



Science For A Better Life

Survival Design Strategies in an Estimands Framework

Jonathan Siegel, Bayer

The Pharmaceutical Industry Working
Group on Estimands in Oncology
Lifetime Data Science Conference, Pittsburgh, PA
May 30, 2019

Disclaimer

The views presented in this presentation are those of the author alone and do not represent the views of Bayer, AG or any of its constituent companies, partners, or affiliates; the Pharmaceutical Industry Working Group on Estimands in Oncology, or any other organization.

Regulatory Context

- The International Conference for Harmonization (ICH) is a collective effort of major pharmaceutical industry regulatory agencies including the FDA (US), EMA/CHMP (EU), PMDA (Japan), and others to develop a joint regulatory approach.
- Among its activities it produces guidance documents jointly agreed upon by the agencies.
- The ICH E9 Guidance, “Statistical Principles for Clinical Trials”, issued in 1998, has been the industry’s primary guidance for statistical practice in clinical trials.
- The draft ICH E9 R1 guidance “Addendum on estimands and sensitivity analyses in clinical trials,” represents the first comprehensive update to regulatory guidance on statistical practice since the E9 guidance was issued.

Background to Change

- A key theme of the guidance is to address a conceptual revolution that has taken place in the characterization and handling of “missing data.”
- Historically, the statistics community has simply assumed that missing data is random and uninformative and can be ignored or imputed based on very simple assumptions.
- This is particularly true in lifetime data analysis, where censoring and its assumption of uninformativeness lies at the heart of almost all our methods.

Is Missing Data Uninformative?

- Clinical trialists have increasingly disputed the validity of assumptions of uninformativeness in realistic clinical trial settings.
- In their view, the reasons why data may not be available may be systematic and can be highly informative
 - Treatment withdrawal and loss to follow-up are often due to safety events or perceived lack of treatment efficacy
 - Patients in the worst condition (e.g. greatest pain) may have the most difficulty filling out questionnaires
 - If safety events delay assessments, they can lengthen TTE efficacy indicators and make them appear more efficacious

The Prevention and Treatment of Missing Data

- In 2010, the US National Research Council issued the white paper “The Prevention and Treatment of Missing Data in Clinical Trials” proposing a rethinking and redesign of clinical trials to address more realistic assumptions, including:
 - Trial designs and operational strategies that follow patients consistently, simplify data collection, and avoid missing data where possible
 - Collecting data on reasons and mechanisms for missing data
 - Clearly articulating and justifying assumptions made regarding expected missing data mechanisms in the clinical trial’s context at trial design
 - Endpoint definitions and trial design strategies incorporating these assumptions
 - Post-hoc sensitivity analyses to ensure assumptions made were justifiable.

The Estimands Guidance

- The purpose of the estimands guidance is largely to implement this new thinking. It introduces new concepts and terminology to address the paradigm shift.
- ***Intercurrent events*** are defined as events that occur after treatment initiation and either preclude observation of the variable or affect its interpretation.
- An ***estimand*** attempts to address how the outcome of treatment observed in the study compares to what would have happened under different treatment conditions
- Clinical trial design requires ***strategies*** for reliable evaluation of estimands in the presence of intercurrent events.

Oncology Estimands WG

- initiated and led by Evgeny Degtyarev (Novartis) and Kaspar Rufibach (Roche), first TC Feb 2018
- Main purpose: ensure **common understanding and consistent definitions** for key estimands in Oncology across industry
- 32 members (14 from Europe and 18 from US) representing **20** companies
 - 5 subteams: causal, treatment switching, censoring mechanisms, case studies in solid tumors, case studies in hematology
- **EFSPi SIG** (Nov 2018) and **ASA Biopharm Section SWG** (Apr 2019)
- in dialogue with regulators from EMA, FDA, Japan, China, Taiwan, Canada



