Estimands in the context of External Comparator (EC) Studies

Or: Do we need more estimand attributes?

Dr. Gerd Rippin, Senior Director Biostatistics, Real World Solutions, IQVIA

Basel, Oct 27th , 2023

Agenda

- Introduction to External Comparator Cohort studies
- Revisiting the 5 estimand attributes
- Applying the 5 estimand attributes to ECs
- Do we need further attributes?
- Conclusion
- Q&A

Background

- Increased use of single-arm trials (SATs) in regulatory (and payer) decision making in recent years has driven the need to contextualize outcomes

- Real-world external comparators (ECs) obtained from data derived from standard care have been utilized to provide this context

- Randomization to standard of care or placebo in some situations may be either impractical and/or unethical (e.g., rare disease)

- Increasing use of the EC approach, **typically in cases of a *large* expected treatment effect* differences**, where residual bias after statistical adjustment is not considered to be critical. No intention at all to replace the proven RCT gold standard design

- Limited guidance established to date by regulators specific to ECs until recently, when FDA issued a draft guideline (Feb 2023)

RCT: Randomized Controlled Trial, FDA: US Food and Drug Administration

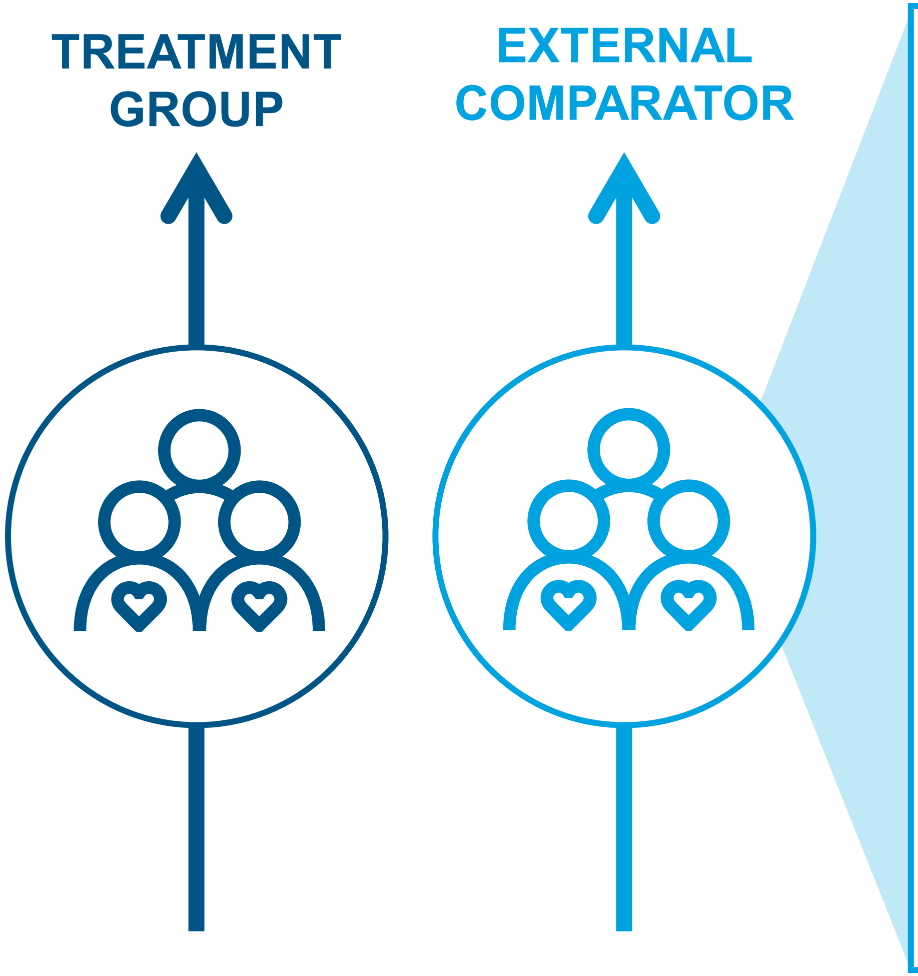
*:U.S. Food and Drug Administration (2018). Framework for FDA's real-world evidence. https://www.fda.gov/media/120060/download

An external comparator can help to establish context

INTERNAL CONTROL

Not feasible because:

- Unethical to randomize
- Ultra-rare indication
- Patients unwilling to participate due to risk of not receiving drug



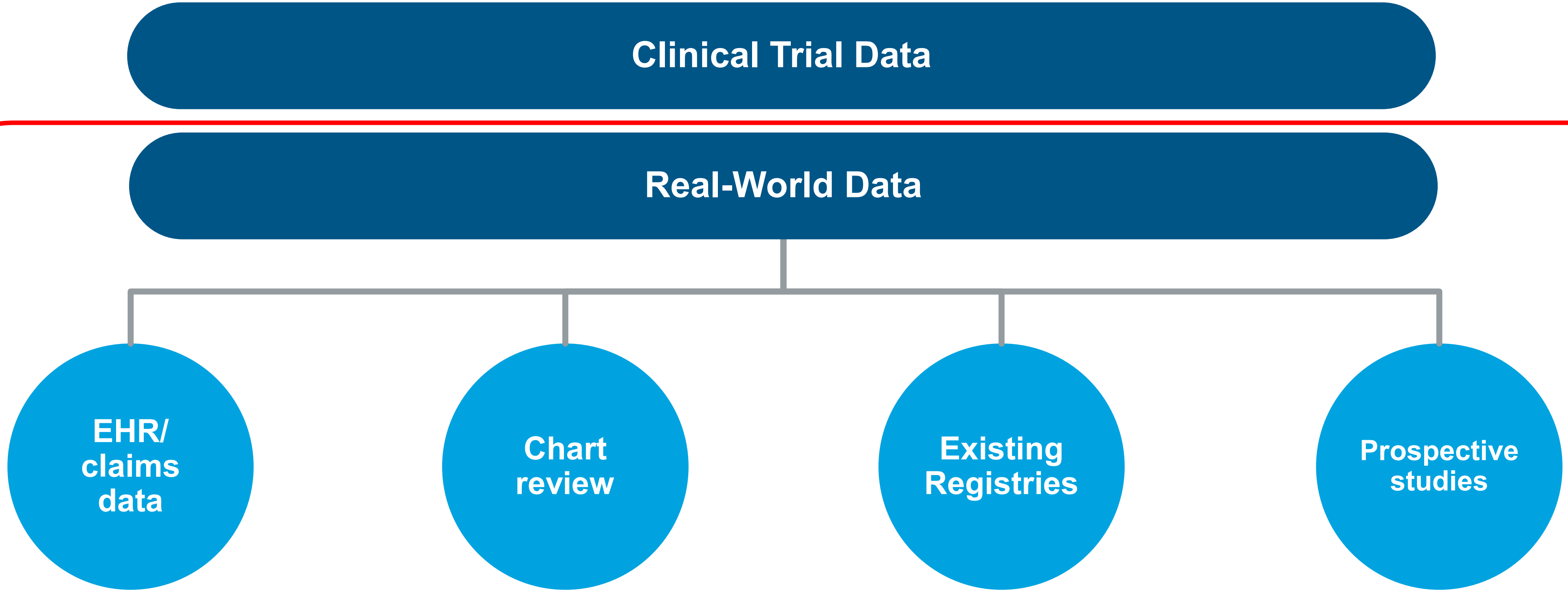
TREATMENT GROUP

EXTERNAL COMPARATOR

- Patient cohort derived from real-world data (RWD) or historical trials
- Patients selected according to the trials' main inclusion / exclusion criteria
- Compare the outcomes from the treatment group to a cohort of similar patients

