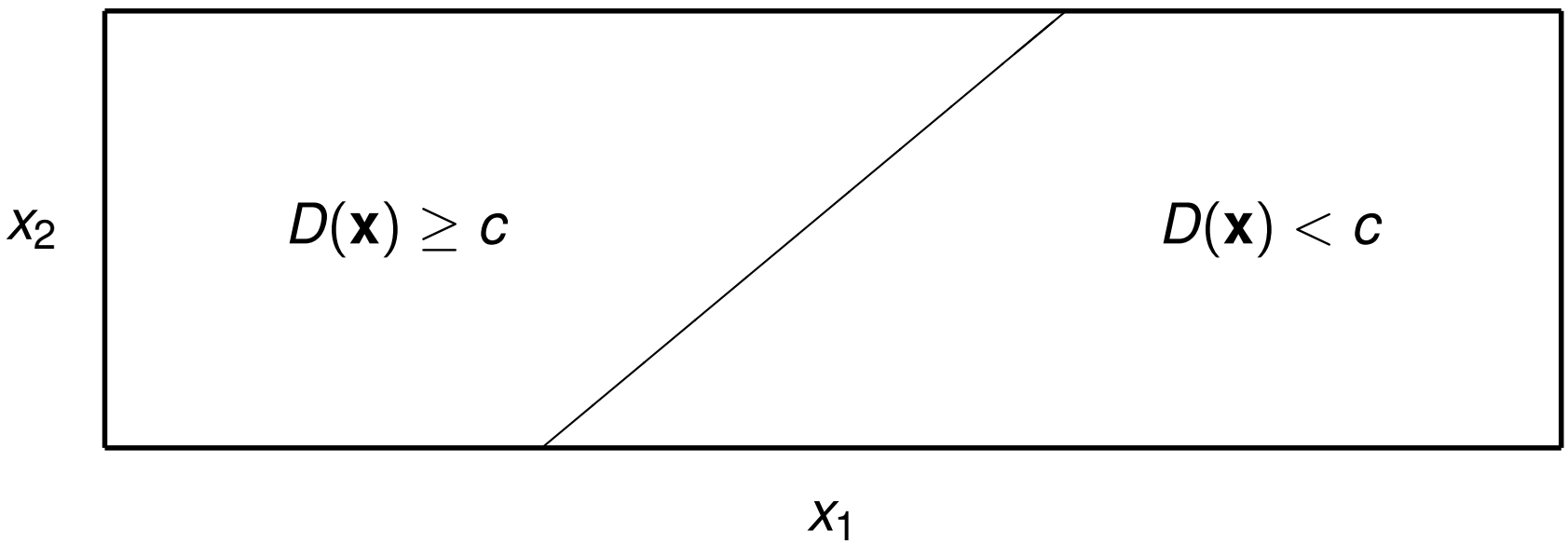
For **uncontrolled** studies let $Y(\mathbf{x})$ be the outcome of a subject with covariates $\mathbf{x}$ and set

$$D(\mathbf{x}) = g\{E[Y(\mathbf{x})]\}$$

Subgroups in terms of a PITE

- Define the appropriate predicted individual treatment effect $D(\mathbf{x})$ between test and control in a patient with covariates $\mathbf{x}$
- Identify patients with predicted treatment effect above a threshold $c > 0$ based on their covariates:

$$S(c) = \{\mathbf{x}; D(\mathbf{x}) \geq c\}$$



Comparison of options

Subsets of covariates based option

- Pros
  - Marginal thresholds easy to interpret
  - Comparisons within subgroups do not require modeling
- Cons
  - Potentially large number of subgroups
  - Requires dichotomization of continuous covariates
  - Decision driven by significance rather than relevance

PITE based option

- Pros
  - Defines subgroups in terms of relevant effects
  - Divides the covariance space into 2 subgroups only
- Cons
  - Requires modeling (like tests for interactions)
  - Requires the definition of a relevant threshold
  - Unintuitive relationship between covariates and predictions

Table of Contents

Renal safety of contrast media: CIN data

- Analysis on occurrence of CIN (dichotomized creatinine data) in the absence of creatinine measurements

Medium	CKD	DM	Patients	Events	Rate (%)
1	Y	Y	115	4	3.5
0	Y	Y	116	18	15.5
1	Y	N	247	6	2.4
0	Y	N	255	13	5.0
1	N	Y	178	1	0.6
0	N	Y	158	3	1.9
1	N	N	842	8	1.0
0	N	N	811	13	1.6