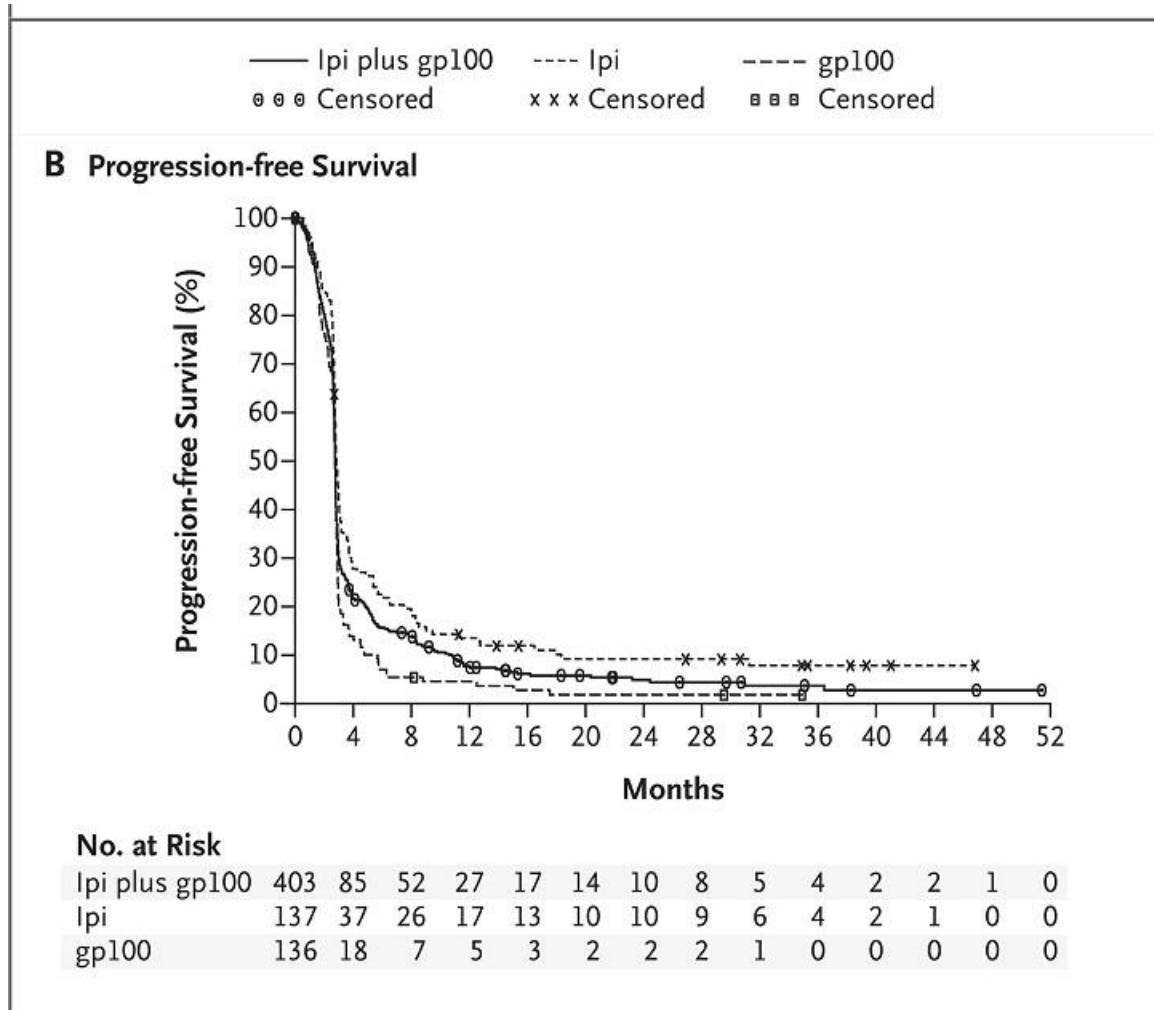
A Simple Example

- In 3-arm Phase III trial of ipilimumab + gp100 peptide vaccine, ipilimumab plus gp100 placebo, or gp100 plus ipilimumab placebo, the following results for median PFS were reported (Hodi et al., 2010):

Treatment	Median (months)	95% CI
Ipilimumab + gp100	2.76	(2.73, 2.79)
Ipilimumab + gp100 placebo	2.86	(2.76, 3.02)
Gp100 + ipilimumab placebo	2.76	(2.73, 2.83)

- Conclusions:
 - Medians reliably estimated in all 3 arms (narrow CIs)
 - Medians similar in all 3 arms
- Question: Correct?