RWD: real-world data

Main data sources for External Comparators*



Clinical Trial Data

Real-World Data

EHR/ claims data

- Curated data
- Variables and data recorded for administrative purposes
- Limited to data recorded

Chart review

- Can be conducted across countries/regions
- Retrospective design
- More flexibility in data than EHR/claims but limited to SoC assessments

Existing Registries

- Often include disease specific outcomes
- Often single-country/region
- Requires agreements with data stewards

Prospective studies

- Expensive/ long duration
- High control over design and outcomes assessed
- Can be conducted across countries/regions
- May run concurrently to a clinical trial

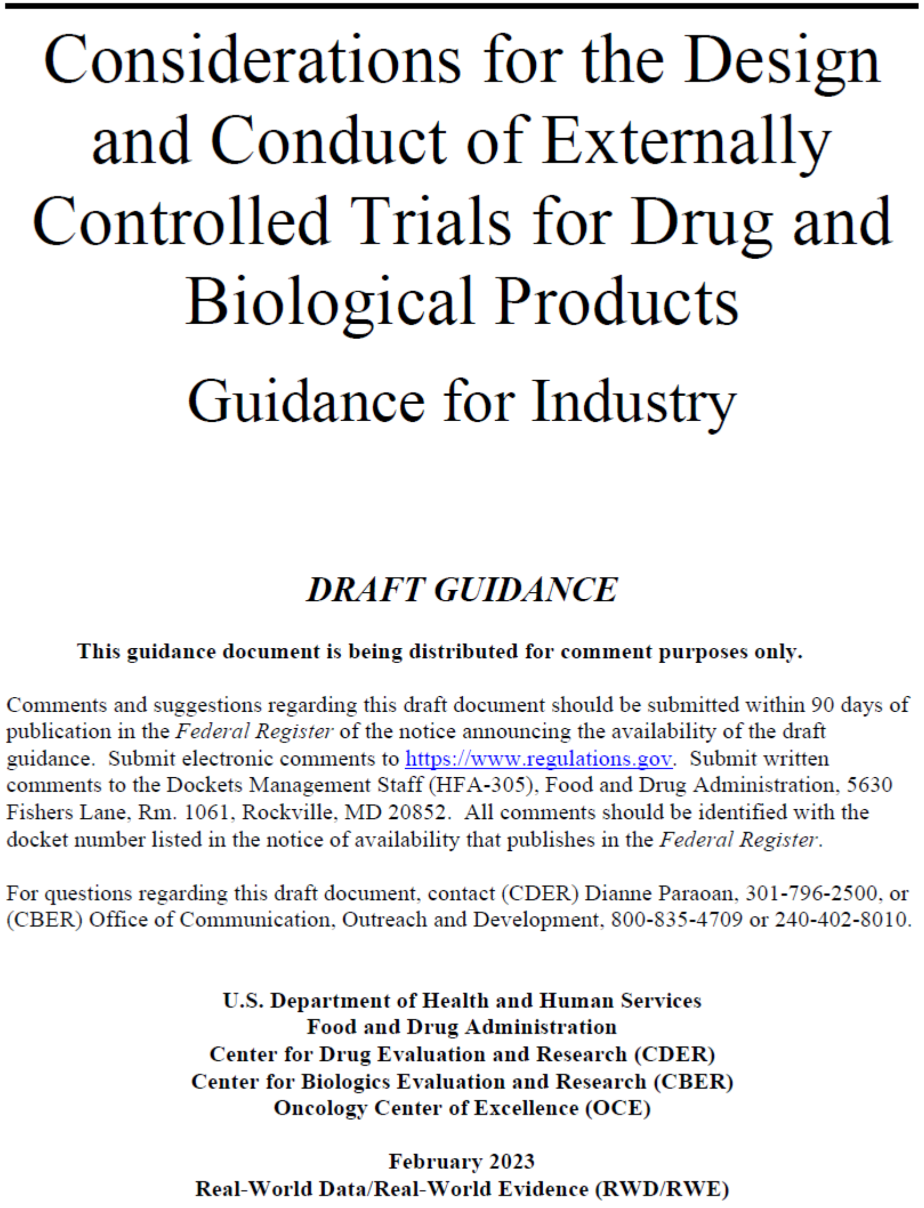
*Excluding literature reviews, these are not labelled as external comparators

SoC: standard of care; EHR: electronic health records

New FDA draft guidance

- Issued in Feb 2023
- IQVIA provided comments to the FDA in May 2023
- Topics of the guidance include:
    - Pre-planning the analysis
    - Estimands & intercurrent events
    - Differential exposure, diagnostics and other characteristics
    - Blinded validation of endpoints like progression to be considered
    - Missing data, measurement error and misclassification
    - Other topics
    - No specific recommendations about specific statistical approaches

- Local Health Technology Agency (HTA) requirements should be consulted in addition, when there is a HTA purpose for the ECA study



Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products

Guidance for Industry

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 90 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff (HFA-305), Food and Drug Administration, 5630 Fishers Lane, Rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

