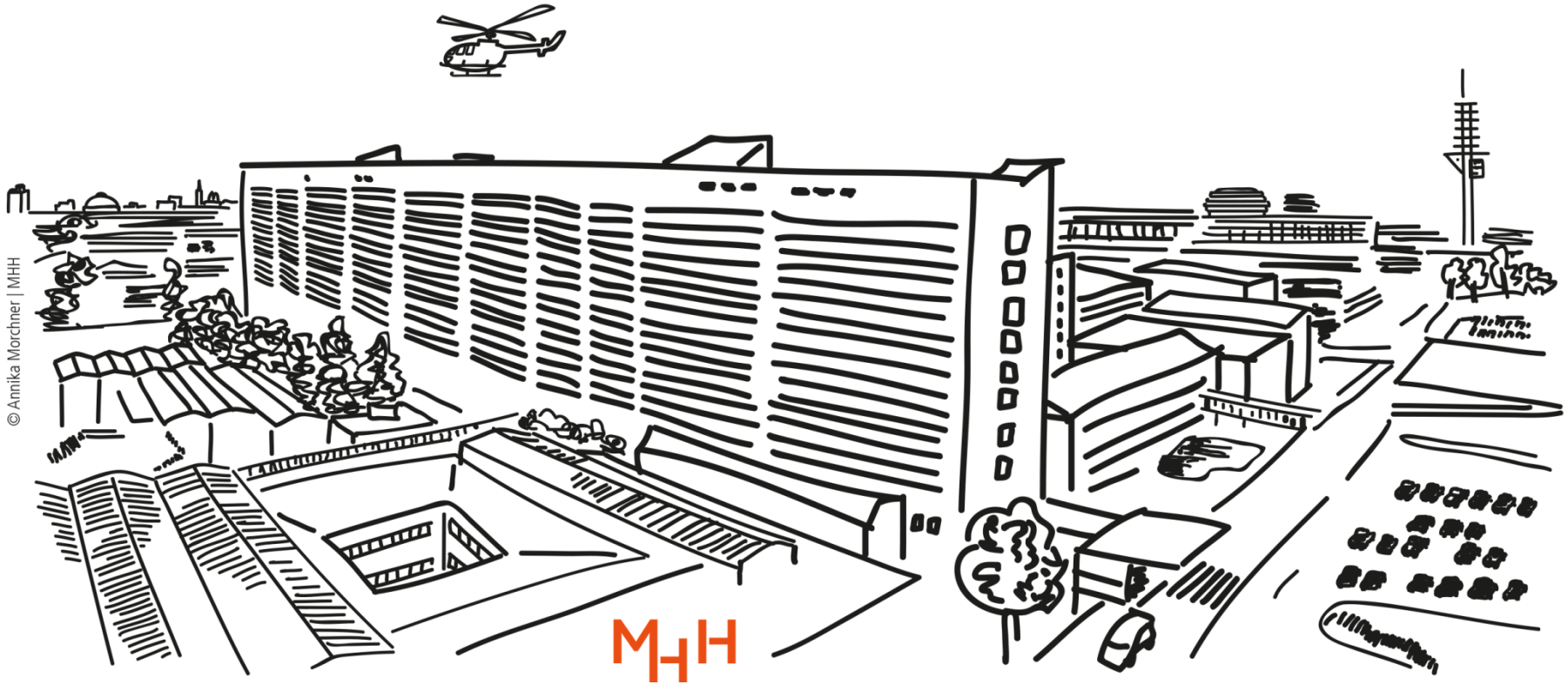
Significance, relevance, and a proportionate amount of evidence for decision making in drug licensing

Armin Koch



Communalities / where we met

We

- both came from math, moved to stats and worked in drug licensing / regulation,
- first met on a case in my former life (and actually also Uli's former life), where biostats could really contribute to decision making,
- discussed thereafter on multiple occasions (APF, DIA, EFSPI),
- met again during the EU-funded ASTERIX-project, where Kit Roes found industry view for an academic project of relevance,
- contributed to the EFSPI-stats workshop, which now is an established platform for dialog between stakeholders,
- and joined forces during the first round of discussions about ICH-E20 in the "defense against the dark arts".