Simple Example: PFS Median



A Simple Example (Cont.)

- Answer: Not so. The results are not what they seem.
- In this study, the first assessment was scheduled at 12 weeks
- By the 12th week, a majority of patients in all 3 arms had progressed
 - The median in all 3 arms actually occurred before 12 weeks without being measured.
- The reported median simply reflects when the first measurement occurred.
- The reported CIs simply reflect patient scheduling variation around the target first assessment time, not variation in the PFS itself.
- Features of the study design (measurement schedule) are confounded with, and being reported as, features of the thing measured.

Features Illustrated

- This simple, extreme example illustrates several features that the estimands guidance's authors want study designers to think about very carefully:
 - Intercurrent events, events which prevent or confound observation of the endpoint of interest.
 - In this case the visit schedule itself, or an incompatibility between the visit schedule and the quantity estimated (median PFS), is the cause of the intercurrent event.
 - The fundamental question of estimands, ***what are we actually measuring? What do our results mean?***
 - This requires us to look at the interaction between study design, intercurrent events, the effect of interest, and analysis methods
- We want to detect situations where results aren't what they appear to be.
- We want to proactively design studies so that they report what we want.

An estimand

- In the E9 R1 guidance, an ***estimand*** defines the target of estimation for a particular trial objective (“what is to be estimated”) through specification of:
 - The population to be evaluated
 - The variable being assessed
 - The handling of intercurrent events
 - The population-level summary for the variable (the statistic)
- In general, these four elements are treated as a package. They must be pre-specified and justified
- The trial design must align with the strategy for handling intercurrent events.
- The guidance requires post-hoc sensitivity analyses to check that the strategy for handling intercurrent events was appropriate

Strategies for handling intercurrent events

- The guidance articulates 5 strategies for handling intercurrent events.
 - The discussion illustrates how changing the strategy changes:
 - The meaning of what is being estimated and conclusions that can be drawn
 - The required trial design and visit and follow-up schedule.
 - The analysis method
- In addition, the guidance recommends general practices for the prevention and treatment of intercurrent events:
 - Following patients consistently to the extent appropriate for the strategy
 - Visit schedule frequency appropriate for reliable estimation
 - Obtaining reasons for withdrawal/dropout appropriate to assessing correlation with outcome
 - Simplified assessments to minimize errors

Strategies in Brief

- Treatment policy strategy
 - Ignore intercurrent events, follow patients until event of interest
- Composite strategy
 - Treat intercurrent events as events (composition with event of interest)
- Hypothetical strategy
 - Define scientific question as what would have happened if intercurrent event had not occurred
- Principal stratum strategy
 - Define scientific question by limiting population of interest to patients in whom intercurrent events do not occur
- While on treatment strategy
 - Define scientific question as only concerned with time prior to intercurrent events.

How Strategy Changes Interpretation:

Example: Time to Progression

- Time to progression (TTP) is sometimes used as an endpoint in early-stage cancer where death due to the disease is considered unlikely.
- The traditional approach is to use Kaplan-Meier, Cox, etc. and censor deaths.
- But what exactly does this mean?
- It can be interpreted as a hypothetical strategy (Rufiback, 2018)
 - Censoring dead patients treats them as still at risk, and assumes their risk of subsequent progression is similar to progression in live patients
 - The scientific question is what would have happened if the patient hadn't died, and had progression hazard after death similar to progression hazard in the same timeframe in patients who didn't die.

Time to Progression: Hypothetical vs. While on Treatment

- Consider, instead, a while on treatment strategy (Rufiback, 2018)
- In this case, we are only interested in what happens before patients die.
 - A classic example is a palliative medication. Palliative efficacy is defined as relief of symptoms during the period prior to death. Shorter or longer survival time is irrelevant to palliative efficacy.
 - So once dead, a patient is removed from the risk set
- This means a while-on-treatment strategy for TTP should be modeled using a competing risk model, e.g. CIF.
 - Death is modeled as a competing risk event.
- Changing the strategy changes the scientific question. It results in a different analysis method, a different estimate, and a different interpretation of what the estimate means.