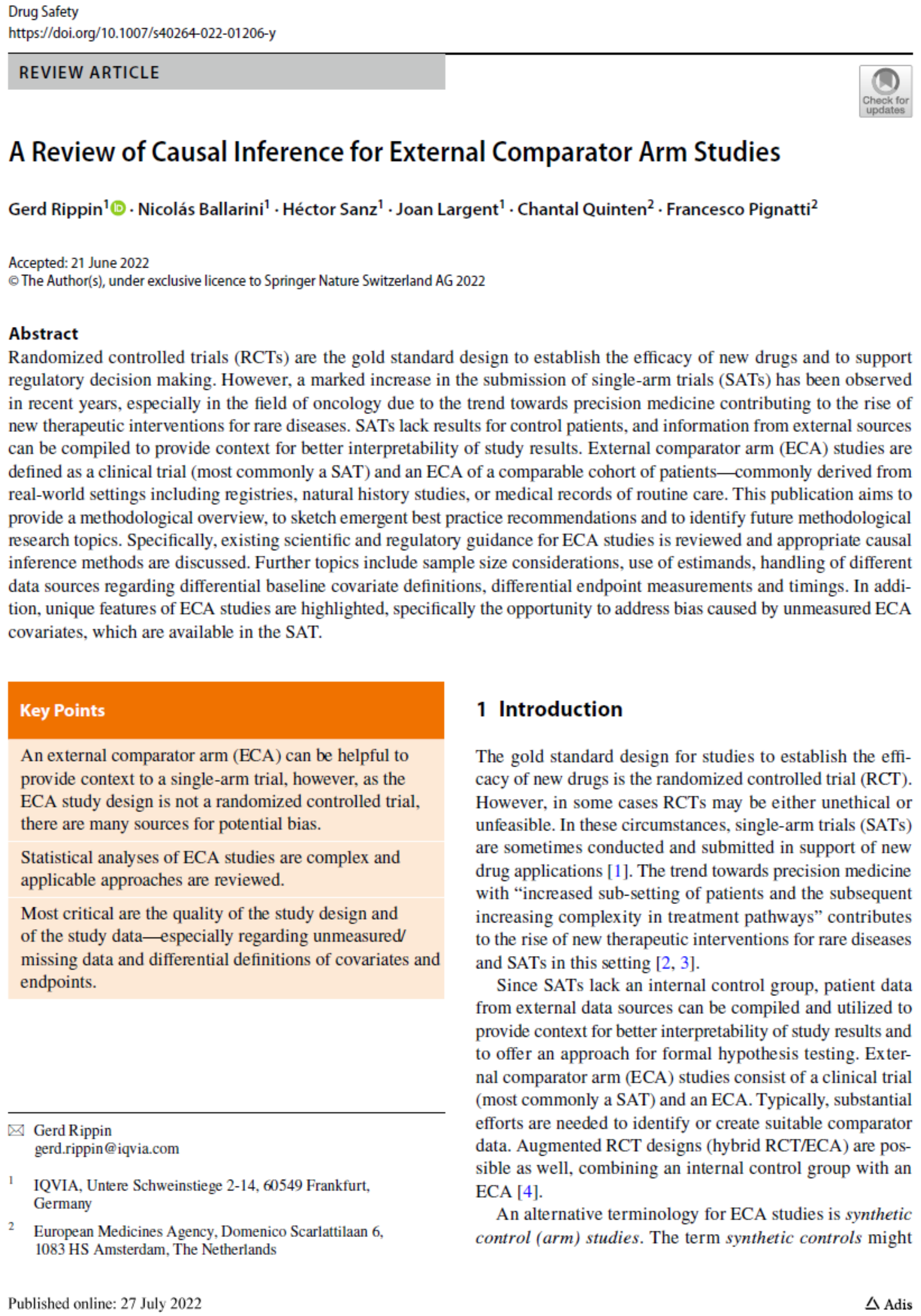
For questions regarding this draft document, contact (CDER) Dianne Paraoan, 301-796-2500, or (CBER) Office of Communication, Outreach and Development, 800-835-4709 or 240-402-8010.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Oncology Center of Excellence (OCE)

February 2023
Real-World Data/Real-World Evidence (RWD/RWE)

Joint IQVIA-EMA publication about External Comparators

This presentation is largely based on the joint IQVIA-EMA publication you see on the right-hand side



Drug Safety
https://doi.org/10.1007/s40264-022-01206-y

REVIEW ARTICLE

A Review of Causal Inference for External Comparator Arm Studies

Gerd Rippin[1] · Nicolás Ballarini[1] · Héctor Sanz[1] · Joan Largent[1] · Chantal Quinten[2] · Francesco Pignatti[2]

Accepted: 21 June 2022

Abstract

Randomized controlled trials (RCTs) are the gold standard design to establish the efficacy of new drugs and to support regulatory decision making. However, a marked increase in the submission of single-arm trials (SATs) has been observed in recent years, especially in the field of oncology due to the trend towards precision medicine contributing to the rise of new therapeutic interventions for rare diseases. SATs lack results for control patients, and information from external sources can be compiled to provide context for better interpretability of study results. External comparator arm (ECA) studies are defined as a clinical trial (most commonly a SAT) and an ECA of a comparable cohort of patients—commonly derived from real-world settings including registries, natural history studies, or medical records of routine care. This publication aims to provide a methodological overview, to sketch emergent best practice recommendations and to identify future methodological research topics. Specifically, existing scientific and regulatory guidance for ECA studies is reviewed and appropriate causal inference methods are discussed. Further topics include sample size considerations, use of estimands, handling of different data sources regarding differential baseline covariate definitions, differential endpoint measurements and timings. In addition, unique features of ECA studies are highlighted, specifically the opportunity to address bias caused by unmeasured ECA covariates, which are available in the SAT.

Key Points

An external comparator arm (ECA) can be helpful to provide context to a single-arm trial, however, as the ECA study design is not a randomized controlled trial, there are many sources for potential bias.

Statistical analyses of ECA studies are complex and applicable approaches are reviewed.

Most critical are the quality of the study design and of the study data—especially regarding unmeasured/missing data and differential definitions of covariates and endpoints.

1 Introduction

The gold standard design for studies to establish the efficacy of new drugs is the randomized controlled trial (RCT). However, in some cases RCTs may be either unethical or unfeasible. In these circumstances, single-arm trials (SATs) are sometimes conducted and submitted in support of new drug applications [1]. The trend towards precision medicine with "increased sub-setting of patients and the subsequent increasing complexity in treatment pathways" contributes to the rise of new therapeutic interventions for rare diseases and SATs in this setting [2, 3].

Since SATs lack an internal control group, patient data from external data sources can be compiled and utilized to provide context for better interpretability of study results and to offer an approach for formal hypothesis testing. External comparator arm (ECA) studies consist of a clinical trial (most commonly a SAT) and an ECA. Typically, substantial efforts are needed to identify or create suitable comparator data. Augmented RCT designs (hybrid RCT/ECA) are possible as well, combining an internal control group with an ECA [4].

An alternative terminology for ECA studies is *synthetic control (arm) studies*. The term *synthetic controls* might

✉ Gerd Rippin
gerd.rippin@iqvia.com

[1] IQVIA, Untere Schweinstiege 2-14, 60549 Frankfurt, Germany

[2] European Medicines Agency, Domenico Scarlattilaan 6, 1083 HS Amsterdam, The Netherlands

Published online: 27 July 2022

Revisiting estimands

What is an Estimand?

An estimand is a “precise definition of the treatment effect reflecting the clinical question posed by the trial objective” (EMA, 2017, ICH E9 Addendum).

- An estimand has 5 attributes:
  - **Attribute 1: Population**
    - › The population targeted for the indication (as specified by in- and exclusion criteria)
  - **Attribute 2: Treatment conditions**
    - › Dosage, route, frequency, …
  - **Attribute 3: Specification of the endpoint**
    - › Overall Response, Overall Survival, …
  - **Attribute 4: Population-level summary**
    - › E.g., Hazard Ratio (HR), Restricted Mean Survival Time, proportion of objective response,…
  - **Attribute 5: Handling of intercurrent events (IEs)**
    - › Intercurrent events are post-baseline events like:
      - » Starting a subsequent therapy, experiencing an AE, premature end of treatment, …

Estimand Attribute 1 – The Analysis Populations

Which population to take - should we rely on the ITT population?

- There is no Intention-to-treat (ITT) (or PP) population in a RW dataset
  - If it is possible to use a previous RCT control arm, the ITT population should be available in the dataset
- As a general principle, there is a preference to compare populations which share a common definition
  - Compare like with like (“apples with apples”)
  - This leads most likely to use the Safety population
    - Comparing patients who actually took the drug
    - Straightforward results interpretation, but outside common RCT standards

Estimand Attribute 1 – The Analysis Populations

Which population to take - should we rely on the ITT population?

