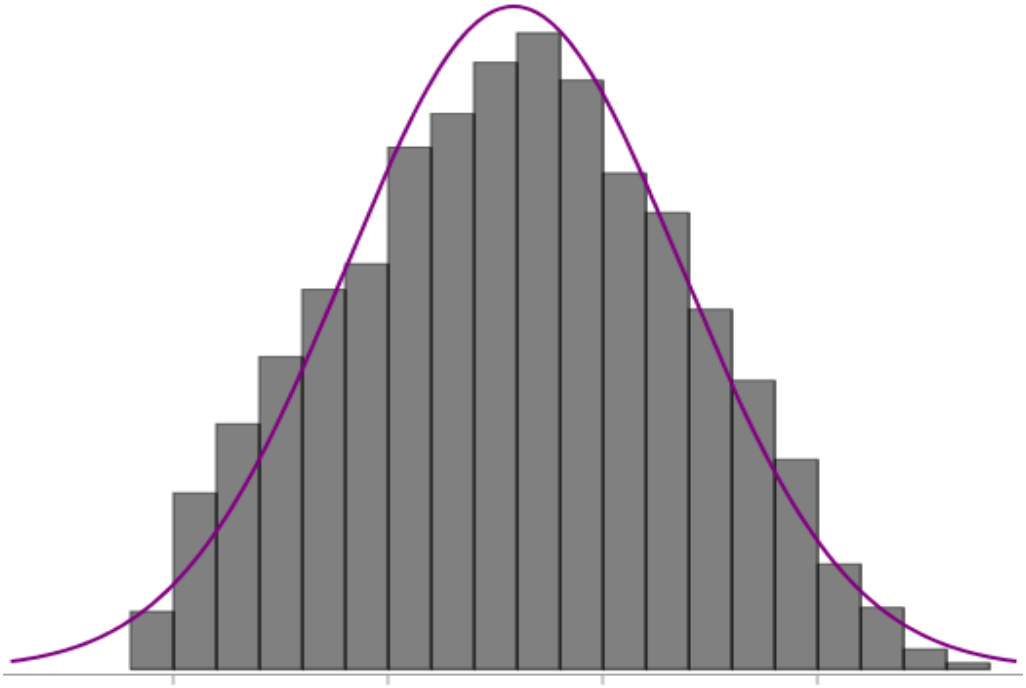
2021

Statistical Concepts in Clinical Research

Denái R. Milton MS
The University of Texas MD Anderson Cancer Center

STATISTICAL CONCEPTS IN CLINICAL RESEARCH



Cancers Diseases Hospital/MD Anderson
Clinical Research Workshop

Prepared by: Denái R. Milton, M.S.
Principal Biostatistician
Department of Biostatistics

Intention

- Seek First to Understand, Then to be Understood (Habit #5)
- Begin with the End in Mind (Habit #2)

The 7 Habits of Highly Effective People **~** Stephen Covey

This reference guide was created to provide understanding of basic statistical concepts used in clinical research to equip you with the skills necessary to succeed in this area including effective collaborations with statisticians.

<u>The overall objectives of the reference guide are:</u>

- To introduce or review concepts to consider when designing a clinical trial;
- To introduce or review the four phases of clinical trials including different types of designs for Phase I and Phase II clinical trials;
- To introduce or review observational studies;
- To introduce or review analysis of categorical, continuous, and time-to-event measures as well as Bayesian methodology.

Thank you for participating in this clinical research workshop and we wish you much success in your careers.

I. Protocol Development
   a. Prospective Studies (concepts to think about when designing a clinical trial/writing a protocol)
      i. Disease to be treated/trial entry criteria.
      ii. Treatments/doses/schedules/treatment combinations.
      iii. Main goal(s) of the trial (i.e., how trial results may be used for planning future studies or changing clinical practice).
      iv. Main clinical outcome(s).
      v. Secondary goals.
      vi. Anticipated accrual rates/sample size.
   b. Observational Studies
   c. Common Efficacy Endpoints - Definitions
   d. Sample Size/Power Determination
II. Clinical Trial Designs
   a. Phase I Dose Finding
      i. Ruled-based designs
      ii. Model-based designs
      iii. Model-assisted designs
   b. Phase II Single Arm Binary
      i. N-stage group sequential designs
      ii. Model-based designs
   c. Phase III Registration
   d. Phase IV Post Marketing
   e. Pilot Studies
III. Observational Studies
   a. Cohort Studies
   b. Case-Control Studies
   c. Cross-Sectional Studies
IV. Data quality
V. Statistical methodology
   a. Analysis of Categorical Measures
      i. Fisher's exact test/chi-squared test
      ii. Logistic regression
   b. Analysis of Continuous Measures
      i. Parametric (e.g., t-test)/non-parametric (e.g., Wilcoxon rank-sum test)
      ii. Linear regression
   c. Analysis of Time-to-event/survival Measures
      i. Kaplan-Meier method
      ii. Cox proportional hazards regression
      iii. Competing risks
   d. Bayesian methods (basic concept)
VI. Miscellaneous (p-values, confidence intervals, multivariate vs. multivariable)

PROTOCOL DEVELOPMENT

1. Protocol Development

Clinical research provides the foundation for the practice of medicine. Ideally, the principles of medical practice should be based on sound scientific rationale and evidence. Clinical studies are useful to the extent that they yield valid inferences. The goal of proper study design is to minimize the errors that threaten the scientific validity of conclusions based on these inferences.

1.1. Prospective Studies

Well-conducted randomized controlled trials, with adequate numbers of subjects; blinding of therapies, subjects and researchers; and carefully standardized methods of measurement and analysis are the best evidence for a cause-and-effect relationship.

The protocol is a document that describes how a clinical trial will be conducted and ensures the safety of the trial subjects and integrity of the data collected. Protocols should be clear, unambiguous and maintain scientific integrity. The protocol should describe the background, rationale, objectives, design, methodology, statistical considerations, and organization of a clinical study.

Some concepts to think about when designing a clinical trial/writing a protocol:

1. Main Goal(s) of the Trial
    a. How the trial results may be used for planning future studies or changing clinical practice.
    b. What are the primary objective(s) and endpoint(s) (i.e., what is the primary question you would like the study to address)?

2. Other Goals of the Trial
    a. What are the secondary and exploratory objectives and endpoints?

3. Study Design
    a. What is the best study design to address the study objectives?
    b. What disease group(s) are of interest (inclusion/exclusion criteria)?
    c. What treatments/doses/schedules/combinations will be investigated (best control group for study population)?
    d. How many visits are required?
    e. What assessments will be administered at each visit?
    f. **How many subjects are needed to address the primary objective/how many are possible based on funding?**
    g. How many sites will be participating?
    h. What's the anticipated accrual rate?

4. How will the data be collected?
    a. Excel, REDCap, Prometheus, DMI, MOCLIP

1.2. *Observational Studies*

Studies that do not use random assignment to allocate subjects into comparative groups are collectively referred to as non-experimental or observational studies. Observational studies are also non-interventional,

meaning the treatment and care of the subject are not influenced by the study but are conducted as in usual practice. These studies reflect less artificial and more naturalistic circumstances; people's lives and behaviors are not being modified by restrictive rules or specific recommendations, and the natural history of disease occurrence and progression can be better observed. As such, observational studies may provide opportunities to evaluate the effectiveness of treatment in people who are more like those who are in need of treatment in the community (i.e., more generalizable). Similar to clinical trials, protocols for observational studies should be clear, unambiguous and maintain scientific integrity. The protocol should describe the background, rationale, objectives, methodology, and statistical considerations for the study.

1. Main Question the Study will Address
    a. What disease group(s) are of interest?
    b. How the study results may be used for planning future studies or changing clinical practice.
    c. What are the primary objective(s) and endpoint(s)?

2. Other Goals of the Study
    a. What are the secondary and exploratory objectives and endpoints?

3. Sample Size Determination

4. Statistical Methodology
    a. Addressing potential biases

1.3. *Study Objectives vs. Endpoints*

The objective of a study is an active statement about how the study will address specific research question(s). For example, a primary objective of a study could be to compare the efficacy of Drug X to Drug Y in subjects diagnosed with multiple myeloma. There could be many endpoints for this objective, including overall survival, progression-free survival or objective response rate. An endpoint is not an objective but can be included in the objective; the primary objective of the study is to compare **overall survival** of Drug X to Drug Y in subjects diagnosed with multiple myeloma.