This presentation is about the specifics of biostatistics in drug regulation

Biostatistics in drug regulation

Coming from Mathematics to Biostats in drug regulation, you realize:

- “significant” is something really important,
- but beyond that “clinical relevance” is of equal importance,
- biostatistical discussions in drug regulation are on a very high level because
  1. we are experimenting with patients,
  2. we share a responsibility for future patients in need of treatment,
  3. clinical trials are horribly expensive,
  4. there is a legal background,
- Clarity is urgently needed in the multi-disciplinary context,

Disclaimers:

- I am speaking about Phase III clinical trials (late stage drug development).
- For risks and adverse effects ask your preferred regulator (or your senior manager).
- Opinions are mine and do not necessarily reflect the institutions, where I (try to) contribute to decision making.
- “Defense against the dark arts” reflects my fight against being imprecise, lenient, or steeling „evidence“ ☺

The mandate / What statisticians should know about drug regulation

German Drug Law (Arzneimittelgesetz):

§ 25 Entscheidung über die Zulassung

(2) Die zuständige Bundesoberbehörde darf die Zulassung nur versagen, wenn

1. die vorgelegten Unterlagen, einschließlich solcher Unterlagen, die auf Grund einer Verordnung der Europäischen Gemeinschaft oder der Europäischen Union vorzulegen sind, unvollständig sind,
2. das Arzneimittel nicht nach dem jeweils gesicherten Stand der wissenschaftlichen Erkenntnisse ausreichend geprüft worden ist oder das andere wissenschaftliche Erkenntnismaterial nach § 22 Abs. 3 nicht dem jeweils gesicherten Stand der wissenschaftlichen Erkenntnisse entspricht,
3. das Arzneimittel nicht nach den anerkannten pharmazeutischen Regeln hergestellt wird oder nicht die angemessene Qualität aufweist,
4. dem Arzneimittel die vom **Antragsteller** angegebene therapeutische Wirksamkeit fehlt oder diese nach dem jeweils gesicherten Stand der wissenschaftlichen Erkenntnisse vom **Antragsteller** unzureichend begründet ist,
5. das Nutzen-Risiko-Verhältnis ungünstig ist,

Regulators do not license drugs!

… not tested according to scientific standards

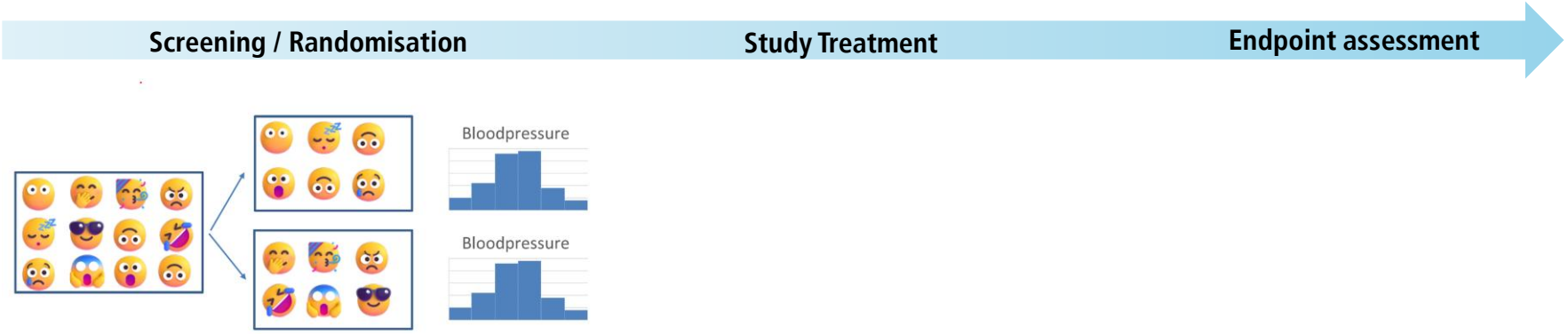
… efficacy is missing

… or insufficiently justified according to scientific standards

… benefit/risk is unfavourable

For reference to EC-Legislation see:
Guideline on the investigation of subgroups in confirmatory clinical trials