- It is also possible to argue to use the trial’s ITT population and the external Safety population
  - Stating that the analysis is a conservative analysis
  - And that it would be meaningful to incorporate the initial drop-out rate (especially when not being very low)
    - The initial drop-out rate occurs after randomization but before the treatment is taken
- Example: Chimeric antigen receptor (CAR) – T cell treatment
  - There is a longer time between randomization and actual treatment start date
- If helpful, more than one estimand (as supplementary analyses) can be set-up

Estimand Attribute 1 – The Analysis Populations

Eligibility Criteria (I)

- The EC approximates the trial’s eligibility criteria as much as possible.
- However, some baseline (index date, time zero) information is likely to not be available in RW:
  - HIV test, ECOG,…
  - These variables are typically not or not always measured / documented in RW datasets

Estimand Attribute 1 – The Analysis Populations

Eligibility Criteria (II)

- RW measurements to derive the analysis population are usually taken from a time window (look-back period) before the index date
  - E.g., 3, 6 or 12 months before true baseline
- The longer eligibility measurements are retrieved from the past, the more likely the occurrence of measurement error
  - Either random measurement error (without a systematic shift), or
  - a systematic shift in case the patient’s condition is deteriorating fast
- This leads to another dimension of approximation, in addition to the eligibility criteria approximation

Estimand Attribute 1 – The Analysis Populations

Different ways to apply criteria for making populations more similar

- If the trial results are already available (or at least baseline measurements), it can be considered to restrict the covariate value ranges according to what is observed in the trial
  - Instead of just applying the trial's eligibility criteria to the EC only
  - Following to principle to compare like with like (Pocock, 1976, Gray et al, 2020)
- An example:
  - › Eligibility says age>18, but the actual range is between 43 and 82
  - › Could restrict the age range for the ECC correspondingly
  - Could do this approach for all important covariates
- There is no guidance yet whether such an approach would be considered preferable

Estimand Attribute 1 – The Analysis Populations

Attribute 1 Summary

- Clearly, the analysis population is in most cases (if not in all) an approximation of the population of interest
- Hence, in the context of ECs the difficulties already seen for the estimand attribute number 1 make the whole estimand an approximation
- Refinement of the analysis population according to observed trial value ranges is possible but there is no regulation / guideline yet which is assessing this option “officially”

Estimand Attribute 2 – Treatment conditions

Doses, route of administration,…

- Eligible doses, route of administration, etc., need to be specified unambiguously
- If different doses have different expected treatment effects, separate analyses must be performed
  - This relates to the “consistency assumption” of causal inference methodology (see also next page)

Estimand Attribute 2 – Treatment conditions

Can we test against SoC, which is an umbrella term for diverse treatments?

- SoC deserves a thorough description by means of statistical tables
  - Descriptive statistics of SoC treatments
  - Note that the definition of SoC may vary by country/region
    - May not be as consistent in RW as it is in a controlled setting

- The validity of testing against SoC depends critically on whether the single treatments summarized under the umbrella term of SoC do have the same expected treatment effect
  - Statistically speaking, we need to check for a violation of the consistency assumption in causal inference:
    - Check for homogeneity of SoC treatment effects at least for the most common treatments
    - Homogeneity of SoC treatment effects is a (partially) testable model assumption

- Note that this topic is also related to attribute 1 “Analysis population”
  - Sometimes the SoC definition needs to be narrowed down because some RW SoC are classified to not be appropriate, and then some patient groups are no longer included in the analysis population

Estimand Attribute 2 – Treatment conditions

Exposure - be mindful of RW treatment exposure, which may be lower compared to a SAT setting

- In the controlled setting of a SAT there is a likelihood for higher (close to perfect) exposure compared with the non-ideal RW setting
  - This is due to a strict SAT protocol, and strong monitoring of sites and patients to follow the protocol
  - Thus, differential exposure / compliance / adherence across the two data sources is a definitive possibility
- Always describe treatment exposure in detail for highest transparency
  - E.g., exposure times, cycles
- It needs to be evaluated for the concrete study at hand whether any needed **minimum number of treatment cycles** or **minimum exposure time** across both treatment arms is a reasonable approach to overcome potentially substantial exposure differences
  - Exposure time is post-baseline, so this is non-trivial to handle statistically
    - › This is also related to the estimand attribute 5 how to handle intercurrent events
    - › Requesting a minimal exposure introduces immortal time bias
  - Possible to perform a supplementary analysis applying a minimum required exposure

Estimand Attribute 3 – Specification of the Endpoint

Are endpoints comparable across cohorts?

- While endpoints like Overall Survival are typically comparable across cohorts, others may be less comparable:
  - Progression-free survival
    - Progression is measured differently in RW: Typically, no RECIST or Lugano-standardized measurements
    - Comparability will be limited
  - For composite time-to-event (TTE) endpoints in general (consisting of two or more outcomes) - maybe somewhat surprisingly - different censoring proportions for the earlier part of the composite endpoint can bias estimations
    - Assume an example where PFS occurs a lot earlier than death, but there is 100% missingness for progression dates and 0% missingness for survival. Then of course, the estimate for the endpoint PFS is heavily biased, and actually the endpoint survival is estimated instead of PFS
      - Hence PFS, Time to next treatment or death, Duration of response until progression or death and other composite endpoints may not be comparable with trial data

Estimand Attribute 3 – Specification of the Endpoint

Are endpoints comparable across cohorts?

- Safety outcomes are typically not as exhaustively documented in RW data sources as in controlled trials, so are typically not comparable.
  - There may be though a (close to) complete list of major / life-threatening events or hospitalizations / doctor’s visits in some RW data sources, which may allow to compare key safety events

Estimand Attribute 3 – Specification of the Endpoint

How to handle different timings of follow-up?