Strategies and Types of informativeness

- **Positively informative** intercurrent events tend to provide qualitative information about the event of interest
 - Scientific question is what actually happened, including the intercurrent event
 - Goal of improvement is to better incorporate the intercurrent event into the analysis
 - Strategy: composite
- **Counterfactual** intercurrent events tend to confound the event of interest
 - Scientific question is what would have happened if intercurrent event had not occurred.
 - Intercurrent events rendered uninformative conditioned on a model
 - Strategies: hypothetical and principal stratum
- **Uninformative** events permit traditional missing-data assumptions
 - Intercurrent events can be treated as irrelevant to scientific question
 - Strategies:
 - Treatment policy strategy (intercurrent event represents noise)
 - While-on-treatment strategy (occurrence of intercurrent event renders scientific question irrelevant)

Planning for Intercurrent Events

- Predicting what intercurrent events are likely to occur, and whether or not they are likely to be informative, becomes critical in an estimands framework
- Before beginning every clinical trial, it becomes important to the study team, working as an interdisciplinary team including clinicians, statisticians and study operations management, to:
 - Predict and document the types of intercurrent events likely to affect key endpoints in the trial
 - Estimate their likely type and degree of informativeness
 - Develop design and trial conduct strategies for preventing them
 - Develop estimands and analysis strategies for handling them

Collecting Data About Informativeness

- Collecting data about intercurrent events is critical.
- Key needs:
 - Assess the appropriateness of assumptions made in the trial
 - Support the development of more reliable future trials,
 - Data currently collected on withdrawals etc. needs to be rethought and redesigned to meet this new purpose.
- Reasons for patient dropout, failure to come to the clinic, etc. should be documented for all key endpoints
- Reason classes should be formulated by likely causality, to distinguish likely informative reasons (patient gives up on treatment) from likely uninformative reasons (missed visit due to scheduling conflict)
 - An analogy to assigned causality in classifying adverse events might be appropriate

Current Standard: The ITT approach

- The general approach to survival analysis in clinical trials has been in place for some time, with disagreements only on censoring rules and details
- The basic ITT approach in a randomized trial is to:
 - Analyze all randomized patients
 - Use the randomized treatment however a patient was actually treated
- Current EMA guidelines (2012) and FDA guidance (2018) on PFS, follow Fleming et al. (2009) and also recommend.
 - Following patients until event under all circumstances
 - Independent of changes in treatment, missed assessments, etc.
 - On a schedule that's independent of changes in the treatment schedule
 - Censoring at last assessment date in all patients without event

Opening the Box

- The estimands framework opens up the possibility of alternatives to the current standard approach.
- Some of the included strategies have previously not generally been acceptable in a regulatory context
- The estimands framework provides a set of concepts and logic to evaluate when alternative strategies might be acceptable
- When the traditional strategy's assumptions about intercurrent events are particularly inconsistent with the clinical context, alternatives may result in less bias or better reflect the research question.

What strategy does the current ITT approach correspond to?

- The ITT approach endorsed by Fleming et al. (2009) corresponds to a treatment policy approach.
- It ignores intercurrent events, and assumes they are non-informative
- Under the estimands framework, using the strategy requires checking that the underlying assumptions hold, and trial designs to maximize the likelihood of their holding.
- The design requirement is that patients have to consistently follow the assessment schedule, as planned, independently of the treatment schedule.
 - Patients visit the clinic for efficacy assessments on time despite treatment delay
 - Patients continue assessments on schedule despite subsequent therapy
- This is not what most current clinical trials do, despite their use of methods assuming it.

Is treatment-independent follow-up feasible?

- Feasibility of following patients independently of treatment depends on study-specific conditions
- May be infeasible or unethical if patients
 - Enter a new drug trial with new experimental therapy after withdrawal
 - Receive frequent/extensive/intrusive/radioactive imaging or similar methods are required (Continuing invasive/risky assessments beyond benefit may not be ethical).
 - Live far away from the clinic
- May be more likely if
 - Treatment safety delays are not long, and/or intermediate safety visits are appropriate
 - Patients receive subsequent and/or supportive therapy at the same clinic
 - The study's required measurements and clinic visits after treatment withdrawal are not much more frequent/intrusive/difficult than under standard-of-care.
 - Patients are informed about and motivated to continue participation
- There are a variety of other practical considerations