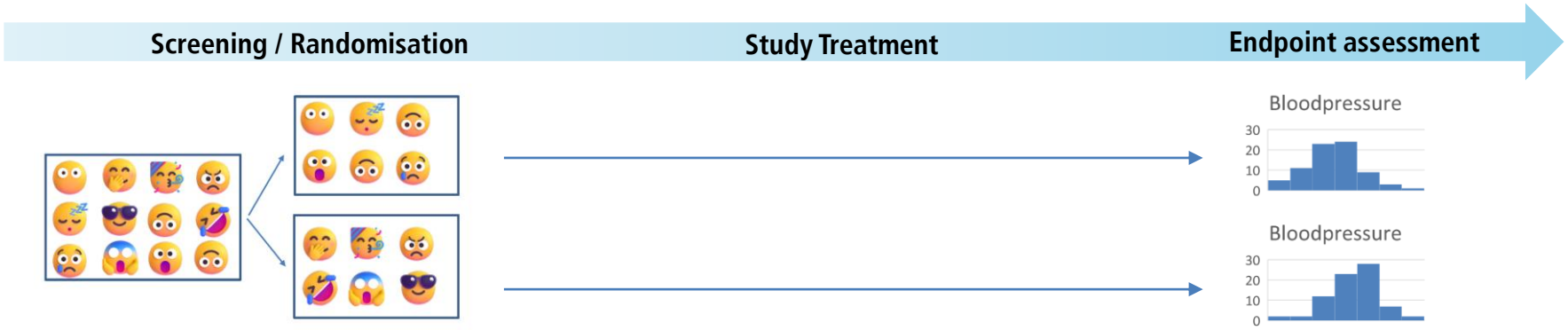
The decision model and prerequisites



Randomization

- generates balanced groups regarding all known / measured but also unknown / unmeasured risk factors *at baseline*,
- allows probability statements about differences in a measured *baseline* variables between randomized groups (e.g. „large differences are unlikely“),
- enables quantification of what „large“ and „unlikely“ should mean,
- doesn‘t safeguard against differential post-baseline interventions.

The decision model and prerequisites



If

we wish to conclude from a large group difference in the primary endpoint *at the end* of the trial that the drug is no placebo / likely efficacious, we assume

- patients have been properly randomized,
- all patients have been observed and treated the same and experimental or control treatment is the only difference between groups,
- a pre-specified plan to decide about study success (Christopher Colon).

Frequentist decision making:

If:

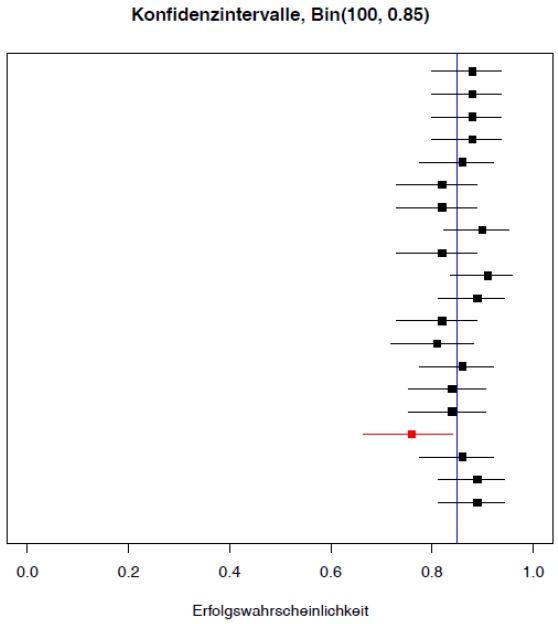
- the experimental drug E is as efficacious as placebo P (i.e. H_0: $p_E = p_P$), and
- You are willing to accept a proportion of 2,5% false positive decisions, (like with diagnostics, you have to accept some false positives), and
- You conduct 100 identical valid studies/experiments and analyze them with appropriate statistical methodology,