Table 2: Occurrence of contrast-induced nephropathy (CIN) defined as a rise in Cr $\geq$ 0.5mg/dl (Cr = creatinine, CKD = chronic kidney disease, DM = diabetes mellitus)

Analysis by subgroups

Subgroup	Events/patients	OR (95% CI)	*p*-value
All	66/2722	0.38 (0.22–0.66)	**0.0005**
CKD = Y	41/733	0.31 (0.15–0.65)	**0.0017**
CKD = N	25/1989	0.53 (0.23–1.21)	0.1301
DM = Y	26/567	0.29 (0.08–0.56)	**0.0019**
DM = N	40/2155	0.52 (0.27–1.00)	0.0511
DM = Y, CKD = Y	22/231	0.20 (0.06–0.60)	**0.0043**
DM = N, CKD = Y	19/502	0.46 (0.17–1.24)	0.1254
DM = Y, CKD = N	4/336	0.29 (0.03–2.84)	0.2885
DM = N, CKD = N	21/1653	0.59 (0.24–1.43)	0.2414

Table 3: Odds ratios of the risk of CIN under contrast medium 1 relative to medium 0. Odds ratios and confidence intervals identical to those in Table 4 of [2], *P*-values differ from those of Fisher's exact test which are presented in the paper.

Critique of analysis

Conclusion of the paper [2]

- “The largest difference (!) in the incidence of CIN found between patients given [medium 1] and those given [medium 0] was in subgroups with CKD or CKD+DM.”
- “Patient-related predictors of CIN were found to be CKD and CKD+DM, but not DM alone.”

- The authors did not provide an analysis on the subgroup defined by the presence or absence of DM
- Some authors [9, 10] prefer interaction tests over comparisons of within subgroup effects

Modeling

Analyzing interactions or predictive individual treatment effects requires modeling.

Let X_1, X_2 denote the indicator variables for DM and CKD, respectively, and let Z denote the treatment indicator and p the probability of CIN

$$\begin{aligned}\text{logit}(p(z,\mathbf{x})) = \alpha + \underbrace{\beta_1 x_1 + \beta_2 x_2 + \gamma x_1 x_2}_{\text{prognostic effects}} \\ + (\delta + \underbrace{\epsilon_1 x_1 + \epsilon_2 x_2 + \eta x_1 x_2}_{\text{predictive effects}})z \end{aligned} \tag{1}$$

Then the PITE is given by

$$D(\mathbf{x}) = \text{logit}\{p(1,\mathbf{x})\} - \text{logit}\{p(0,\mathbf{x})\} = \delta + \epsilon_1 x_1 + \epsilon_2 x_2 + \eta x_1 x_2$$

Predicted individual treatment effect (PITE)

Note that

- the PITE does not depend on the prognostic effects
- $\exp\{D(\mathbf{x})\} = \dfrac{p(1,\mathbf{x})[1-p(0,\mathbf{x})]}{[1-p(1,\mathbf{x})]p(0,\mathbf{x})} = \text{OR}(\mathbf{x})$

Let $\hat{\delta}$, $\hat{\epsilon}_i$ and $\hat{\eta}$ be the maximum likelihood (ML) estimators of δ, ϵ_i and η, respectively, then

$$\hat{D}(\mathbf{x}) = \hat{\delta} + \hat{\epsilon}_1 x_1 + \hat{\epsilon}_2 x_2 + \hat{\eta} x_1 x_2$$

is the ML estimator of $D(\mathbf{x})$.

Identify a subset $\hat{S}(c)$ of the covariate space with

$$\hat{S}(c) = \{\mathbf{x}; \hat{D}(\mathbf{x}) \geq c\}$$

for some clinically relevant $c > 0$

Testing within subgroups versus testing for interaction

Testing within subgroup effects

- $H_{00} : D((0,0)) = \delta = 0$
- $H_{01} : D((0,1)) = \delta + \epsilon_1 = 0$
- $H_{10} : D((1,0)) = \delta + \epsilon_2 = 0$
- $H_{11} : D((1,1)) = \delta + \epsilon_1 + \epsilon_2 + \eta = 0$