Treatment Policy and Feasibility (cont)

- Trial designers should look carefully at the specific clinical context and decide how likely patients are to remain on assessments.
 - This feasibility evaluation requires careful dialog and cooperation between statisticians, clinicians, and study managers.
- If meeting the design requirements of a treatment policy strategy is not feasible, designers should consider an alternative strategy.
 - The treatment policy strategy, although when biased, might still be better than the alternatives.
 - There often is no perfect strategy given study goals and the clinical trial context.
 - The strategy to select will often be not the ideal but the least bad.

Alternatives to Treatment Policy:

Composite strategy

- If an intercurrent event is expected to be highly correlated with the event of interest, then a composite strategy may be a reasonable approach
- A possible example, in some cases, is clinical progression and radiological progression
- It is plausible that in at least some contexts clinical deterioration is predictive of radiological progression and/or death.
 - There may be cases where it is both infeasible to follow patients beyond clinical progression, and clinical progression is highly informative.
- In these cases, a composite of clinical and radiological progression might be the most reliable endpoint under the circumstances
 - One size does not necessarily fit all.

Treatment Policy and Crossover

- Crossover is a common problem in Phase III oncology clinical trials
- If an oncology treatment has received accelerated approval based on an endpoint like response or PFS, it generally must complete survival trials in order to receive full approval.
- If patients withdraw from study treatment, they generally must be permitted to receive any approved treatment as subsequent therapy.
- This means patients in the survival trial generally have to be permitted to take the preliminarily-approved experimental treatment once they withdraw from the study.
- Quite commonly, a substantial percentage of patients in the comparator arm end up receiving the study treatment as subsequent therapy.

Treatment Policy and Crossover (Cont)

- The estimands framework allows for a hypothetical strategy which potentially includes, under the right conditions, causal inference methods addressing what would have happened if the patient had not changed therapy.
- Causal inference methods (e.g. inverse probability weighting) depend heavily on assumptions, particularly the assumption of no unmeasured confounders.
- Because of heavy dependence on unverified assumptions, regulatory use of a causal-inference based hypothetical strategy may be an uphill climb
 - In oncology we often do not understand what causes the cancer and cannot predict which patients will respond to treatment
 - Assuming we possess all relevant information can be a strong assumption.
- Nonetheless, the estimands framework identifies the steps necessary to establish the appropriateness of a new strategy. It might, perhaps with improved methodology requiring fewer assumptions, lead to eventually bridging the gap.

Trial circumstances influence research purposes

Example: Sequence of Treatments

- In the Bayer 16507 study, 6 doses of Xofigo (Ra-223) at standard dose were compared with either 12 doses at standard dose or 6 doses at a higher dose, all q4w.
 - For the 6 dose/12 dose comparison, both regimens are the same for the first 6 doses (~24 weeks)
- In CAR T-cell therapy (e.g. axicabtagene ciloleucel and tisagenlecleucel), the treatment is manufactured from the patient's blood, with a gap between apheresis and infusion. (Neelapu et al, 2017; Schuster et al., 2019)
 - Patients are given bridging therapy during the time between these two events
- In both cases, patients' therapy in at least one study arm changes after the start of the intervention.

Sequence of Treatments (Cont)

- In both examples, the solution is to redefine both the population of interest and the treatment effect
 - Define the population of interest as the patients who are eligible to start further treatment at the change point
 - In the Xofigo trial, the population is patients who previously received 6 doses of Xofigo.
 - In the CAR-T trials, the population is patients who have received apheresis and completed CAR-T treatment manufacture while receiving SOC therapy in the meanwhile.
 - In both cases, this population is different from the patients who started the trial. It excludes patients who e.g. had events, dropped out, or became ineligible prior to the change point.
 - Define the treatment effect as the effect of treatment after the change point
 - In the Xofigo trial, the trial compares patients receiving up to 6 further doses against patients receiving no further treatment
 - In planned CAR-T trials, the trial compares patients receiving CAR-T treatment against patients receiving comparator therapy.
 - Define the start of TTE endpoints as the time of the therapy change point.