then

- you would see „large difference / significance“ in only 2,5% of the cases.
- we call this „unlikely“ and if we see such a larger „significant“ difference between treatment groups, we tick:

✓ formal proof of efficacy demonstrated.

Frequentist decision making



but:

- we usually don‘t do / don‘t have 100 identical studies
- we just have one.

It is just a model which we use to support decision making!

We might have chosen other models:

- we license, if the drug tastes well,
- we license, if the drug doesn‘t harm,
- we license if the treatment effect has the right sign,

… but we decided for this one (and went even further).

Study-wise type-1-error

Statistician's mandate is control of the type-1-error ☺:

"Throughout this document the term 'control of type I error' rate will be used as an abbreviation for the control of the family-wise type I error in the strong sense, i.e., there is control on the probability to reject at least one true null hypothesis, regardless which subset of null hypotheses happens to be true."

Statisticians speak about "families of hypotheses", but leave it to others to define, what family should mean.

However, in 2002 Joachim Röhmel and others made an important decision:

„Control of the study-wise rate of false positive conclusions at an acceptable level alpha is an important principle and is often of great value in the assessment of confirmatory clinical trials."

(CPMP/EWP/908/99)

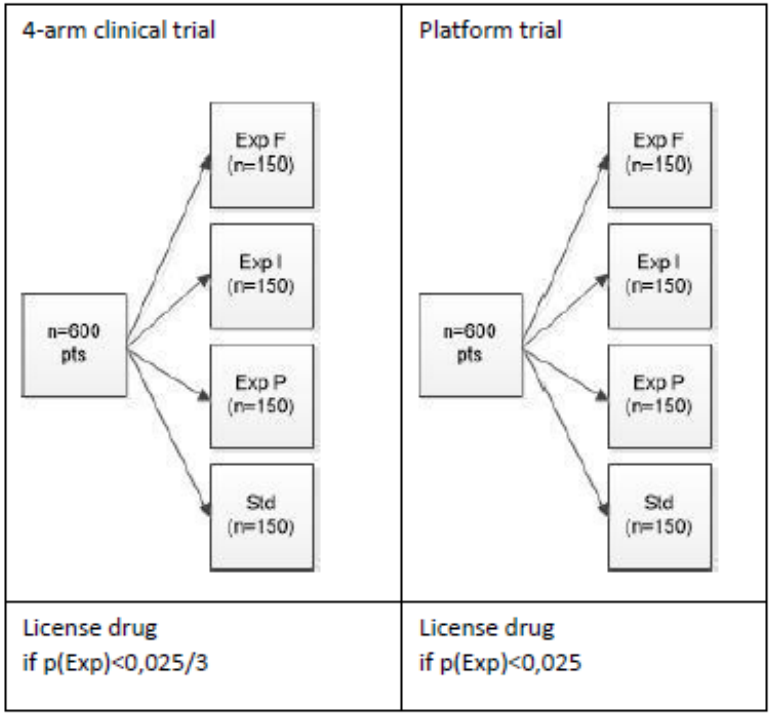
Thinking from the perspective of questions posed to "the study" defined by eligible patients being randomly allocated to a number of treatments is the most scientific principle guiding decision making in the presence of multiplicity.

Recent example raised discussion:

Three badly recruiting trials in a rare condition motivated one of the agencies to propose the conduct of a multi-arm clinical trial,

however,

with just changing the label of the trial applicants justified changes to the assessment rules for formal proof of efficacy / licensing.