Testing interactions

- $K_{10,00} : D((1,0)) - D((0,0)) = \epsilon_1 = 0$
- $K_{01,00} : D((0,1)) - D((0,0)) = \epsilon_2 = 0$
- $K_{11,00} : D((1,1)) - D((0,0)) = \epsilon_1 + \epsilon_2 + \eta = 0$
- $K_{11,01} : D((1,1)) - D((0,1)) = \epsilon_1 + \eta = 0$
- $K_{11,10} : D((1,1)) - D((1,0)) = \epsilon_2 + \eta = 0$

Estimates of model parameters and contrasts

Parameter	Estimate	Stderr	*p*-value
α	-4.1172	0.2796	$< .0001$
β_1	0.1724	0.6465	0.7898
β_2	1.1932	0.3990	**0.0028**
γ	1.0572	0.7515	0.1595
δ	-0.5296	0.4521	0.2414
ϵ_1	-0.7013	1.2447	0.5731
ϵ_2	-0.2394	0.6755	0.7230
η	-0.1585	1.4581	0.9135
$\delta + \epsilon_1$	-1.2309	1.1597	0.2885
$\delta + \epsilon_2$	-0.7690	0.5019	0.1254
$\delta + \epsilon_1 + \epsilon_2 + \eta$	-1.6288	0.5699	**0.0043**
$\epsilon_1 + \eta$	-0.8598	0.7594	0.2575
$\epsilon_2 + \eta$	-0.3979	1.2922	0.7582
$\epsilon_1 + \epsilon_2 + \eta$	-1.0992	0.7274	0.1308

Table 4: Parameter/contrast estimates for the model (1) with standard errors and *p*-values for the hypotheses that a parameter/contrast is zero.

Results based on the PITE

Decision rule

Contrast medium 1 is to be preferred over medium 0 in patients with an observed reduction of the (relative) CIN risk by 50% or more

DM	CKD	OR($\mathbf{X}$) (95% CI)	≤ 0.5	OR($\mathbf{X}$)/OR($\mathbf{0}$) (95% CI)	≤ 0.5
1	1	0.20 (0.06–0.60)	Y	0.50 (0.04–5.69)	Y
0	1	0.46 (0.17–1.24)	Y	0.79 (0.21–2.96)	N
1	0	0.29 (0.03–2.84)	Y	0.33 (0.08–1.39)	Y
0	0	0.59 (0.24–1.43)	N	—	—

Table 5: Odds ratios of the (relative) risk of CIN under contrast medium 1 relative to medium 0 from Table 3; $\text{OR}(\mathbf{X}) = \exp\{\hat{D}(\mathbf{X})\}$

- According to the decision rule based on OR($\mathbf{x}$), all patients with DM or CKD or both should be given contrast medium 1 (Table 5).
- However, according to the rule based on OR($\mathbf{x}$)/OR($\mathbf{0}$), only patients with DM alone or in combination with CKD should be given medium 1.

Stabilizing the PITE rule

Uncertainty and instability of the decision rules can be accounted for by bagging (bootstrap aggregating) [11]:

1. Let $N(z, \mathbf{x})$ the number of subjects and $n(z, \mathbf{x})$ the number of subjects with CIN for $Z = z$ and $\mathbf{X} = \mathbf{x}$.
2. For $b = 1, \ldots, B$, draw $n_b^*(z, \mathbf{x}) \sim \text{Binom}(n(z, \mathbf{x})/N(z, \mathbf{x}), N(z, \mathbf{x}))$ and calculate

$$\text{OR}_b^*(\mathbf{x}) = \frac{n_b^*(1, \mathbf{x})[N(0, \mathbf{x}) - n_b^*(0, \mathbf{x})]}{[N(1, \mathbf{x}) - n_b^*(1, \mathbf{x})]n_b^*(0, \mathbf{x})}$$

3. Contrast medium 1 is preferred if

$$P^*(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} I[\text{OR}_b^*(\mathbf{x}) \leq 0.5] \geq \frac{1}{2} \quad \text{or}$$

$$Q^*(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} I[\text{OR}_b^*(\mathbf{x})/\text{OR}_b^*(\mathbf{0}) \leq 0.5] \geq \frac{1}{2}$$

Results from bagging

DM	CKD	$P^*(X)$	$Q^*(X)$
1	1	**0.9527**	**0.7017**
0	1	**0.5308**	0.2523
1	0	**0.6370**	0.4548
0	0	0.3269	—

Table 6: Proportions of bootstrap samples leading to a preference of contrast medium 1 ($B = 10000$).