Sequence of Treatments (Cont)

- By redefining the population of interest, the treatment effect, and the start point of TTE variables:
 - The complications of modeling changes in hazards resulting from change in therapy are avoided
 - The ability to retain standard proportional hazards assumptions is preserved.
- The downside is that the research question is redefined into one that may not be of primary interest to a treating physician or patient
 - They may be interested in assessing the complete effect of the entire regimen
 - By starting measuring the treatment effect in the middle of the regimen, this purpose is frustrated.
- It may be necessary to design the study to answer a question that can be answered reliably under the circumstances, not the one we really want to answer.
- It may also be appropriate to have different estimands of the “same” endpoint for different research purposes and audiences.

Intercurrent event related confounding may be a reason to change endpoints

- The Bristol-Myers Squibb CA184024 pivotal trial evaluated ipilimumab+dacarbazine vs. placebo+dacarbazine in metastatic melanoma (Robert et al. 2011)
- This trial began with PFS as primary endpoint
- During the course of the trial, a number of potentially confounding intercurrent events became better known, including
 - Immune-mediated tumor swelling that could mimic progression (Wolchok et al 2009; Seymour et al, 2017).
 - Potential for delayed effects that could potentially reduce power in a faster-maturing endpoint (Siegel 2010)
 - Study design problem: First assessment occurred at week 12, after historical median for dacarbazine (Robert et al., 2011) , similar issues to initial example (Hodi et al., 2010)
- These issues are now standard features of immunotherapy design.

Changing endpoints Cont.

- Primary endpoint was switched to OS mid-study (ClinTrials.gov disclosure April 20, 2009)
- Study based on OS was successful (HR 0.72, $p < 0.001$)
- Secondary PFS was also significant (HR 0.76, $p = 0.006$)
- Although treatment effect was sufficiently strong that PFS benefit remained significant, switch to OS appears to have been prudent
 - Visit schedule had first visit at 12 weeks, similar to first example
 - Occlusion of all events until past median reduced study power, although not enough to prevent significance.
 - Identity of medians is an example of need to ensure visit schedule can support needed estimates.

Greater need for cooperation between clinicians, study managers, and statisticians

- Traditionally, clinicians wrote the study design sections of the protocol and statisticians wrote the analysis sections
 - Decisions about when and how to follow patients were often made in isolation, based on clinician traditions and rules of thumb,
 - Decisions about endpoint definitions and censoring rules were often also made in isolation, based on statistician traditions and rules of thumb
 - Sensitivity analyses were often a slowly growing grab bag of ones traditionally done.
- The estimands framework requires a much greater fit between research question, study design, analysis method, censoring rules, and sensitivity analysis
- This in turn requires much closer dialogue and much greater cooperation between the members of a study design team.
 - They now need to have common objectives and speak a common language.

Need for research into new methods

- The basic impetus for the estimands framework is that traditional statistical methods frequently depend on simplistic assumptions about intercurrent events which are unrealistic in the context of real clinical trials
- Lifetime data analysis, with its dependence on assumptions like non-informative censoring, proportional hazards, etc., is particularly vulnerable to the estimands critique
- New methods are needed that are less dependent on these assumptions.
- Dialogue between practitioners and researchers is critical.

Summary

- The estimands framework subjects common traditional statistical assumptions to empirical verification in a realistic clinical setting.
 - Patients in whom events were not observed cannot simply be assumed to be non-informatively censored.
 - This has a huge potential impact on lifetime data analysis, whose basic methods have historically depended on non-informativity assumptions
- It imposes a new conceptual framework based on strategies for addressing intercurrent events and sensitivity analyses to verify the appropriateness of the strategies used.

Summary cont.

- The first line of defense is the prevention of intercurrent events
- Comprehensive data on intercurrent events needs to be collected
 - Reasons for missed visits
 - Reasons for withdrawal from each type of follow-up (treatment, assessments, survival follow-up) should be
 - Reason categories should be designed to support assigning a likely informativity status where possible.
- In order to assume data collection and censoring are non-informative, the study should be designed, to the extent feasible, so that
 - The data collection schedule is independent of the treatment schedule
 - Patients are followed, to the extent possible, until the event occurs or the study closes, regardless of anything else.