1.4. *Common Efficacy Endpoints*

Endpoints	Definition	Advantages	Limitations
Overall survival (OS)	Time from randomization/start of treatment until death from any cause	• Universally accepted measure of direct benefit • Easily and precisely measured	• May require a larger trial population and longer follow-up to show statistical difference between groups • May be affected by crossover or subsequent therapies • Includes deaths unrelated to cancer
Progression-free survival (PFS)	Time from randomization/start of treatment until disease progression or death	• Requires small sample size and shorter follow-up time compared with OS • Includes measurement of stable disease (SD) • Not affected by crossover or subsequent therapies • Generally based on objective and quantitative assessment	• Validation as a surrogate for survival can be difficult in some treatment settings • Not precisely measured (i.e., measurement may be subject to bias) • Definition may vary among trials • Requires frequent radiologic or other assessments • Requires balanced timing of assessment among treatment arms
Time to progression (TTP)	Time from randomization/start of treatment until objective tumor progression; does **not** include deaths		
Recurrence-free survival (RFS)	Time from date of response to the first of either recurrence or relapse, second cancer, or death	• Similar to PFS; may be useful in evaluation of	• Similar to PFS

		highly toxic therapies	
Objective response rate (ORR)	Proportion of subjects with reduction in tumor burden of a predefined amount (typically includes complete remission and partial response)	• Can be assessed in single-arm trials • Requires a smaller population and can be assessed earlier, compared with survival trials • Effect is attributable directly to the drug, not the natural history of the disease	• Not a comprehensive measure of drug activity
Duration of response (DoR)	Time from documentation of tumor response to disease progression		

1.5. Sample Size/Power Determination

The field of statistics exists because it is usually impossible to collect data from all individuals of interest (population). Thus, the only solution is to collect data from a subset (sample) of the individuals of interest, but the real desire is to know the "truth" about the population. It is imperative in medical research to ensure that reported comparisons are based on a sufficient number of subjects to be statistically valid. Small samples may lack sufficient statistical power to detect important differences or associations.

1.5.1. Hypothesis Testing

Hypothesis testing is a process in statistics whereby an assumption regarding a population parameter is tested. Hypothesis testing is used to assess the plausibility of a hypothesis (i.e., assumption) by using sample data. In hypothesis testing, there are the null hypothesis (H_0) and the alternative hypothesis (H_1 or H_a). The null hypothesis is usually a hypothesis of equality between population parameters (e.g., there is no difference in response rate between the experimental treatment and standard of care). The alternative hypothesis is effectively the opposite of a null hypothesis (e.g., the response rate in the experimental treatment is higher than the standard of care). Thus, they are mutually exclusive, and only one can be true.

The following table presents a 2x2 representation of the truth from the population and the decision based on the study sample. The significance (α) level is the probability that the decision based on the study sample is that there is a difference when in fact there is no difference. Whereas, the power is the probability that the decision based on the sample correctly concludes there is a difference.

	DECISION (based on study sample)	
TRUTH (population)	There is no difference	There is a difference
There is no difference (H_0)		**Significance (α) level**
There is a difference (H_a)		**Power**

A scientific conclusion is always drawn from the statistical testing of hypothesis, in which the chosen significance (α) level, is used for decision-making. However, the probability of committing false statistical inferences is considerably increased when more than one hypothesis is simultaneously tested, which therefore requires proper adjustment of the significance level.

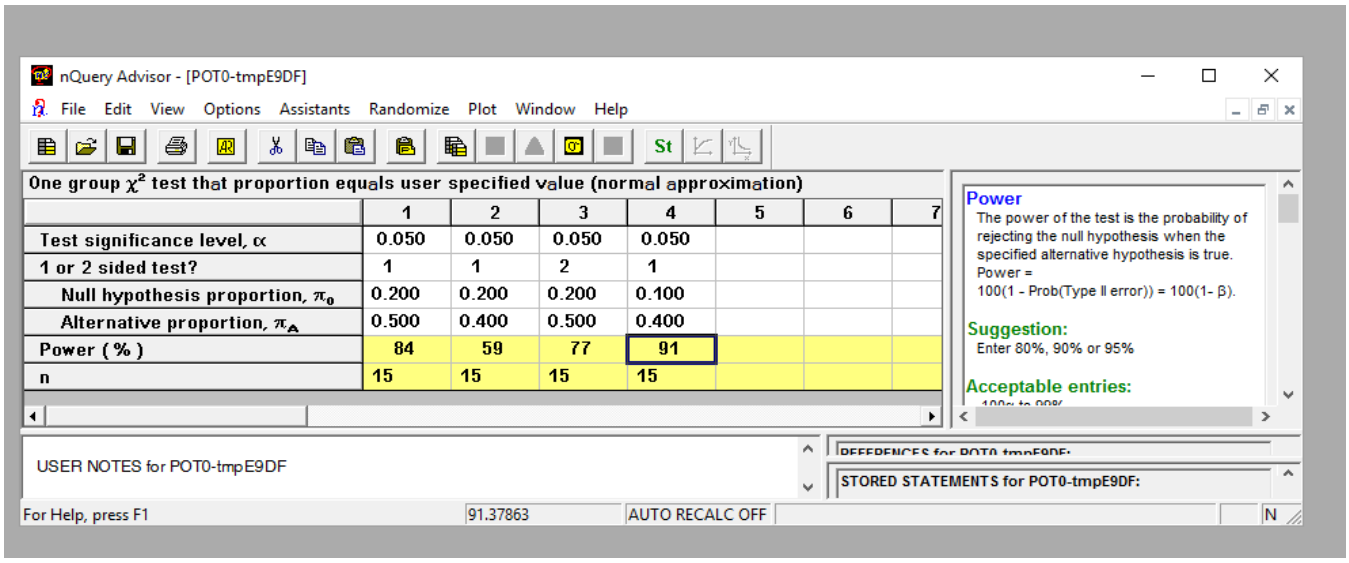
Examples

If the primary objective of your study is to test if an experimental drug is superior to the standard of care or a historical control, the sample size/power computation required will be based on hypothesis testing.

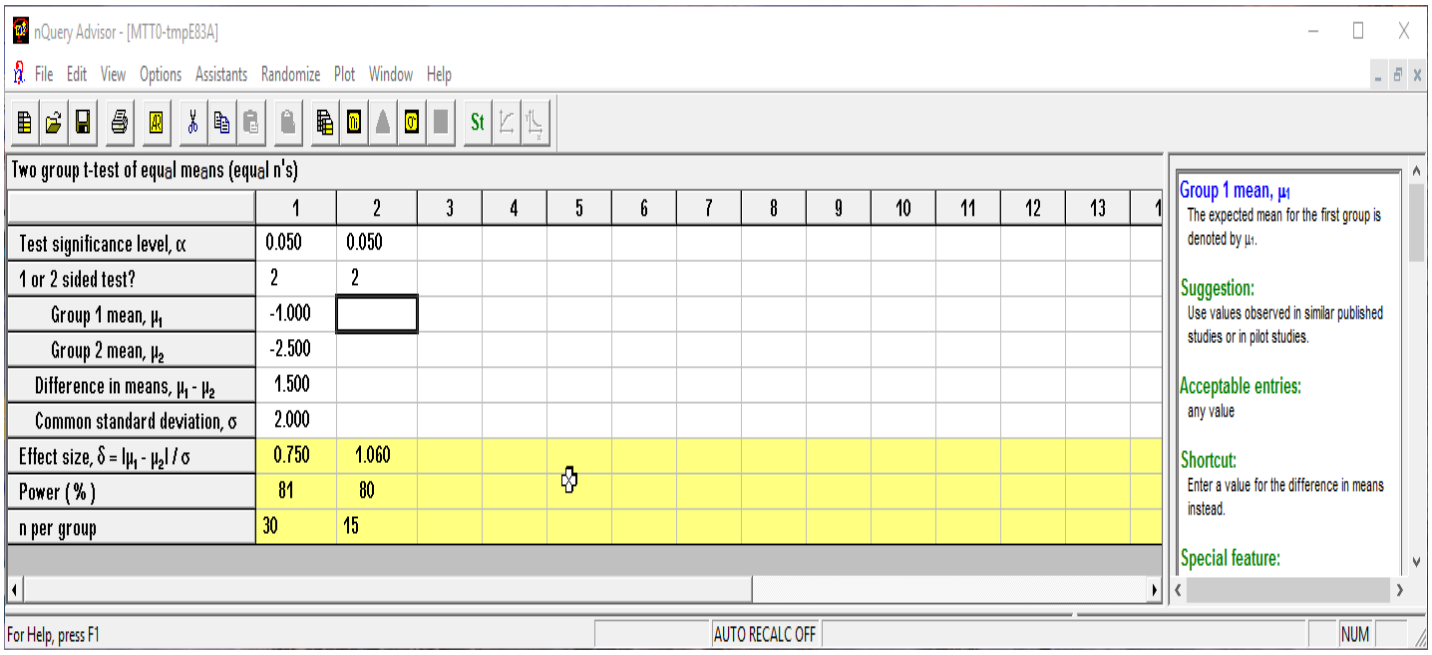
Information required for sample size/power computations includes:

a. The number of groups (e.g., one: experimental vs. historical control; two: experimental vs. active control [e.g., standard of care]).
    - Obtained from the study design.
b. Parameter estimates for each group (e.g., response rates, change in tumor size, standard deviation of tumor size, effect size [i.e., the absolute value of difference between group means divided by the common standard deviation], median survival or survival rates).
    - Primary objective endpoint(s) obtained from the primary study objective. Estimates are obtained from literature or previous studies. Often, this information is unknown. In those cases, use estimates that are considered to be clinically meaningful.
c. Significance level or *α* level (one-sided or two-sided).
    - Typically 5%; one-sided or two-sided depends on whether your alternative hypothesis is $ORR_A > ORR_B$ (one-sided), $ORR_A < ORR_B$ (one-sided), or $ORR_A \neq ORR_B$ (two-sided).
d. Power or sample size (depending on what is being computed)
    - ≥ 80% power is common; although 70% is also acceptable for Phase II trials