Remark: For the third subgroup we have 1/178 for medium 1 and 3/158 events for medium 0. In case of 2/158 instead we would obtain $\text{OR}(1, 0) = 0.44 < 0.5$ but also $P^*(1, 0) = 0.4683 < 0.5$, i.e., the preference for medium 1 suggested by the original data would not be supported after bagging.

Table of Contents

Simulation specifications

Z	X_1	X_2	Pr[event]	odds ratio
1	1	1	0.100	—
0	1	1	0.050	2.111
1	1	0	0.075	—
0	1	0	0.050	1.541
1	0	1	0.060	—
0	0	1	0.050	1.213
1	0	0	0.050	—
0	0	0	0.050	1.000

Table 7: Simulation parameters

- 10% significance level for test for difference within subgroups
- 10% significance level for interaction test
- Prefer treatment 0 over 1 if $\exp\{D(\mathbf{X})\} \geq 1.2$
- Prefer treatment 0 over 1 if $\exp\{D(\mathbf{X}) - D(\mathbf{0})\} \geq 1.2$

Simulation results

n	x_1	x_2	sig diff	sig inter	$P^*(\mathbf{x}) \geq 0.5$[1]	$Q^*(\mathbf{x}) \geq 0.5$[2]
100	1	1	0.3524	0.2006	0.8432	0.7404
100	1	0	0.1572	0.1130	0.6530	0.6122
100	0	1	0.0892	0.0864	0.4930	0.5014
100	0	0	0.0754	—	0.3822	—
100	1	1	0.3524	0.3078	0.8432	0.8316
100	1	0	0.1572	0.1412	0.6530	0.6562
100	0	1	0.0892	0.0880	0.4930	0.5006
1000	0	0	0.1046	—	0.2002	—

Table 8: Proportion of 5000 simulations where subjects with covariates $\mathbf{x}$ are denied treatment 1 for parameters in Table 7, $B = 200$ bootstrap draws.

[1] $P^*(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} I[\exp\{\hat{D}_b^*(\mathbf{x})\} \geq 1.2]$

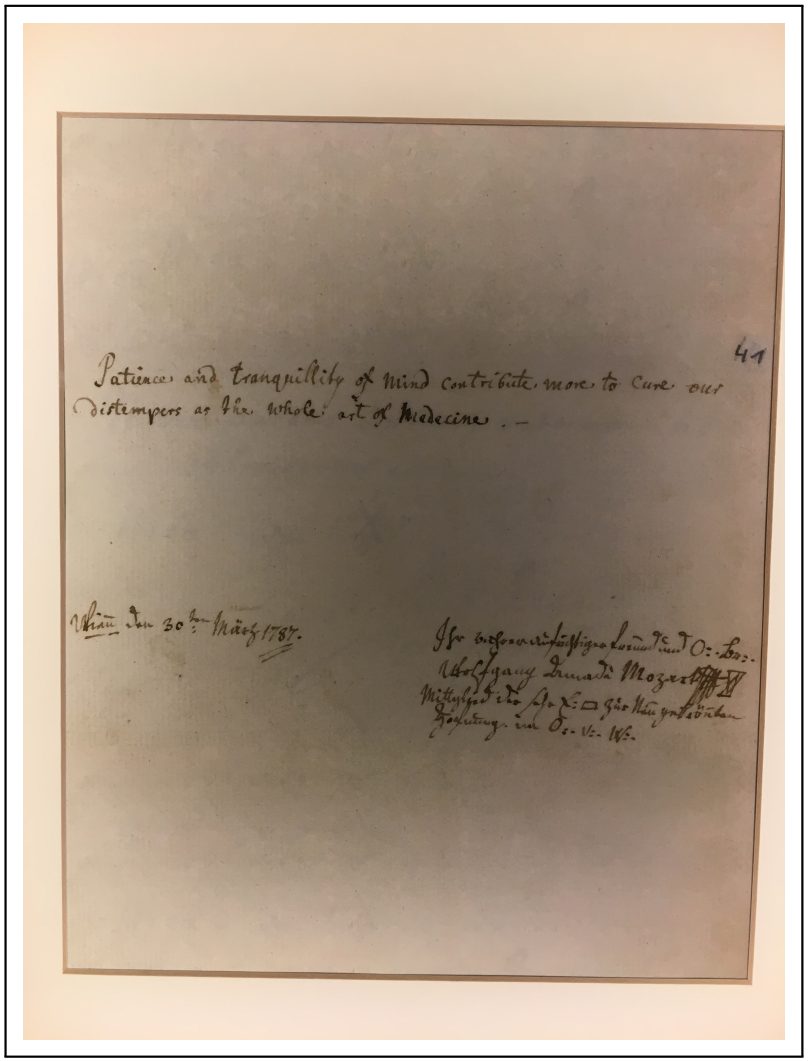
[2] $Q^*(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} I[\exp\{\hat{D}_b^*(\mathbf{x}) - \hat{D}_b^*(\mathbf{0})\} \geq 1.2]$