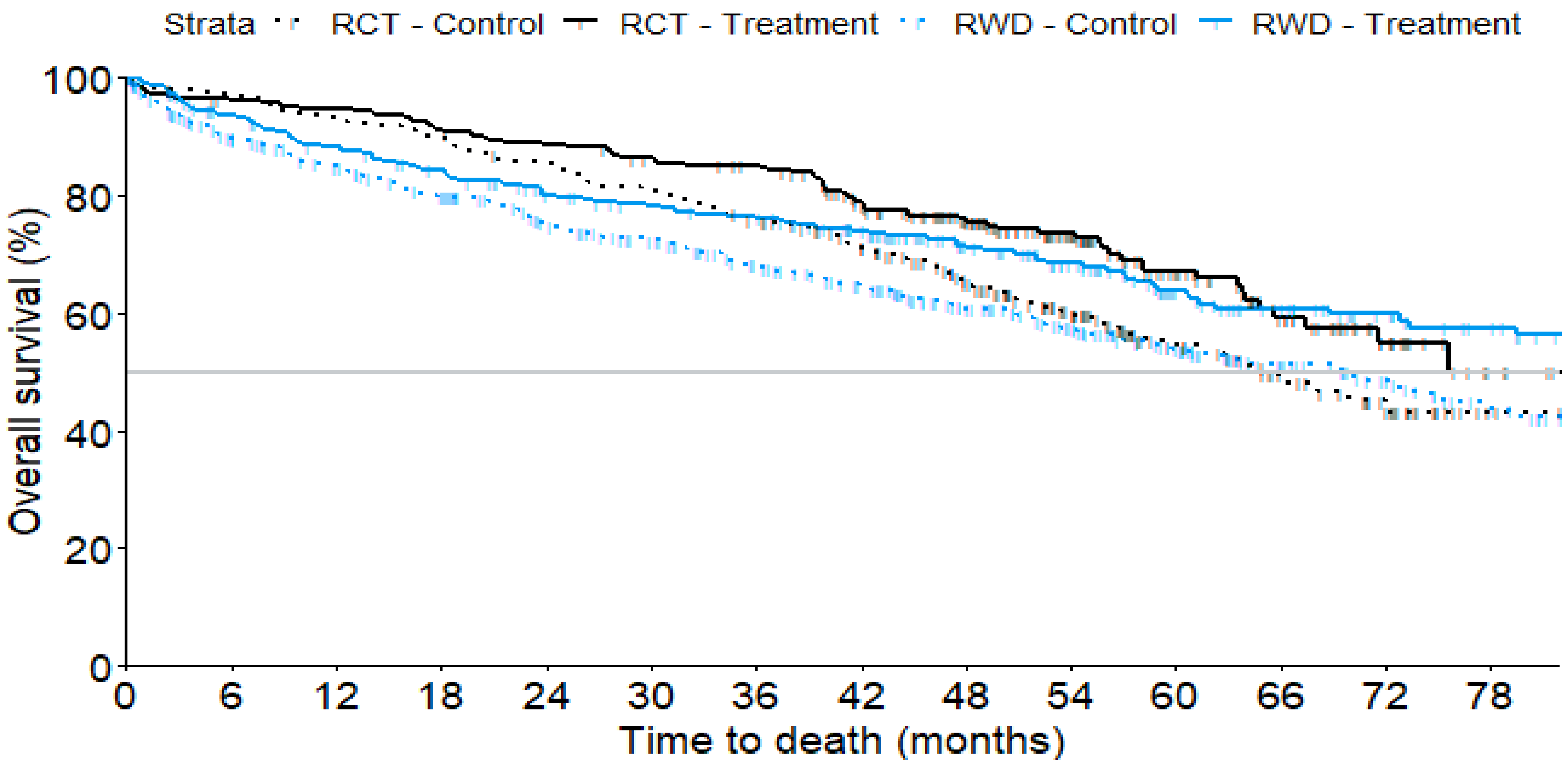
- Could define a time window around a specific target date for defining acceptable outcome measurements
  - E.g., 3 months after baseline +/- 2 weeks
- Could use interval-censoring methods for some time-to-event endpoints
  - Traditionally, progression is assumed to have occurred at the time of progression measurement
    - This constitutes a simplification which works less well for RW data (compared to RCTs) with heterogeneous follow-up schedules,
    - This simplification has been criticized before: Collet (2023), Zhang et al. (2017), Bogaerts et al. (2021)
      - There can be more frequent follow-up visits for more severe cases (intensity bias)
      - Recommendation to use interval-censoring methods

Estimand Attribute 4 – Population-level Summary

Certain assumptions can be more fragile in ECC studies

- The proportional hazards assumption of the Cox model is easily lost when comparing treatments across data sources (see next slide for an example)
- Recommendation to estimate rather restricted mean survival time differences (RMSTDs)
  - Different ways to implement
  - Restricted Mean Survival Time (RMST) models as one option
    - Andersen et al. (2004), Tian et al. (2014)
- Note that the Cox regression model was criticized the last decade anyway
  - Aalen et al. (2015), Hernán (2010), Mao et al. (2018), Rufibach et al. (2019), Martinussen et al. (2020), Stensrud et al. (2019)
- Also outside of time-to-event analyses some assumptions for population-level summaries may become more fragile due to the different setting of the trial and RW data.
  - For example, variances may become more easily heteroscedastic

Kaplan-Meier survival curves for different multiple myeloma RCT and RWD cohorts



Number at risk

Strata														
RCT - Control	229	221	212	204	194	183	170	151	119	86	51	34	23	3
RCT - Treatment	242	230	227	218	211	203	197	169	139	99	62	39	20	6
RWD - Control	461	389	356	327	293	271	241	215	183	157	128	103	85	69
RWD - Treatment	284	254	236	222	202	192	186	170	139	118	84	71	57	47

Hoogendoorn WE, Sanz H, Quinten C, Pignatti F, Largent J, Rippin G. A methodology study to evaluate ECA study results versus RCT results: Multiple Myeloma case study. Poster at 38th ICPE, August 2022.

Estimand Attribute 5 – Handling of Intercurrent Events (IEs)

IEs can differ more dramatically compared to RCTs

- IEs can differ strongly across cohorts due to different settings and data temporality
  - For example, the kind of subsequent therapies may be completely different
- The more long-term the endpoint the more important to consider IEs
  - This would suggest to look out for alternatives to the treatment policy estimand
    - Potentially as supplementary analyses
- For short-term endpoints IEs may be less relevant
  - In this case the rationale for the treatment policy estimand may be stronger

We discussed the 5 estimand attributes.
Are we done now?

Estimand Attribute 6? – The Definition of Baseline

The estimand is not completely defined without the specification of baseline (index date, time zero)

- How to define the baseline date for the SAT in an EC study?
  - Usually taken as the start of index treatment
- However, there are situations where the treatment start date is unclear or not available
  - Use of combination treatments with multiple start dates
  - Treatment is unknown/unlikely to be effective
    - Treatment may be taken much later because it is not considered an effective therapy
  - Multiple baselines may be possible
    - For example, at the start date of the third and higher lines of treatment (LoTs) in oncological settings
    - Using all baselines may be an efficient approach (Backenroth, 2021, Hatswell et al., 2022)
  - No treatment is available at all (test against non-users)
    - May take a progression date or another important clinical event (myocardial infarct, stroke) as baseline
    - However, immortal time bias is introduced for the trial participants
      - Time from index date (e.g., progression) to treatment start date
      - Need to handle immortal time bias statistically
    - An index date which is far in the past has disadvantages, as it adds unnecessary “white noise” to the analysis.