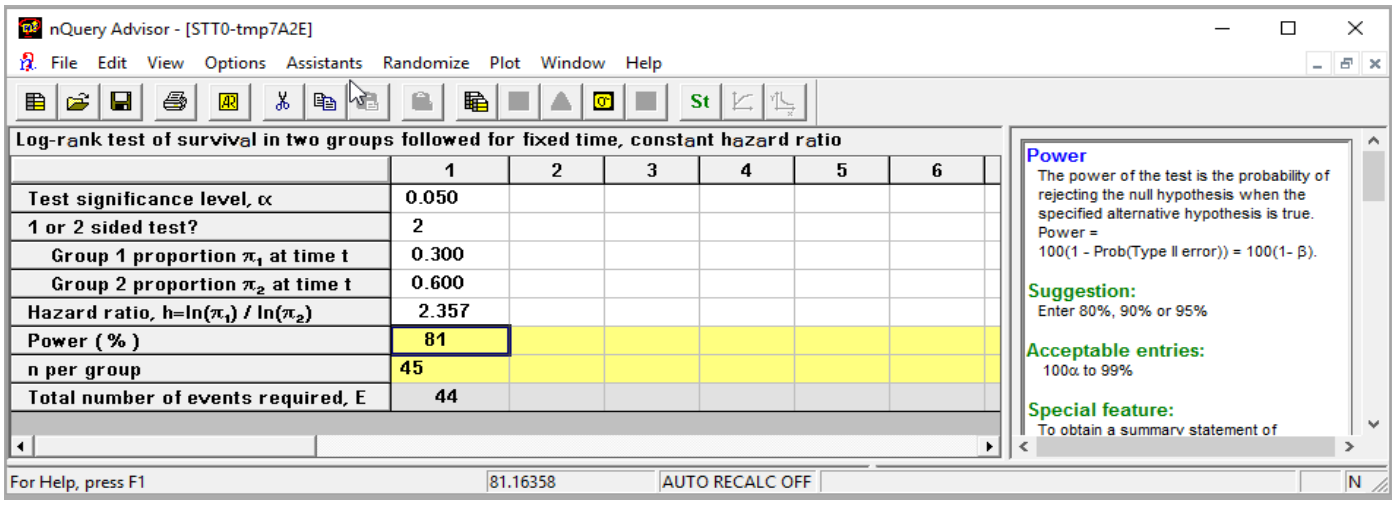
Example for one group vs. historical control for response rates:



Example for two groups for change in tumor size:



Example for two groups for survival:



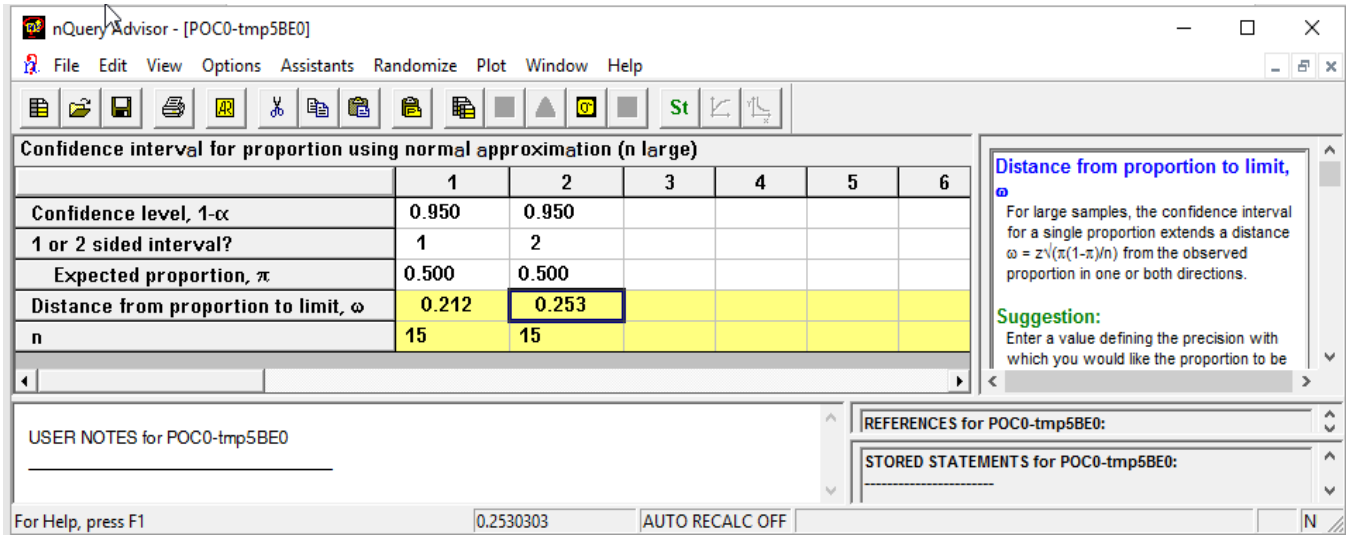
1.5.2 Estimation – Confidence Interval

If the primary objective of your study is to estimate a parameter, perhaps because this is a phase II study with small sample sizes, then your sample size justification will be based on precision via a confidence interval.

Information required for precision computations includes:

a. The number of groups (e.g., one: experimental vs. historical control; two: experimental vs. control).
   - Obtained from the study design
b. Parameter estimates (e.g., proportion for binary, effect size for continuous)
c. Confidence interval (one-sided or two-sided).
   - 95% is common; one-sided or two-sided

Example for one group for response rate:



CLINICAL TRIAL DESIGNS

A clinical trial is a planned prospective experiment involving human subjects from a specified population designed to evaluate an intervention in order to determine appropriate interventions for future members from the same population.

1. Phase I (Dose Finding) Clinical Trials

The focus in phase I trials is looking at what the drug does to the body and what the body does with the drug. Phase I trials aim to find the best dose of a new drug with the fewest side effects. These studies also determine how a drug is absorbed, distributed, metabolized and excreted as well as the duration of its action. Very low doses of the drug are given initially to subjects while higher doses are given to subsequent subjects until side effects become too severe or the desired effect is seen. The number of subjects included in phase I trials are typically small (e.g., 15 to 30). The drug may help subjects, but phase I trials are to test a drug's safety as opposed to efficacy. If a drug is found to be safe enough, it can be tested in a phase II clinical trial.

A phase I trial design has many components, including starting dose, dose increment, dose escalation method, number of subjects per dose level, specification of dose-limiting toxicities (DLT) and assessment period, target toxicity level, definition of the maximum tolerated dose (MTD) and recommended dose for phase II trials (RP2D).

Dose escalation methods for phase I cancer clinical trials fall into two broad classes: rule-based designs, which include the traditional 3+3 design and its variations, and model-based designs. Rule-based designs assign subjects to dose levels according to pre-specified rules based on actual observations of target events (e.g., the DLT) from the clinical data. Typically, the MTD or RP2D is determined by the pre-specified rules as well. On the other hand, the model-based designs assign subjects to dose levels and define the RP2D based on the estimation of the target toxicity level by a model depicting the dose–toxicity relationship.

1.1. Algorithm (Rule)-Based Designs

Algorithm-based designs are a class of conventional designs that use a set of simple, prespecified rules to determine the dose escalation and de-escalation. Examples include the conventional 3+3 design and its extensions, such as the accelerated titration design and the rolling 6 design. The conventional 3+3 design remains the predominant method for conducting phase I cancer clinical trials. It requires no modeling of the dose–toxicity curve beyond the classical assumption for cytotoxic drugs that toxicity increases with dose. This algorithm-based design proceeds with cohorts of three subjects; the first cohort is treated at a starting dose that is considered to be safe based on extrapolation from animal toxicological data, and the subsequent cohorts are treated at increasing dose levels that have been fixed in advance. The traditional 3+3 algorithm is described below.