Summary cont.

- The feasibility of complete, treatment-independent follow-up should be dispassionately assessed
 - It may depend on a variety of factors including difficulty and intrusiveness of visit schedule and assessments, the nature of subsequent therapy, and the indication, patient culture, and other factors.
- Where follow-up past key intercurrent events (e.g. clinical progression is infeasible), consider alternatives
- It may sometimes be appropriate to change the research question to one that can be feasibly answered in the context
- It may sometimes be better to try a different endpoint.

Summary cont.

- When the intercurrent events are considered highly predictive of the endpoint of interest, a composite strategy might be considered.
- Although hypothetical strategies require strong assumptions, the estimands framework makes them potential parts of the armamentum.
- The estimands framework requires increased cooperation between statisticians, clinicians, and study operations managers.
- Within statistics, it requires greater cooperation between methods researchers and applied practitioners.
- It opens a path to regulatory consideration of new approaches.
 - But this path is neither a sure nor an easy one.

Conclusions

- The estimands framework will hopefully result in more reliable trials under realistic clinical conditions
- It requires much closer cooperation/interaction between
 - Clinicians, statisticians, and trial managers
 - Statistical theory and practical clinical and operational aspects of clinical trials
- Statistical theory directly affects the choice of endpoints and study design
- The set of alternative strategies widens, and more thought and skill is required to select among them and justify the selection
- New opportunities are opened up.
- A path to regulatory consideration of new approaches is opened. But this path is neither a sure nor an easy one.

Questions?

Thank You!

References

- International council on harmonization of technical requirements for registration of pharmaceuticals for human use, Statistical Principles for Clinical Trials (E9) (1998)
- National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials*. National Academies of Science, Engineering, and Medicine. (2010)
- International council on harmonization of technical requirements for registration of pharmaceuticals for human use, Step 2 Draft Guideline on Estimands and Sensitivity Analyses in Clinical Trial Statistical Principles for Clinical Trials (E9 [R1]) (2017)
- Hodi, FS et al. Improved survival with ipilimumab in patients with metastatic melanoma. *N Engl J Med*. 363:711-23 (2010)
- Rufiback, K. Treatment effect quantification for time-to-event endpoints—Estimands, analysis strategies, and beyond. *Pharmaceutical Statistics*. 18:1–21 (2018).
- European Medicines Association. Appendix 1 to the guideline on the evaluation of anticancer medicinal products in man - methodological consideration for using progression-free survival (PFS) or disease-free survival (DFS) in confirmatory trials. Revision 1 (2012).
- Food and Drug Administration. Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics: Guidance for Industry: December 2018.

References (Cont)

- Fleming, TR, Rothmann, MD, and Lu, HL. Issues in using progression-free survival when evaluating oncology products. *J Clin Oncol* 27:2874-2880 (2009)
- Standard dose versus high dose and versus extended standard dose radium-223 dichloride in castration-resistant prostate cancer metastatic to the bone. ClinTrials.gov ID NCT02023697 (Last update Sep 18 2018).
- Sternberg et al. A randomized phase 2 study investigating 3 dosing regimens of radium-223 dichloride (Ra-223) in bone metastatic castration-resistant prostate cancer (mCRPC). *J Clin Oncol* suppl 36:5008 (ASCO) 2018.
- Neelapu, SS et al. Axicabtagene ciloleucel CAR T-cell therapy in refractory large B-cell lymphoma. *N Engl J Med* 377:2531-2544 (2017)
- Schuster, SS et al. Tisagenlecleucel in adult relapsed or refractory diffuse large B-cell lymphoma. *N Engl J Med* 380:45-56 (2019)
- Robert, C. Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. *N Engl J Med* 364:2517-2526 (2011)
- Wolchok J et al. Guidelines for the evaluation of immune therapy activity in solid tumors: immune-related response criteria *Clin Cancer Res* 15:7412–20 (2009)

References (Cont)

- Seymour, L, et al. iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. *Lancet Oncol* 18: e143–52 (2017)
- Siegel, J. Random delayed effects and the power of the unweighted log-rank test in survival trials. *Joint Statistical Meetings Abstract* #307648 (2010)



Thank you!