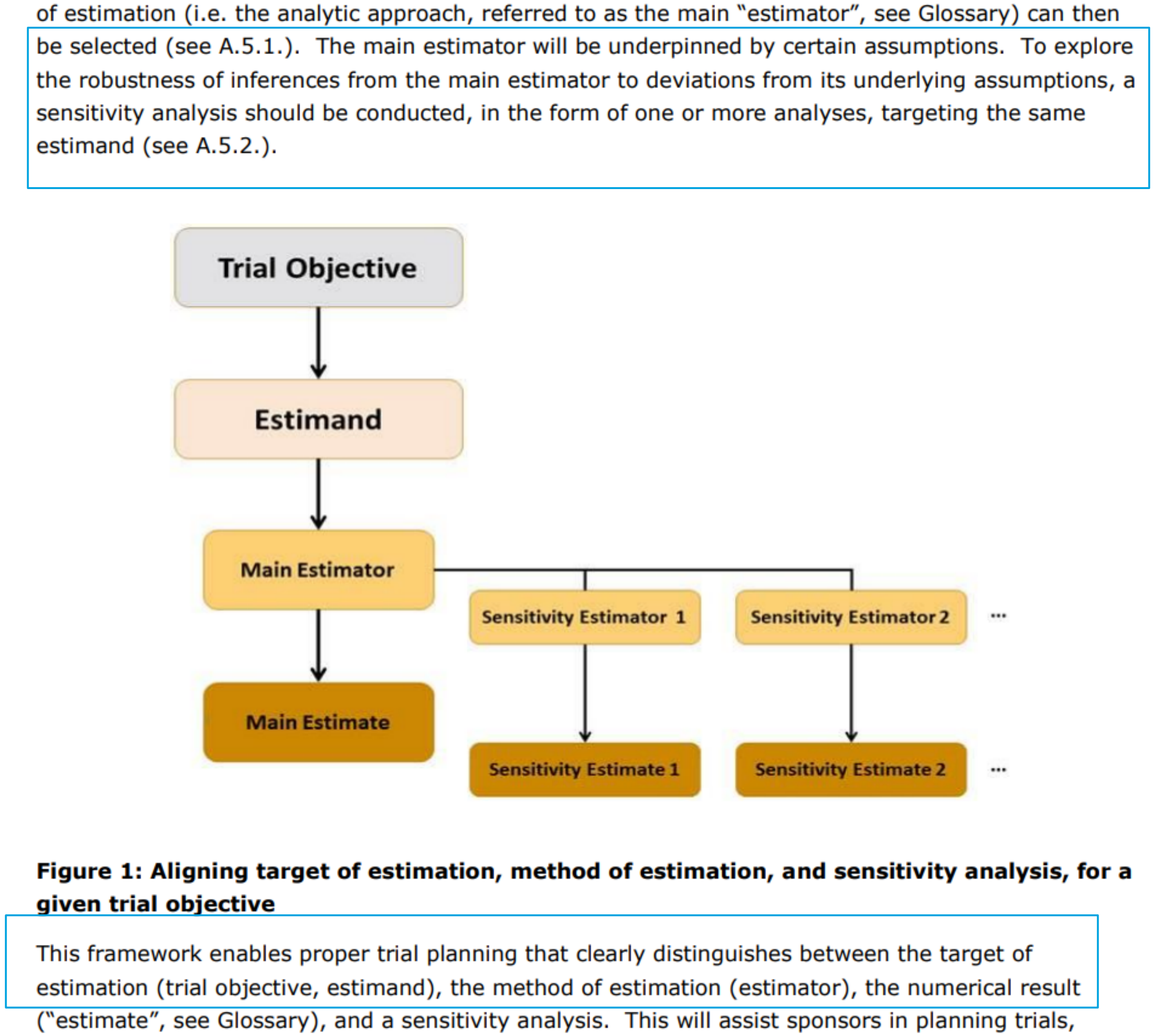
Estimand Attribute 6? – The Definition of Baseline

Issues with the validity of baseline measurements

- Covariates measured at the original SAT baseline may become post-baseline covariates if the index date is moved to an earlier time point (e.g., disease progression)
  - This is problematic from a statistical theory perspective, as only baseline variables can be used in traditional statistical models (like propensity score models) for covariate adjustment
- How to define what constitutes a valid baseline measurement?
  - As stated before, the ECA typically defines a look-back period in order to define which measurements are eligible to be considered baseline
    - E.g., define the last 3 / 6 / 12 months before index date as an acceptable look-back period
    - This period may be chosen to be different for specific classes of measurements (like laboratory values, ECOG, severity of disease like Gleason score or number of bone lesions, genetic testing, etc.)
    - Often, multiple reasonable definitions of look-back periods are possible, which may lead to defendable but also to somewhat arbitrary choices
    - Even if the look-back period is defined in an optimal way in some sense, still, measurement error may arise by actual values at true baseline being different from previously recorded values
    - This is a threat to the validity of causal inference methods

Estimand & Estimator Relationship

ICH E9(R1) Figure & Explanations:



of estimation (i.e. the analytic approach, referred to as the main "estimator", see Glossary) can then be selected (see A.5.1.). The main estimator will be underpinned by certain assumptions. To explore the robustness of inferences from the main estimator to deviations from its underlying assumptions, a sensitivity analysis should be conducted, in the form of one or more analyses, targeting the same estimand (see A.5.2.).

Figure 1: Aligning target of estimation, method of estimation, and sensitivity analysis, for a given trial objective

This framework enables proper trial planning that clearly distinguishes between the target of estimation (trial objective, estimand), the method of estimation (estimator), the numerical result ("estimate", see Glossary), and a sensitivity analysis. This will assist sponsors in planning trials,

Estimand Attribute 7? – The Marginal Estimator

The estimand is not completely defined without the specification of the marginal estimator

- An RCT is estimating the ATE under mild conditions
- ECC studies are designed to estimate one (or more) of the marginal estimators of ATE, ATT, ATU or ATO
  - ATE: Average Treatment Effect
  - ATT: Average Treatment Effect on the Treated
  - ATU: Average Treatment Effect on the Untreated
  - ATO: Average Treatment Effect in the Overlap Population
- Potentially, the marginal estimator can be handled as per the estimand attribute “population-level summary”
  - by saying that the “ATE hazard ratio” or the “ATT relative risk” is the population-level summary
    - › However, a new attribute seems to be a natural solution, because the population-level summary (e.g., a relative risk) is specified independently from the marginal estimator
- Another alternative is to say that the marginal estimator should be implemented on the estimator level (and not as an additional estimand attribute).
    - › However, the estimand level seems to be appropriate, because the quantity which is estimated by different marginal estimators is expected to change.
    - › A different estimator can change the estimate because of different assumptions, but estimating the ATE, ATT, etc. is conceptually different (estimating a different estimand).
- What do you think? Please state your opinion in the discussion! ☺

Conclusions for External Comparator Cohort Studies

1 The estimand framework applies also for ECs.

2 However, ECs need nuanced discussions of the 5 estimand attributes.

3 Further estimand attributes may be considered for the framework (baseline definition, marginal estimator).

4 Apply supplementary analyses for a robust description of study results .

Questions for the audience

An estimand is a “precise definition of the treatment effect reflecting the clinical question posed by the trial objective” (EMA, 2017, ICH E9 Addendum).

1 Do you think that the specification of baseline is a worthy / needed estimand attribute to have a “precise definition of the treatment effect” in place?

2 Do you think that the marginal estimator should be specified in the population-level summary attribute, on the estimator level or as an additional estimand attribute?

Thank you!

Dr. Gerd Rippin

Senior Director, Biostatistics

Real-World Solutions, IQVIA

Gerd.Rippin@iqvia.com

References

Estimands

- ICH E9(R1) Expert Working Group. ICH E9(R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. EMA/CHMP/ICH/436221/2017, 2020. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf.

- U.S. Food and Drug Administration (2023). Considerations for the design and conduct of externally controlled trials for drug and biological products. Draft Guidance for Industry. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-design-and-conduct-externally-controlled-trials-drug-and-biological-products.

- Rippin G, Ballarini N, Sanz H, Largent J, Quinten C, Pignatti F: A review of causal inference for external comparator arm studies. Drug Saf. 2022;45(8):815-837.

- Mao H, Li L, Yang W, Shen Y. On the propensity score weighting analysis with survival outcome: Estimands, estimation, and inference. *Stat Med* 2018;37(26):3745-3763.