- Enroll 3 subjects at the starting dose level
- If 0 of the 3 subjects experiences a DLT at a given dose level, proceed to the next higher dose level with a cohort of 3 subjects
- If 1 of 3 subjects experiences a DLT at a given dose level, enter 3 additional subjects at the current dose level
- If 1 of 6 subjects experiences a DLT at a given dose level, proceed to the next higher dose level with a cohort of 3 subjects

- If at least 2 of 3 or 2 of 6 subjects experience a DLT at a given dose level, then the MTD has been exceeded
- Once the MTD has been exceeded, treat another 3 subjects at the previous dose level if there were only 3 subjects treated at that dose level
- The MTD is the highest dose level in which 6 subjects have been treated with at most 1 experiencing a DLT

The main advantages of algorithm-based methods are that they are easy to implement and do not require special software. However, their performance (operating characteristics) is not guaranteed and they have some drawbacks. For example, these designs may be inefficient in establishing the dose that meets a specific target toxicity level. In addition, the decision of dose allocation for future subjects as well as the definition of the RP2D rely on information from the current dose level and do not use all available information. As such, the RP2D is then selected from the pre-specified dose levels depending on which one best fits the definition of acceptable toxicity set *a priori*. However, although not ideal, the algorithm-based methods have been successful in establishing safe recommended doses for phase II trials during the past several decades for anticancer agents that were eventually used worldwide in clinical practice.

1.2. Model-Based Designs

An alternative dose escalation method for phase I clinical trials is to use statistical models that actively seek a dose level that produces a prespecified probability of dose-limiting toxicity by using toxicity data from all enrolled subjects to compute a more precise dose–toxicity curve. This method is typically carried out using Bayesian models. Bayesian models require an initial estimation of DLT rate (also called prior distribution of θ), which characterizes the shape of the dose–toxicity curve. The occurrence of toxicity (or not) in subjects enrolled at each dose level provides additional information for the statistical model and results in an adjustment of θ (i.e., posterior distribution of θ) according to Bayes' theorem. The posterior distribution is then evaluated to identify the dose closest to the target toxicity level, and this dose is used to treat future subjects and to set the recommended dose for phase II trials. These model-based designs use all of the available data to model the dose–toxicity curve, and they provide a confidence interval for the RP2D at the end of the trial.

Some model-based designs include continual reassessment method (CRM; O'Quigley, J., Pepe, M., Fisher, L. Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics* 1990; 46, 33-48) and EffTox method (Thall, Peter F., Cook, John D. Dose-Finding Based on Efficacy-Toxicity Trade-Offs. *Biometrics* 2004; 60, 684-693).

1.3. Model-Assisted Designs

Model-assisted designs were developed to combine the advantages of algorithm-based designs and model-based designs (Yuan, Ying, Lee, Jack J., Hilsenbeck, Susan G. Model-Assisted Designs for Early-Phase Clinical Trials: Simplicity Meets Superiority. JCO Precision Oncology 2019;3, 1-12). Similar to the model-based design, the model-assisted design uses a statistical model (e.g., the binomial model) to derive the design for efficient decision making; however, like the algorithm-based design, its dose escalation and de-escalation rule can be predetermined before the onset of the trial and, thus, can be implemented in as simple a way as the algorithm-based designs.

Some model-assisted designs include modified toxicity probability interval method (mTPI; Ji Y, Liu P, Li Y, Bekele BN. A modified toxicity probability interval method for dose-finding trials. *Clinical Trials* 2010;

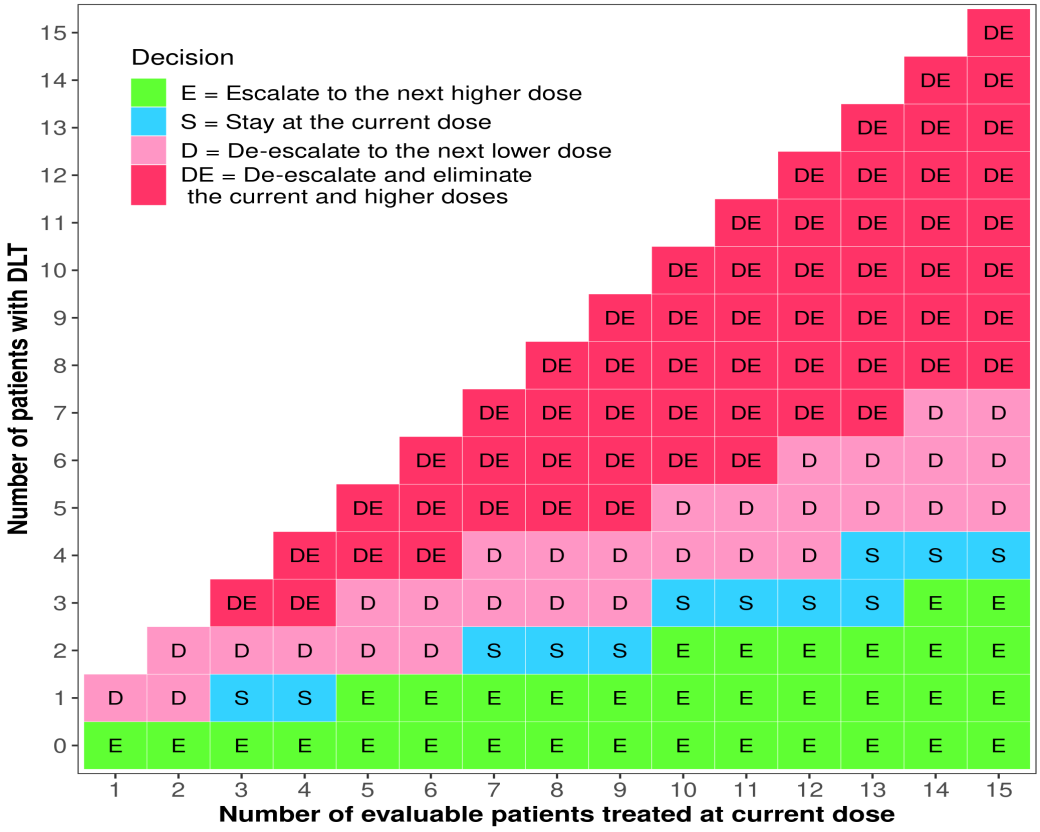
7: 653-663) and Bayesian optimal interval (BOIN; Liu S. and Yuan, Y. Bayesian Optimal Interval Designs for Phase I Clinical Trials. *Journal of the Royal Statistical Society: Series C* 2015*;* 64, 507-523).

1.3.1. Bayesian Optimal Interval (BOIN) Design *(Example)*

The BOIN design is implemented in a simple way similar to the traditional 3+3 design, but is more flexible and possesses superior operating characteristics. In this example, the phase I trial explores 5 dose levels and the maximally accepted DLT rate is 30%. A total of 15 subjects will be enrolled in cohorts of size 3 starting at the 2nd dose level.

The target toxicity rate for the MTD is $\phi = 0.3$ and the maximum sample size is 15. We will enroll and treat subjects in cohorts of size 3. To guide dose-escalation decisions, if the observed DLT rate at the current dose is ≤ 0.236, the next cohort of subjects will be treated at the next higher dose level; if it is ≥ 0.359, the next cohort of subjects will be treated at the next lower dose level. For the purpose of overdose control, doses j and higher levels will be eliminated from further examination if $\Pr(p_j > 0.3 \mid \text{data}) > 0.95$, where p_j is the true DLT rate of dose level $j, j = 1,\cdots, 5$. When the lowest dose is eliminated, stop the trial for safety. Figure 1 presents the dose escalation/de-escalation rules for the study. Note that although subjects are enrolled in cohorts of size 3, Figure 1 includes decision rules for all subjects (i.e., it is not necessary to wait for all of the subjects in the next cohort to complete the DLT assessment period before making decisions for each enrolled subject). For example, if 4 subjects have been enrolled in the trial with one subject experiencing a DLT, the decision for the next subject would be to remain at the current dose level.

Figure 1. Dose escalation/de-escalation rules for the BOIN design



After the trial is completed, select the MTD based on isotonic regression as specified in Liu and Yuan (2015). This computation is implemented by the shiny app "BOIN" available at http://www.trialdesign.org. Specifically, select as the MTD the dose for which the isotonic estimate of the toxicity rate is closest to the target toxicity rate. If there are ties, select the higher dose level when the isotonic estimate is lower than the target toxicity rate and select the lower dose level when the isotonic estimate is greater than or equal to the target toxicity rate.

Operating Characteristics

Table 1 shows the operating characteristics of the trial design based on 1000 simulations of the trial using shiny app "BOIN" available at http://www.trialdesign.org. The operating characteristics show that the design selects the true MTD, if any, with high probability to the dose levels with the DLT rate closest to the target of 0.3.