Table of Contents

Concluding remarks

- Personalized safety is in it's infancy
- Focus on specific safety aspects rather than safety picture at large
- Analysis methods to be selected thoughtfully
  - Significance tests for within group differences may be conservative and difficult to interpret
  - Interaction test suffer even more from lack of power
  - Effect size based selection rules have a high detection chance at the expense of a high false positive rate
- Modeling in the presence of many covariates challenging
  - Sparse modeling is a must for adequate prediction precision and interpretability of results
  - Issues of post selection inference [12]
- Extent of corroboration of results is low for predictive efficacy [13], not expected to be better for safety

Some advice



“Patience and tranquility of mind contribute more to cure our distempers as the whole art of medicine”

Vienna, March 30, 1787
Wolfgang Amadeus Mozart

References I

[1] FDA.
Paving the way for personalized medicine.
Food and Drug Administration, 2013.

[2] P A McCullough, M E Bertraud, J A Brinker, and F Stacul.
A meta-analysis of the renal safety of isomolar iodixanol compared with low-osmolar contrast media.
Journal of the American College of Cardiology, 48:692–699, 2006.

[3] P N Sylaja, W Dong, J C Grotta, M K Miller, K Tomita, S Hamilton, C Semba, and M D Hill.
Safety outcomes of alteplase among acute ischemic stroke patients with special characteristics.
Neurocritical Care, 6:181–185, 2007.

[4] The NINDS t-PA Stroke Study Group.
Intracerebral hemorrhage after intravenous t-pa therapy for ischemic stroke.
Stroke, 28:2109–2118, 1997.

[5] M E Farkouh, H Kirshner, R A Harrington, S Ruland, F W A Verheugt, T J Schnitzer, G T Burmester, E Mysler, M C Hochberg, M Doherty, E Ehrsam, X Gitton, G Krammer, B Mellein, A Gimona, P Matchaba, C J Hawkey, and J H Chesebro on behalf of the TARGET Study Group.
Comparison of lumiracoxib with naproxen and ibuprofen in the Therapeutic Arthritis Research and Gastrointestinal Event Trial (TARGET), cardiovascular outcomes: randomised controlled trial.
Lancet, 364:675–684, 2004.

[6] J B Singer, S Lewitzky, E Leroy, F Yang, X Zhao, L Klickstein, T M Wright, J Meyer, and A Paulding.
A genome-wide study identifies HLA alleles associated with lumiracoxib-related liver injury.
Nature Genetics, 42:711–716, 2010.

[7] T Cai, L Tian, P H Wong, and L J Wei.
Analysis of randomized comparative clinical trial data for personalized treatment selections.
Biostatistics, 12:270–282, 2011.

[8] S Chen, L Tian, T Cai, and M Yu.
A general statistical framework for subgroup identification and comparative treatment scoring.
Biometrics, 2017.

References II

[9] S F Assmann, S J Pocock, L E Enos, and L E Kasten.
Subgroup analyses and other (mis)uses of baseline data in clinical trials.
Lancet 2000, 355:10641069.

[10] S T Brookes, E Whitely, M Egger, G D Smith, P A Mulheran, and T J Peters.
Subgroup analysis in randomized trials: risks of subgroup-specific analyses: Power and sample size of interaction test.
Journal of Clinical Epidemiology, 57:229–236, 2004.

[11] L Breiman.
Bagging predictors.
Machine Learning, 24:123–140, 1996.

[12] R J Tibshirani, J Taylor, R Lockhart, and R Tibshirani.
Exact post-selection inference for sequential regression procedures.
Journal of the American Statistical Association, 111:600–620, 2016.

[13] J D Wallach, P G Sullivan, J F Trepanowski, K L Sainani, E W Steyerberg, and J P A Ioannidis.
Evaluation of evidence os statistical support and corroboration of subgroup claims in randomized clinical trials.
JAMA Internal Medicine, 177:554–560, 2017.