- Bornkamp B, Rufibach K, Lin J et al. Principal stratum strategy: potential role in drug development. Pharmaceutical Statistics 2021;20(4):737-751

References

Interval-censoring

- Collett D. Modelling survival data in medical research. Fourth edition, 2023. Chapman & Hall/CRC, New York. https://doi.org/10.1201/9781003282525.
- Bogaerts K, Komarek A, Lesaffre E. Survival analysis with interval-censored data. A practical approach with examples in R, SAS and BUGS, 2017. Chapman & Hall/CRC, New York. http://dx.doi.org/10.1201/9781315116945.
- Zhang X, Pillenayegum E, Chan KKW. The impact of ignoring interval censoring in progression-free survival in cancer trials: a systematic review. University of Toronto J of Publ Health, 2021;2(2). http://dx.doi.org/10.33137/utjph.v2i2.36844.

RMST model

- Andersen PK, Hansen MG, Klein JP. Regression analysis of restricted mean survival time based on pseudo-observations. Lifetime data analysis 2004;10:335-350.
- Tian L, Zhao L, Wei LJ. Predicting the restricted mean event time with the subject’s baseline covariates in survival analysis. Biostatistics 2014;(15(2):222-233.

References

Criticism of the Cox model

- Aalen OO, Cook RJ, Roysland K. Does Cox analysis of a randomized survival study yield a causal treatment effect? Lifetime Data Anal, 2015;21(4):579-593. http://dx.doi.org/10.1007/s10985-015-9335-y.
- Hernán MA. The hazards of hazard ratios. Epidemiology, 2010;21(1):13-15. http://dx.doi.org/10.1097/EDE.0b013e3181c1ea43.
- Mao H, Li L, Yang W, Shen Y. On the propensity score weighting analysis with survival outcome: Estimands, estimation, and inference. Stat Med, 2018;37(26):3745-3763. http://dx.doi.org/10.1002/sim.7839.
- Rufibach K. Treatment effect quantification for time-to-event endpoints – estimands, analysis strategies, and beyond. Pharm Stat, 2019;18(2):145-165. http://dx.doi.org/10.1002/pst.1917.
- Martinussen T, Vansteelandt S, Andersen PK. Subtleties in the interpretation of hazard contrasts. Lifetime Data Anal, 2020;26(4):833-855. http://dx.doi.org/10.1007/s10985-020-09501-5.
- Stensrud MJ, Aalen JM, Aalen OO, Valberg M. Limitations of hazard ratios in clinical trials. Eur Heart J, 2019;40:1378-1383. http://dx.doi.org/10.1093/eurheartj/ehy770.

References

External Comparators

- Pocock, S. (1976). The combination of randomized and historical controls in clinical trials. J Chronic Dis. 29, 175–88. https://doi.org/10.1016/0021-9681(76)90044-8.
- Gray, C.M., Grimson, F., Layton. D. et al. (2020). A framework for methodological choice and evidence assessment for studies using external comparators from real-world data. Drug Saf 43, 623–33. https://doi.org/10.1007/s40264-020-00944-1.
- Ghadessi, M., Tang, R., Zhou, J. et al. (2020). A roadmap to using historical controls in clinical trials – by Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG). Orphanet J Rare Dis. 15 (1), 1-19. http://dx.doi.org/10.1186/s13023-020-1332-x.
- Burger, H.U., Gerlinger, C., Harbron, C. et al. (2021). The use of external controls: To what extent can it currently be recommended? Pharm Stat 20 (6), 1002-1016. http://dx.doi.org/10.1002/pst.2120.

References

External Comparators

- Seeger, J.D., Davis, K.J., Innacone, M.R. et al. (2020). Methods for external control groups for single arm trials or long-term uncontrolled extensions to randomized clinical trials. Pharmacoepidemiol Drug Saf. 29, 1382-1392. http://dx.doi.org/10.1002/pds.5141.
- Skovlund, E., Leufkens, H.G.M., Smyth, J.F. (2018). The use of real-world data in cancer drug development. Eur J Cancer 101, 69-76. http://dx.doi.org/10.1016/j.ejca.2018.06.036.
- Thorlund, K., Dron, L., Park, J.J.H., et al. (2020). Synthetic and external controls in clinical trials – a primer for researchers. Clin Epi 12, 457-467. http://doi.org/10.2147/CLEP.S242097.

Selection of baseline (index date, time zero)

- Backenroth, D. (2021). How to choose a time zero for patients in external control arms. Pharm Stat 20, 783–92. http://dx.doi.org/10.1002/pst.2107.
- Hatswell, A.J., Deighton, K., Snider, J.T. et al. (2022). Approaches to selecting “time zero” in external control arms with multiple potential entry points: a simulation study of 8 approaches. Medic Dec Making 42 (7), 893–905. http://dx.doi.org/10.1177/0272989X221096070.

Back-up Slides

Gerd Rippin, PhD

Senior Director, Biostatistics

Real-World Solutions, IQVIA

Gerd.Rippin@iqvia.com

Revisiting estimands

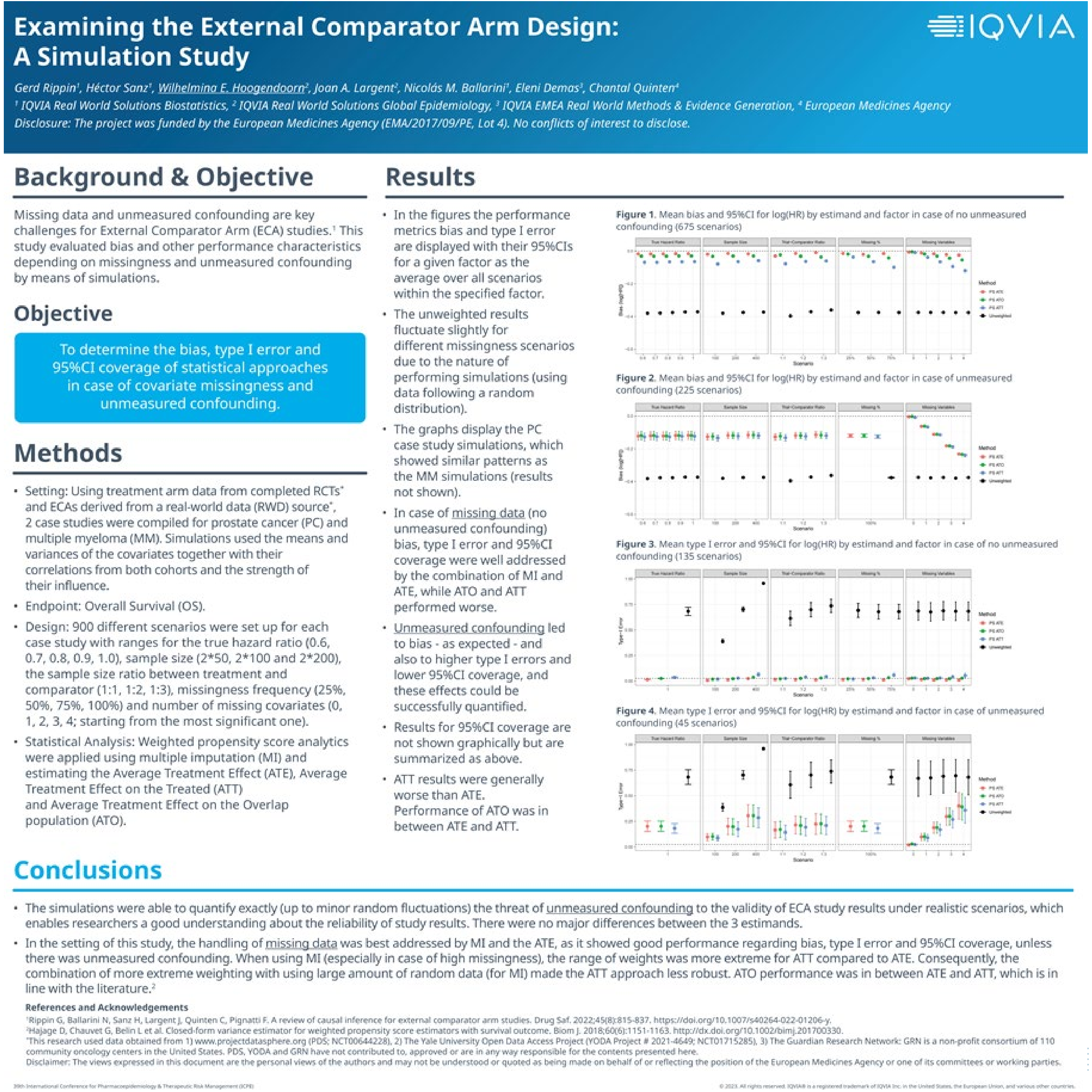
What is an Estimand?