Table 1. Operating characteristics of the BOIN design

	Dose 1	Dose 2	Dose 3	Dose 4	Dose 5	Number of Subjects	% Early Stopping
Scenario 1							
True DLT rate	0.01	0.05	0.1	0.2	0.3		
Selection %	0	0.7	12.8	41.4	45.1		0
% Pts treated	0	24.1	30.0	29.4	16.5	15	
Scenario 2							
True DLT rate	0.05	0.1	0.2	0.3	0.4		
Selection %	0.4	10.8	30.4	39.6	18.8		0
% Pts treated	1.0	31.8	36.2	23.7	7.3	15	
Scenario 3							
True DLT rate	0.1	0.2	0.3	0.4	0.5		
Selection %	6.3	31.1	36.1	22.9	3.6		0
% Pts treated	6.4	44.5	33.8	13.6	1.8	15	
Scenario 4							
True DLT rate	0.2	0.3	0.4	0.5	0.6		
Selection %	23.6	45.8	21.7	7.7	0.6		0.6
% Pts treated	17.3	52.0	24.5	5.6	0.6	15	
Scenario 5							
True DLT rate	0.3	0.4	0.5	0.6	0.7		
Selection %	46.0	35.8	9.6	1.7	0		6.9

% Pts treated	33.1	50.4	14.6	1.8	0.01	14.7	
Scenario 6							
True DLT rate	0.5	0.6	0.7	0.8	0.9		
Selection %	41.8	8.1	0.3	0	0		49.8
% Pts treated	55.7	40.8	3.4	0.1	0	12.7	

Simulations have shown that model-based and model-assisted methods, which use all toxicity information accumulated during the trial, achieve good estimations of the target probability of dose-limiting toxicity at the RP2D without treating too many subjects at suboptimal doses. Some of the challenges presented by some model-based designs include the need for biostatistical expertise and available software on site to perform model fitting in real time, as well as an expedited collection of data from each cohort of subjects to fit the model. As such, implementation of these designs may not be straightforward. In addition, the model may fail to reach the RP2D if the prior distributions for the parameters of the dose–toxicity curve are inadequate, or conversely, if the prior assumptions are overbearing.

2. Phase II Clinical Trials

Phase II trials further assess safety as well as if a drug works. The drug is often tested among subjects with a specific type of cancer. Phase II trials are done in larger groups of subjects compared to phase I trials. Subjects are closely watched to see if the drug works. If a drug is found to work, it can be tested and compared to the current (standard-of-care) drug in a phase III clinical trial.

Since the number of subjects included in phase I trials is typically small (e.g., 15 to 30) and the safety profile of the drug is limited to a few subjects evaluated at the MTD/RP2D, many phase II trials include formal toxicity and/or futility monitoring which will stop the trial early (i.e., before all subjects have been enrolled) if the accumulated data indicate the likelihood of excessive toxicity is high and the likelihood of acceptable efficacy is low, respectively.

2.1. N-stage Group Sequential Designs

In an N-stage design, the subjects are enrolled in N stages with a binary endpoint. For a two-stage design, at the completion of the first stage, an interim analysis is performed to determine if the second stage should be conducted. The endpoint typically evaluated is response rate (responders vs. non-responders). If the number of subjects responding is greater than a certain amount, the second stage is conducted. Otherwise, it is not.

2.1.1. Simon’s Two-Stage Design (Example)

The primary objective of this example study is to assess the efficacy of Drug X in subjects with melanoma and brain metastases. The primary endpoint is the ORR to this regimen defined as the percentage of number of complete response or partial response in total number of subjects treated. The trial will be conducted by the Simon's optimal two-stage design and the ORR will be estimated accordingly.

It is assumed that Drug X will have a target ORR of 35%. An ORR of 19% or lower is considered a failure

and Drug X will be rejected under this circumstance. When the probability of accepting a 'bad' regimen (i.e. ORR ≤ 19%) is 0.10 and the probability of rejecting a 'good' regimen (i.e. ORR ≥ 35%) is also 0.10, Simon's design requires 23 subjects to enter in the first stage. If 4 or fewer subjects respond to the treatment, the trial will be stopped and the regimen will be declared as ineffective. If there are 5 or more responses, 34 more subjects will be entered in the study to reach a total of 57 subjects. At the end of the study, Drug X will be rejected if ORR is less than or equal to 14/57 and will be accepted otherwise. The operating characteristics of the trial are given as follows: When the true ORR is 0.19 the probability of stopping the trial early is 55%. On the other hand, if the true ORR is 0.35, the probability to stop the trial early is 6%. The expected sample sizes are 38.3 and 55.1 when the true ORRs are 0.19 and 0.35, respectively.

This design has the optimal property of minimizing the expected sample size under the null hypothesis that the new regimen is ineffective.

2.2. Model-Based Designs

As in phase I designs, an alternative monitoring method for phase II clinical trials is to use statistical models. Some model-based designs include methods by Thall et. al. (Thall PF, Simon RM, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in Medicine* 1995; 14:357-79 and Thall, PF and Sung, H-G. Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. *Statistics in Medicine* 1998; 17:1563-1580), Bayesian predictive probability method (Lee JJ, Liu DD. A predictive probability design for phase II cancer clinical trials. *Clinical Trials* 2008; 5(2):93-106), and Bayesian optimal phase 2 (BOP2; Zhou, H., Lee, J. J., & Yuan, Y. BOP2: Bayesian optimal design for phase II clinical trials with simple and complex endpoints. *Statistics in Medicine* 2017; 36(21):3302-3314).

2.2.1. Bayesian Optimal Phase 2 (BOP2) Design (Example)

In this example study, formal monitoring of safety and efficacy will be performed simultaneously after the first 9 subjects in cohorts of size 3 using the Bayesian optimal phase 2 (BOP2) design (Zhou, Lee and Yuan, 2017). A maximum of 30 subjects will be enrolled and the efficacy endpoint is the ORR and safety endpoint is DLT rate:

We will simultaneously monitor efficacy and safety endpoints. Specifically, let n denote the interim sample size and N denote the maximum sample size. Let Y_{eff} and Y_{tox} respectively denote the efficacy and toxic endpoints, with $Y_{eff} = 1$ and $Y_{tox} = 1$ respectively indicating that subjects experience efficacy and toxicity. Let $p_{eff} = Pr(Y_1 = 1)$, $p_{tox} = Pr(Y_2 = 1)$ and define the null hypothesis H_0: $p_{eff} \leq 0.1$ and $p_{tox} > 0.25$, representing that the treatment is inefficacious or overly toxic. We will stop enrolling subjects and claim that the treatment combination is not promising if

$$Pr(p_{eff} > 0.1|data) < \lambda(\frac{n}{N})^{\alpha},$$

or

$$Pr(p_{tox} \leq 0.25|data) < \lambda(\frac{n}{N})^{\alpha},$$

where λ=0.65 and α=0.9 are design parameters optimized to minimize the chance of incorrectly claiming that an efficacious and safe treatment is unacceptable (i.e., type II error) under the alternative hypothesis

$H_1: p_{eff} = 0.25$ and $p_{tox} = 0.1$, while controlling the type I error rate at 0.05 (i.e., the chance of incorrectly claiming that an inefficacious or overly toxic treatment is acceptable is no more than 5%). Assuming a Dirichlet prior distribution $Dir(0.05,0.05,0.2,0.7)$ for the treatment effect, the above decision rule corresponds to the following stopping boundaries and yields a statistical power of 0.8354 under H_1:

Table 2: Optimized stopping boundaries

# Subjects treated	Stop if # response ≤	OR # toxicity ≥
12	0	5
15	1	5
18	1	6
21	2	6
24	2	6
27	3	7
30	3	7

Based on Table 2, we will perform the interim analysis when the number of enrolled subjects reaches 12, 15, 18, 21, 24, 27. When the total number of subjects reaches the maximum sample size of 30, we will reject the null hypothesis and conclude that the treatment combination is acceptable if the number of responses in the efficacy endpoint are greater than 3, and the number of toxicities are less than 7; otherwise we will conclude that the treatment combination is unacceptable.

Below are the operating characteristics of the design based on 10000 simulations using the BOP2 web application, which is available at http://www.trialdesign.org.

Table 3: Operating characteristics

Pr(Eff)	Pr(Tox)	Pr(Eff & Tox)	Early stopping (%)	Claim acceptable (%)	Sample size
0.10	0.10	0.05	80.60	18.92	18.9
0.10	0.25	0.05	93.87	4.90	16.3
0.10	0.40	0.05	99.56	0.24	13.5
0.25	0.10	0.05	15.76	83.54	27.9
0.25	0.25	0.05	65.70	27.69	21.7
0.25	0.40	0.05	97.45	1.22	14.8
0.40	0.10	0.05	4.12	95.09	29.5
0.40	0.25	0.05	62.23	30.45	22.5
0.40	0.40	0.05	97.03	1.54	14.9

3. Phase III Clinical Trials

Phase III trials compare a new drug to the standard-of-care drug. These trials assess the side effects of each drug and which drug works better. Phase III trials enroll 100 or more subjects. Often, these trials are randomized. This means that subjects are put into a treatment group, by chance. Randomization is

needed to make sure that the people in all trial arms are alike. This ensures that the results of the clinical trial are due to the treatment and not differences between the groups. In the US, when phase III clinical trials (or sometimes phase II studies) show a new drug is more effective and/or safer than the current standard-of-care drug, a new drug application (NDA) is submitted to the Food and Drug Administration (FDA) for approval. The FDA then reviews the results from the clinical trials and other relevant information. Based on the review, the FDA decides whether to approve the treatment for use in subjects with the type of illness on which the drug was tested. If approved, the new treatment often becomes a standard of care, and newer drugs must often be tested against it before being approved.