Everybody is right when recommending multi-arm clinical trials
(not only in rare disease)

Everybody knows:

- A smaller overall sample-size is needed for the multi-arm trial than for the conduct of three independent trials (even with Bonferroni-adjustment for multiple comparisons in the multi-arm trial),

More importantly:

- The number of patients randomized to a supposedly inferior standard is substantially smaller.

Ethics, science and costs speak for a multi-arm trial!

	One 4-arm-trial for licensing 3 drugs	Three 2-arm-trials for licensing 3 drugs
Study-wise T1E (one-sided)	2.5%	2.5%
Comparison-wise T1E (one-sided)	0.83%	2.5%
N per group	185	138
Total N	740 (=185*4)	828 (=138*6)
Total N controls	185	414 (=138*3)

(p_{exp}=0.35; p_{std}=0.20, Pow=80%)

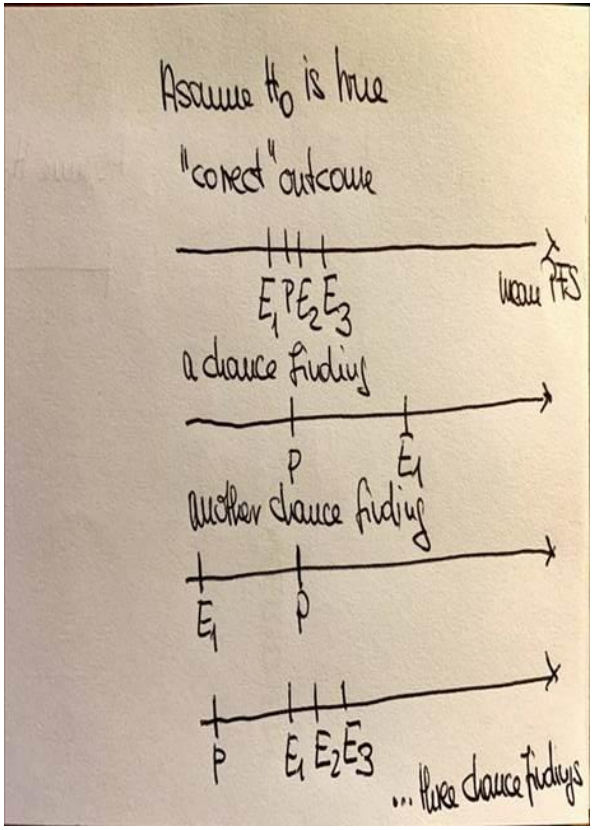
Type-I-error -
The probability of an unlucky landing…

If we feel that we can do multi-arm multi-drug clinical trials w/o adjustment, chance is sweating ☺:

- Chance has to balance 4 groups (obviously more difficult than just balancing two groups).
- As we re-use the placebo-group for > 1 decision making process, the study-wise T1E may not be controlled at the pre-specified level.
- … just as a consequence of the “unlucky landing” of the PBO-group

Fair to say:

- … chance may play against 3 effective drugs,
- … this is not a T1E-issue, but affects power,
- … or, if you factor this into your considerations, then you follow a different concept for T1E-control.



My recommendation:

Implicitly lowering the hurdle (i.e. agreeing to not adjust for multiplicity) is dangerous.

Instead, we should stick to our metrics and discuss alleviations openly:

- if ever 3 applicants come together to do a multi-arm trial, we grant them a study-wise T1E of 7,5% (one-sided)
- if this is so rare, we grant you a T1E of 5% (one-sided)

… but thereafter, we apply statistics in the “normal” way.

ICH-E17:
No adjustment needed 

ICH-E17 did open-up for having multiple trial successes without proper „payment“ in statistical terms:

In this case, because regulatory approvals are based on different primary endpoints by different authorities, no multiplicity adjustment is needed for regulatory decision-making.”