Handling of intercurrent events

- ICH E9 addendum specifies 5 ways of handling intercurrent events
  - Treatment policy, hypothetical, while-on-treatment, composite endpoint, principal stratum
  - Treatment policy: Ignoring IEs
  - Hypothetical: Adjusting for IEs statistically by modelling missing endpoint data
    - Might apply Inverse Probability of Censoring Weights (IPCW) for time-to-event endpoints
    - Simple censoring only valid if the censoring event is uninformative of future survival
    - The hypothetical approach becomes especially important for long-term endpoints like survival, since IEs can be quite different across the 2 data sources, and even more in case of different data temporality
  - While on treatment: Endpoint measurement only considered while on treatment
  - Composite: Incorporating an IE into the endpoint definition (for example Progression-free Survival)
  - Principal stratum: The principal stratum is a post-baseline stratum of patients, in which an IE would not have happened, which is different from the subset of patients actually not exhibiting the IE. This strategy is not widely applied, e.g., as “it relates to a subpopulation of the overall trial population that is not identifiable” (Bornkamp et al., 2021)

ATE or ATT – which to choose?

ATE /ATT differences when using multiple imputation and PS weighting



Examining the External Comparator Arm Design: A Simulation Study

Gerd Rippin[1], Héctor Sanz[1], Wilhelmina E. Hoogendoorn[2], Joan A. Largent[2], Nicolás M. Ballarini[1], Eleni Demas[3], Chantal Quinten[4]

[1] IQVIA Real World Solutions Biostatistics, [2] IQVIA Real World Solutions Global Epidemiology, [3] IQVIA EMEA Real World Methods & Evidence Generation, [4] European Medicines Agency

Disclosure: The project was funded by the European Medicines Agency (EMA/2017/09/PE, Lot 4). No conflicts of interest to disclose.

Background & Objective

Missing data and unmeasured confounding are key challenges for External Comparator Arm (ECA) studies.[1] This study evaluated bias and other performance characteristics depending on missingness and unmeasured confounding by means of simulations.

Objective

To determine the bias, type I error and 95%CI coverage of statistical approaches in case of covariate missingness and unmeasured confounding.

Methods

- Setting: Using treatment arm data from completed RCTs* and ECAs derived from a real-world data (RWD) source*, 2 case studies were compiled for prostate cancer (PC) and multiple myeloma (MM). Simulations used the means and variances of the covariates together with their correlations from both cohorts and the strength of their influence.
- Endpoint: Overall Survival (OS).
- Design: 900 different scenarios were set up for each case study with ranges for the true hazard ratio (0.6, 0.7, 0.8, 0.9, 1.0), sample size (2*50, 2*100 and 2*200), the sample size ratio between treatment and comparator (1:1, 1:2, 1:3), missingness frequency (25%, 50%, 75%, 100%) and number of missing covariates (0, 1, 2, 3, 4; starting from the most significant one).
- Statistical Analysis: Weighted propensity score analytics were applied using multiple imputation (MI) and estimating the Average Treatment Effect (ATE), Average Treatment Effect on the Treated (ATT) and Average Treatment Effect on the Overlap population (ATO).

Results

- In the figures the performance metrics bias and type I error are displayed with their 95%CIs for a given factor as the average over all scenarios within the specified factor.
- The unweighted results fluctuate slightly for different missingness scenarios due to the nature of performing simulations (using data following a random distribution).
- The graphs display the PC case study simulations, which showed similar patterns as the MM simulations (results not shown).
- In case of missing data (no unmeasured confounding) bias, type I error and 95%CI coverage were well addressed by the combination of MI and ATE, while ATO and ATT performed worse.
- Unmeasured confounding led to bias - as expected - and also to higher type I errors and lower 95%CI coverage, and these effects could be successfully quantified.
- Results for 95%CI coverage are not shown graphically but are summarized as above.
- ATT results were generally worse than ATE. Performance of ATO was in between ATE and ATT.

Figure 1. Mean bias and 95%CI for log(HR) by estimand and factor in case of no unmeasured confounding (675 scenarios)

Figure 2. Mean bias and 95%CI for log(HR) by estimand and factor in case of unmeasured confounding (225 scenarios)

Figure 3. Mean type I error and 95%CI for log(HR) by estimand and factor in case of no unmeasured confounding (135 scenarios)

Figure 4. Mean type I error and 95%CI for log(HR) by estimand and factor in case of unmeasured confounding (45 scenarios)

Conclusions

- The simulations were able to quantify exactly (up to minor random fluctuations) the threat of unmeasured confounding to the validity of ECA study results under realistic scenarios, which enables researchers a good understanding about the reliability of study results. There were no major differences between the 3 estimands.
- In the setting of this study, the handling of missing data was best addressed by MI and the ATE, as it showed good performance regarding bias, type I error and 95%CI coverage, unless there was unmeasured confounding. When using MI (especially in case of high missingness), the range of weights was more extreme for ATT compared to ATE. Consequently, the combination of more extreme weighting with using large amount of random data (for MI) made the ATT approach less robust. ATO performance was in between ATE and ATT, which is in line with the literature.[2]

References and Acknowledgements

[1] Rippin G, Ballarini N, Sanz H, Largent J, Quinten C, Pignatti F. A review of causal inference for external comparator arm studies. Drug Saf. 2022;45(8):815-837. https://doi.org/10.1007/s40264-022-01206-y.

[2] Hajage D, Chauvet G, Belin L et al. Closed-form variance estimator for weighted propensity score estimators with survival outcome. Biom J. 2018;60(6):1151-1163. http://dx.doi.org/10.1002/bimj.201700330.

*This research used data obtained from 1) www.projectdatasphere.org (PDS; NCT00644228), 2) The Yale University Open Data Access Project (YODA Project # 2021-4649; NCT01715285), 3) The Guardian Research Network: GRN is a non-profit consortium of 110 community oncology centers in the United States. PDS, YODA and GRN have not contributed to, approved or are in any way responsible for the contents presented here.

Disclaimer: The views expressed in this document are the personal views of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the European Medicines Agency or one of its committees or working parties.

39th International Conference for Pharmacoepidemiology & Therapeutic Risk Management (ICPE)