In many instances, some routine monitoring of trial progress, usually blinded to treatment allocation, is often undertaken as part of a phase III trial. Such monitoring may be undertaken in conjunction with a data and safety monitoring board (DSMB), established to review the information collected. It would therefore appear that assessment of interim treatment differences is a logical and worthwhile extension. However, the handling of treatment comparisons while a trial is still in progress poses problems in medical ethics, statistical analysis and practical organization.

The most appealing reason for monitoring trial data for treatment differences is that, ethically, it is desirable to terminate or change a trial when evidence has emerged that one treatment is clearly superior to the other. This is particularly important when life-threatening diseases are involved. Alternatively, the data may support the conclusion that the experimental treatment and the control do not differ by some predetermined clinically relevant magnitude, in which case it would be desirable, both ethically and economically, to stop the study and divert resources elsewhere. Finally, if information in a trial is accruing more slowly than expected, perhaps because of a low event rate, then extension of recruitment until a large enough sample has been recruited may be appropriate.

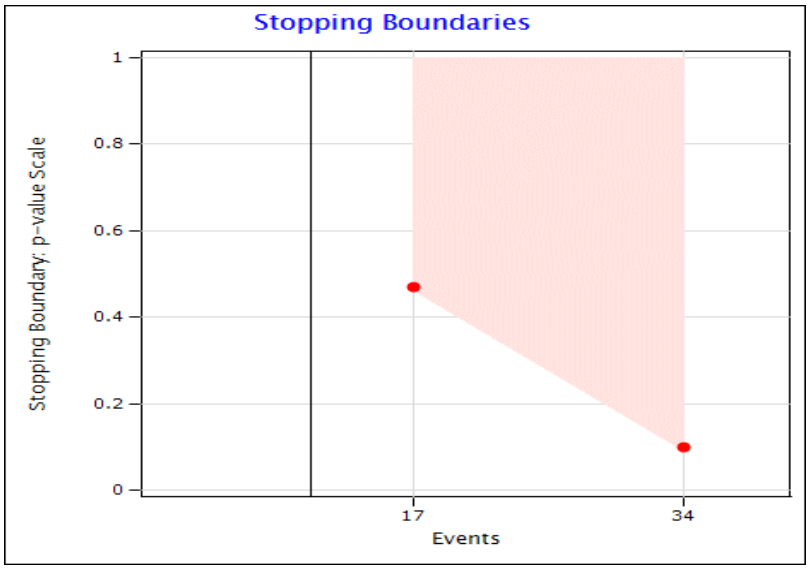
However, multiple analyses of accumulating data lead to problems in the interpretation of results. The main problem occurs when significance testing is undertaken at the various interim looks. Even if the treatments are really equally effective, the more often one analyzes the accumulating data, the greater the chance of eventually and wrongly detecting a difference, thereby drawing incorrect conclusions from the trial.

A second problem concerns the final analysis. When data are inspected at interim looks, the analysis appropriate for fixed sample size studies is no longer valid. Quantities such as *P* values, point estimates and confidence intervals are still well defined, but new methods of calculation are required. If a traditional analysis is performed at the end of a trial that stops because the experimental treatment is found better than control, the *P* value will be too small (too significant), the point estimate too large and the confidence interval too narrow. To remedy these problems, special techniques are required. These can be broadly termed *sequential methods*.

A sequential test monitors a statistic summarizing the current difference between the experimental treatment and control at a series of times during the trial. If the value of this statistic crosses some specified critical value (i.e., stopping rule or boundary), the trial is stopped and an appropriate conclusion drawn. It is possible to look after every subject or to have just one or two interim analyses. If the statistic stays within the test boundary then there is not enough evidence to come to a conclusion at present and a further interim look should be taken. It is the details of the derivation of the stopping rule that introduces much of the variety of sequential methodology. Key early work in the area includes the tests of Pocock and O'Brien & Fleming. A more flexible approach, referred to as the alpha-spending method was proposed by Lan & DeMets and extended by Kim & DeMets.

3.1. Sequential Design – Example

A sample size of 40 subjects will be randomized 1:1 to Drug A or control. Assume the control arm's true median PFS will be approximately 4 months. Assume median PFS in the Drug A arm will be 8.5 months, corresponding to a hazard ratio of 0.471 (Drug A vs. control). The study is anticipated to finish accrual at 36 months with an additional 6 months follow up for a total study duration of 42 months. Under these conditions, we will have 80% power using a 1-sided log rank test with type I error of 0.10. Uniform accrual and exponential PFS distributions are assumed and the expected total PFS event count is 34. We will have an interim look for futility once half the expected events (17 events) are observed. A Lan-Demets spending function using an Obrien-Flemming boundary will be used for futility stopping boundaries. **We will stop for futility at our interim look if our p-value is greater than 0.470.**



4. Phase IV Clinical Trials *(Post Marketing)*

Drugs approved by the FDA are often watched over a long period of time in phase IV studies. Even after testing a new medicine on thousands of people, the full effects of the treatment may not be known. Some questions may still need to be answered. For example, a drug may get FDA approval because it was shown to reduce the risk of cancer coming back after treatment. But does this mean that those who get it are more likely to live longer? Are there rare side effects that haven't been seen yet, or side effects that only show up after a person has taken the drug for a long time? These types of questions may take many more years to answer, and are often addressed in phase IV clinical trials. These studies may also look at other aspects of the treatment, such as quality of life or cost effectiveness.

5. Pilot Studies

Pilot studies represent a fundamental phase of the research process. The purpose of conducting a pilot study is to examine the feasibility of an approach that is intended to be used in a larger scale/main study. A pilot study can be used to evaluate the feasibility of recruitment, randomization, retention, assessment procedures, new methods, and implementation of the novel intervention.

Conducting a pilot study prior to the main study can enhance the likelihood of success of the larger scale study and potentially help to avoid doomed main studies. Pilot studies should be well designed with clear feasibility objectives, clear analytic plans, and explicit criteria for determining success of feasibility. They should be used cautiously for determining treatment effects and variance estimates for power or sample size calculations. Finally, they should be scrutinized the same way as full-scale studies.

References

1. Le Tourneau C, Lee JJ, Siu LL. Dose escalation methods in phase I cancer clinical trials. J Natl Cancer Inst. 2009;101(10):708-720. 10.1093/jnci/djp079.
2. Leon, Andrew C et al. "The role and interpretation of pilot studies in clinical research." *Journal of psychiatric research* vol. 45,5 (2011): 626-9. doi:10.1016/j.jpsychires.2010.10.008.
3. Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, Robson R, Thabane M, Giangregorio L, Goldsmith CH. A tutorial on pilot studies: the what, why and how. *BMC Medical Research Methodology***volume 10**, Article number: 1 (2010). doi:10.1186/1471-2288-10-1
4. Todd S, Whitehead A, Stallard N, Whitehead J. Interim analyses and sequential designs in phase III studies. Br J Clin Pharmacol. 2001 May; 51(5): 394–399.doi: 10.1046/j.1365-2125.2001.01382.x

OBSERVATIONAL STUDIES

1. Cohort Studies

In cohort studies, participants that have a particular common exposure (the cohort) are identified and outcomes are observed over time. In these studies, information about the risk factor or exposure is determined prior to the observation of the outcomes. Cohort studies can either be prospective or retrospective.

1.1. Prospective Cohort Studies

In prospective cohort studies, the risk factor or exposure and subsequent outcomes are observed after the beginning of the study. This is also known as a longitudinal study. The selection of participants is influenced by a variety of factors, including the type of exposure being investigated, the frequency of the exposure in the population, and the accessibility of participants, as well as the likelihood of their continuing participation. Unexposed participants should be sampled from the same (or comparable) source population as the exposed group. Both exposed and unexposed groups should not have the outcome being investigated and be equally susceptible to development of the outcome at the beginning of the study. The baseline characteristics of the exposed group should not differ systematically from those in the unexposed group for the exposure of interest. Equivalent information should be available on exposure and outcomes in both groups. Both groups should be accessible and available for follow-up.

1.2. Retrospective Cohort Studies

Retrospective cohort studies, sometimes called “chart reviews”, are carried out at the present time and look to the past to examine medical events or outcomes. Specifically, a cohort of participants selected based on exposure status is chosen at the present time, and outcome data (e.g. disease status, event status), which was measured in the past, are reconstructed for analysis. The primary disadvantage of this study design is the limited control the investigator has over data collection. The existing data may be incomplete, inaccurate, or inconsistently measured between participants. However, because of the immediate availability of the data, this study design is comparatively less costly and shorter than prospective studies and can provide valuable results to address important clinical research questions.

Comparing effectiveness of interventions in retrospective studies is difficult because usually there are baseline differences between interventions. In randomized controlled trials (RCTs), treatment influences on outcomes are usually considered as causal because the participants taking different treatments are supposed to be exchangeable (i.e., their characteristics, except the intervention that is evaluated, are expected to be the same). However, in retrospective studies the assumption of exchangeability is not valid because participants are prescribed different medications precisely because they differ in prognostic factors. Hence, applying sound statistical methods to reduce confounding – a systematic error in a study that results from confusing the effect of the exposure of interest with other associated correlates of the outcome – is needed when analyzing retrospective studies. Propensity scores are a suitable methodology for adjusting for such differences and, therefore, for obtaining unbiased effectiveness estimates. The goal of propensity scores is to balance observed covariates between participants from the treatment groups in order to mimic what happens in an RCT.