ICH-E17:
No adjustment needed 😠

ICH-E17 did open-up for having multiple trial successes without proper „payment“ in statistical terms:

If agreement cannot be reached due to well-justified scientific or regulatory reasons, a single protocol should be developed with endpoint-related sub-sections tailored to meet the respective requirements of the regulatory authorities. In this case, because regulatory approvals are based on different primary endpoints by different authorities, no multiplicity adjustment is needed for regulatory decision-making.”

ICH-E17:
No adjustment needed 😡

ICH-E17 did open-up for having multiple trial successes without proper „payment“ in statistical terms:

“Agreement on the primary endpoint ensures that the overall sample size and power can be determined for a single (primary) endpoint based on the overall population and also agreed upon by the regulatory authorities. If agreement cannot be reached due to well-justified scientific or regulatory reasons, a single protocol should be developed with endpoint-related sub-sections tailored to meet the respective requirements of the regulatory authorities. In this case, because regulatory approvals are based on different primary endpoints by different authorities, no multiplicity adjustment is needed for regulatory decision-making.”

Learnings:

- „The opposite of „good“ is not „bad“, the opposite is „well intentioned“

  (Kurt Tucholsky, 1890 – 1935)
- Never write „*An ideal clinical trial endpoint is one…”*
- … and the only excuse is that well planned MRCTs are so informative!

ICH-E17:
No adjustment needed 😠

Logic is a problem:

If EMA is interested in two endpoints, they need to be co-primary; if EMA and FDA are interested in different endpoints each, we don‘t care?

- will chance care for even better balance so that the number of false positive conclusions remains the same?
- Captain speaking: „Please be advised that patients in need of secondary prevention for MI should revert to Clopidogrel after landing at IAD because Tigacrelor is not effective in the US“
- in casu pro reo: are we sure that the respective other agency always asks for irrelevant stuff?

More seriously:

- instead of attempting to find a mutual understanding (or paying the prize for no agreement by making endpoints co-primary (and increase power to 90%)
- we all just accept the lowering of the hurdle.

Finis (-2):
We became a bit easy…

P-values based on a randomized clinical trial are called „experimental“ and if calculated from a proper experiment with appropriate methodology will control the probability of a false positive conclusion.

Tons of P-values are calculated every minute, but most are not experimental:

- Single arm trials (no experiment, selectionism difficult to control): P_{SAD}
- Meta-Analyses (type-1-error usually already exhausted): P_{MA}
- Real World Evidence (Observational studies for treatment comparisons leave opportunity that other reasons (beyond chance) cause the “effect”): P_{RWE}
- Bayesian Statistics: $P_{🙏}$

… we pretend they are all the same

Finis (-1):
Biostatistics in drug development

- is embedded in a legal context for decision making,
  - EC-Guidance is not law, however, deviations require justification,
  - EC-Guidance binds regulators, as well,
  - EC-Guidance is supposed to speed-up development and assessment.
- should help to increase clarity, not complexity,
- should not be an expert opinion, but have a say in primary decision making,
- is not only about control of the T1E, but to care that *results are interpretable if the null-hypothesis is rejected*,
- should care that planning, analysis and interpretation of clinical trials are aligned with what is needed for decision making about licensing,
- should defend (if anything) the randomized clinical trial and evidentiary standards of their conduct.

So: if there is a lowering of „the hurdle“, it should be explicit and not implicit.

Finis:

Working with Uli has been very rewarding exactly for the same reasons:

- give thought to the detail,
- not always searching for the minimum (to achieve, to agree, to aim for),
- discussing „a proportionate amount of evidence for decision making“,
- respecting that we plan for formal study success, but someone has to decide that benefit/risk is positive (and may need a bit more),
- always trying to find an independent opinion first before shopping around, what others are thinking.

Thank you for working with us this way!

… and if ever you are bored, just let me know ☺