In a balanced two-arm randomized trial, the propensity score of each subject is equal to 1/2 for every covariate (i.e., subjects with different observed covariates have the same probability of receiving treatment, and reversibly each possible value of the observed covariates is as likely to occur in either of the two groups.) Typically, in retrospective studies there are participants that are more likely to receive an aggressive

treatment because of some of the pre-treatment characteristics included in the observed covariates. Analogously, other participants are more likely to receive a less aggressive treatment given their covariates. However, suppose that we compare two participants who have the same propensity score. These participants could be different in terms of their observed covariates. What is important is that these differences cannot predict which participant has more chance of receiving the aggressive treatment. Given their observed covariates, both have the same probability to be treated despite being quite different in terms of their covariates. Hence, if participants with the same propensity scores are grouped, both aggressively treated and less aggressively treated participants in these groups will have on average covariate patterns similar to those that would occur in a randomized trial.

2. Case-Control Studies

In case-control designs, participants are identified by whether or not they have the outcome of interest. Then a comparison of the groups with respect to exposures or some other attribute is made. These studies begin with case and control participants (i.e., the outcome of interest is known) and look back retrospectively at the participants' exposures to find an association. One of the first steps in this design is to identify and select cases. Case identification should be very specific and the source population should be well defined. The criteria for a case should minimize the likelihood that true cases are missed, while simultaneously avoiding falsely classifying a nonaffected participant as a case. The next key step is to identify and select controls. Ideally, controls are chosen at random from the source population. The selected control group must be at similar risk of developing the outcome.

In addition to confounding being an issue in observational studies, selection bias is also a danger to the internal validity of observational studies; and this bias poses a particular threat to case-control studies. Selection bias occurs when there is a different probability of an participant being chosen to participate in a study or assigned to an intervention, and the characteristics of that participant are confounded with outcomes. Removing biases are central methodological issues in observational studies, therefore, applying sound statistical techniques to address these issues is essential. Ignoring these biases often results in incorrect estimates of the association or effect of the intervention. Of note, selection bias and confounding are not affected by sample size. While large sample sizes provide real advantages in the accurate and powerful detection of associations, their ability to identify causality is not as strong. Indeed, with a very large sample size, a small effect estimate can yield a very low p-value, making many claim cause and effect. Study design can be much more important than p-values in this context. No amount of elaborate statistical analysis can help an experiment that was conducted without attention to key issues such as study design, potential sources of variation, and confounding. Thus, study design and statistical analysis should go hand in hand.

3. Cross-Sectional Studies

In cross-sectional studies, the exposure and outcome information is assessed simultaneously at a single point in time. Unlike in cohort studies (participants selected based on exposure) or case-control studies (participants selected based on outcome), the participants in a cross-sectional study are merely selected based on the inclusion and exclusion criteria set for the study. These designs are often used for population-based surveys and to assess the prevalence of diseases in clinic-based samples. Since cross-sectional studies are a one-time measurement of exposure and outcome, it is difficult to derive causal relationships from the analysis. However, the investigator can study the association between these measures. Cross-sectional studies can typically be conducted relatively fast and are inexpensive.

They may be useful for public health planning, monitoring, and evaluation. For example, the National AIDS Programme may conduct cross-sectional sentinel surveys among high-risk groups and ante-natal mothers every year to monitor the prevalence of HIV in these groups.

References

1. Faries D, Leon AC, Haro JM, and Obenchain RL. 2010. Analysis of Observational Health Care Data Using SAS®.
2. Setia MS. Methodology Series Module 3: Cross-sectional Studies. *Indian J Dermatol*. 2016;61(3):261-264. doi:10.4103/0019-5154.182410.

DATA QUALITY

Investigators seeking statistical analyses for their research should meet with the statistician *prior to* any data collection.

The following are the four approved MDACC databases: REDCap (Research Electronic Data Capture), Prometheus, DMI, MOCLIP

Excel Spreadsheets

Data that require relabeling or editing may prolong the time required to complete the statistical analyses. If data are not correctly entered or coded, inaccurate or incomplete analyses may result. The following is a list of suggested data considerations that may minimize the time required to organize and prepare data for statistical analysis.

1. Each row should have a unique identifier to keep track of which data correspond with a given subject or other experimental unit. Preferably, this identifier will be recorded in the first column. This identifier will typically be a medical record number (MRN) or accession number but may be another identifier. A unique identifier is required so that data queries can be resolved and also so that data may be merged between different datasets, if necessary.

2. Each column should contain only one type of data. Dates, text, and numbers are different types of data.

   For example, if a column contains lab values, as well as entries such as “N/A”, “<0.1”, “could not be determined”, “undetectable”, or “>1000000”, the data may be more difficult to process. For example, including the “>” symbol in one cell of the column will cause a statistical analysis program to read in an entire column as text rather than as numeric values.

3. Each column should contain only one piece of data (i.e., one variable).

   For example, systolic and diastolic blood pressure should be recorded in two separate columns. Instead of entering “144/88” in one column, the systolic pressure of 144 should be recorded in one column and the diastolic pressure of 88 in another.

4. Columns of data should not include units. If units are necessary, the units should be recorded in a separate column or in the column header. For example, if some weights are measured in kilograms and some in pounds, the units of measurement should be entered into the next column. If all entries in the same column are based on the same unit, the unit of measure may be noted in the column header.

5. Coding should be consistent. Multiple spellings and variable lengths must be reconciled prior to analysis.

   For instance, “male”, “Male”, “m”, “ M” and “M” are all different to the computer software and must be coded consistently prior to analysis.

6. The use of Protected Health Information (PHI) should be kept to a minimum and on a need-to-know basis. PHI includes the following information:
   1. Subject names

2. All geographical subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code, if according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or old
4. Phone numbers
5. Fax numbers
6. Electronic mail addresses
7. Social Security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers, including license plate numbers
13. Device identifiers and serial numbers
14. Web Universal Resource Locators (URLs)
15. Internet Protocol (IP) address numbers
16. Biometric identifiers, including finger and voice prints
17. Full face photographic images and any comparable images
18. Any other unique identifying number, characteristic, or code (note this does not mean the unique code assigned by the investigator to code the data)

Per institutional policy ADM0335 Information Security Office Policy for the Use and Protection of Information Resources, PHI should **NOT** be sent outside MD Anderson (i.e., PHI should not be sent to/from non-MDACC email accounts). In general, **DO NOT SEND SUBJECT NAMES** to the statistician.

7. Records should be sorted carefully or not sorted at all.

   Excel has the capability of sorting a column independent of other columns. This means that it is extremely easy to completely scramble the data in a spreadsheet.

8. A key or “Data Dictionary” to define variables and describe possible values for variables should be provided.

   For example, a separate coding sheet should be provided detailing that the column named “N” indicates “Nodal status”, where value 0 corresponds to “negative” status and values 1-3 correspond to “positive” status.

9. Colored cells or colored text should not be used to convey information. Separate columns should be used instead.

   For example, to denote different subject groups, use a separate column with a number or letter to identify each group and include this information in the data dictionary (A=group 1, B=group 2, etc.).

10. Empty cells should not be used to convey meaningful information, such as the lack of a condition. In general, if data are missing for only one possible reason, empty cells can be used to indicate missing data. However, if it's desirable or necessary to distinguish between different reasons for a missing data value, separate codes or database fields should be used to distinguish between these types. For example,

    - Unknown values
    - Values that are not applicable
    - Not done
    - Not recorded

11. Values below the limit of detection for an assay should not be listed as missing/unknown. Instead, the statistician should be consulted as to the appropriate manner to record these data.

12. Data field or variable names should be short, meaningful and unique. Names exceeding two words should be avoided. The length of the name should be within 12 characters. Symbols such as '/', '&' '( )', '?', etc. should be avoided. Variable names may include numbers but cannot begin with a number.

13. The variable names should be presented in the first row of the spreadsheet as the header for each column. The other rows, beginning with row number 2 should contain the data values for each experimental unit (i.e., one row per subject).

14. Cells should not be merged.

15. Cells should not be hidden.

16. The header row should not be duplicated down in the spreadsheet.

17. Summary statistics should not be included in the same spreadsheet as the raw data. Summary statistics or results of preliminary statistical testing may be useful for reference, but these should not be placed in the same spreadsheet as the raw data. Such results should be located in a separate spreadsheet or document, if necessary.

18. For time intervals, start and end dates must be provided rather than a computed interval.
    For example, for survival calculations, the start date as well as the death date or last follow-up date are required, not simply a calculation of time.

Examples of a less than ideal dataset and a better version of the same dataset are displayed below:

"Bad" Example Data Set:

	A	B	C	D	E	F
1	Patient No	DOB	Sex	Race	Status	Creatinine
2	1	19-Aug-55	m	Black	Alive	None
3	2	04/23/1953	F	African American	Dead - 6/12/2008	<.5
4	3	31/10/1942	male	White	Alive	55 mmol/L
5	Patient No	DOB	Sex	Race	Status	Creatinine
6	4	5.6.70	Male	Caucasian	A	N/A
7	5	12-Nov-32	f	A	Dead1/12/2009	2
8	6	09-Aug-52	female	??	D 9/25/2007	200 mmol/L
9	GroupA					
10	GroupB		The average systolic blood pressure for Group A was 125			
11			The average systolic blood pressure for Group B was 115			
12	AVG Creatinine					
13	Group A	0.6				
14	Group B	1.9				
15	TTEST	p = 0.028				

The above example violates data considerations #3, 4, 5, 6, 10, 17 and 18.

"Better" Example Data Set:

ID	Group	DOB	Sex	Race	Status	Death Date	Creatinine	SBP
1	1	08/19/1955	M	1	0			123
2	1	04/23/1953	F	1	1	06/12/2008	0.4	125
3	1	10/31/1942	M	2	0		0.8	127
4	0	05/06/1970	M	2	0			116
5	0	11/12/1932	F	3	1	01/12/2009	2	115
6	0	08/09/1952	F		1	09/25/2007	1.8	114

Key

Group 1 = Treatment
Group 0 = Control
Status 1 = Dead
Status 0 = Alive
Creatinine units = mg/dL
Race 1 = Black
Race 2 = White
Race 3 = Arab

DOB = Date of Birth
SBP = Systolic Blood Pressure

STATISTICAL METHODOLOGY
CATEGORICAL MEASURES

1. Analysis of Categorical Measures

Categorical data analysis is concerned with the analysis of categorical measures (e.g., response), regardless of whether any accompanying explanatory variables are also categorical or are continuous. An important consideration in determining the appropriate analysis of categorical variables is their scale of measurement. The scale of measurement of a categorical variable is a key element in choosing an appropriate analysis strategy. By taking advantage of the methodologies available for the particular scale of measurement, a well-targeted strategy can be chosen. If the scale of measurement is not taken into account, an inappropriate strategy may be chosen that could lead to erroneous conclusions. Categorical variables can be a) nominal; b) dichotomous; and c) ordinal.

- Nominal variables are variables that have two or more categories, but which do not have an intrinsic order (e.g., red, blue, green, yellow). Of note, the different categories of a nominal variable can also be referred to as groups or levels of the nominal variable.
- Dichotomous variables are nominal variables which have only two categories or levels (e.g., male, female)
- Ordinal variables are variables that have two or more categories just like nominal variables only the categories can also be ordered or ranked (e.g., low, medium, high).

1.1. Dichotomous Variables

Dichotomous variables are those that have two possible outcomes (e.g., response vs. no response). The 2 x 2 contingency table (see Table 1) is one of the most common ways to summarize categorical data. Generally, interest lies in whether there is an association between the row variable (Response at Day 90) and the column variable (Treatment). The question of interest in this example is whether the PR or better response rates for Bu-Mel treatment (102/104; 98%) and Mel treatment (95/98; 97%) are the same.

Table 1. Association between Treatment and Response

	Treatment		Total
Response at Day 90	Bu-Mel	Mel	
PR or better	102	95	197
SD/PD	2	3	5
Total	104	98	202

The null hypothesis (i.e., H_0) for this illustration is: There is no association between treatment and response at Day 90. If the sample size in each cell is large enough, the statistic that is used to test the hypothesis is based on the chi-square statistic. However, if the counts in the table are too small to meet the sample size requirements necessary for the chi-square distribution to apply (rule of thumb: < 5 in any cell), exact methods are used to test the hypothesis of no association. Since the number of subjects with SD/PD is < 5 is in two cells in our example, an exact method should be employed.

Fisher's Exact Test

If we assume the margins of the 2 x 2 contingency table are fixed (i.e., 104, 98, 197, 5) then the significance level (i.e., p-value) is the probability of the observed data or more extreme data occurring under the null hypothesis. The two-sided Fisher's exact p-value in our example is 0.675.

Measures of Association

Measures of association are used to assess the strength of an association. For the 2 x 2 contingency table, one measure of association is the odds ratio (OR).

For Table 1, the OR compares the odds of the Bu-Mel subjects having PR or better response to the odds of the Mel subjects having PR or better response. It is computed as:

$$OR = \frac{102/95}{2/3} = 1.6$$

The OR ranges from 0 to infinity. When the OR is 1, there is no association between the two variables. If the OR is greater than 1, the Bu-Mel group is more likely than the Mel group to have PR or better response. If the OR is less than 1, then Bu-Mel is less likely than the Mel group to have PR or better response.

Another measure of association is relative risk, which is the risk of developing a particular condition (e.g., cancer) for one group compared with another group. In our example, the relative risk is computed as:

$$RR = \frac{102/197}{2/5} = \frac{0.518}{0.4} = 1.3$$

1.2. Logistic Regression Modeling

In general, the overall idea of regression modeling is to examine two things: (1) does a set of predictor (explanatory) variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.

Logistic regression is a form of statistical modeling that describes the relationship between a categorical variable and a set of explanatory variables. The explanatory variables can be categorical or continuous. One of the advantages of logistic regression modeling is that model interpretation is possible through odds ratios, which are functions of model parameters.

Example: Is there an association between treatment and progression/death adjusting for age (65 years; > 65 years), cytogenetic risk (high; standard), and R-ISS stage (I-II; III)?

Measure	Odds Ratio (95% CI)	p-value
Bu-Mel vs. Mel	0.77 (0.37, 1.59)	0.48
> 65 years vs. 65 years	0.97 (0.39, 2.38)	0.94
High vs. standard	0.73 (0.29, 1.82)	0.50
III vs. I-II	1.85 (0.73, 4.71)	0.20

Interpretation: In addition to the p-values being large, the odds ratios for each measure is close to 1 and the 95% confidence interval contains 1, indicating that none of the measures were significantly associated with progression and/or death.

Reference

Stokes, M.E., Davis, C.S., and Koch, G.G., *Categorical Data Analysis Using the SAS System*, Cary, NC: SAS Institute Inc., 1995.

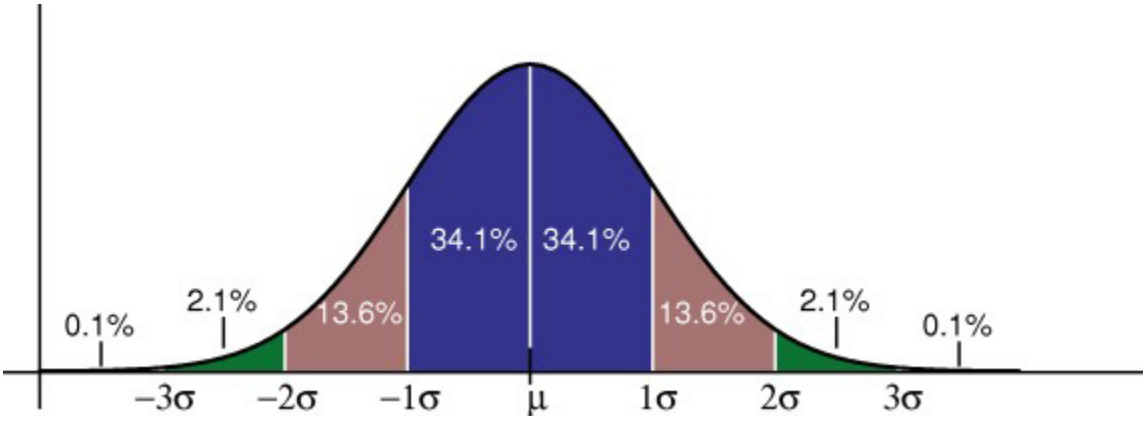
STATISTICAL METHODOLOGY
CONTINUOUS MEASURES

1. Continuous Measures

Continuous variables are also known as quantitative variables and can be categorized as either interval or ratio variables.

- Interval variables are variables for which their central characteristic is that they can be measured along a continuum and they have a numerical value (e.g., temperature). So the difference between 20°C and 30°C is the same as 30°C to 40°C.
- Ratio variables are interval variables, but with the added condition that 0 (zero) of the measurement indicates that there is none of that variable. So, temperature measured in degrees Celsius is not a ratio variable because 0°C does not mean there is no temperature. However, temperature measured in Kelvin is a ratio variable as 0° Kelvin (often called absolute zero) indicates that there is no temperature whatsoever.

It has been observed that the natural variation of many continuous variables tends to follow a bell-shaped distribution, with most values clustered symmetrically near the mean and a few values falling out on the tails. The bell-shaped distribution is also known as the normal (or Gaussian). The shape and location of a normal distribution are completely determined by its mean and standard deviation (SD). In a normal distribution, 68% of the data fall within 1 SD of the mean (34% above, 34% below); 95% within 2 SD and 99.7% within 3 SD of the mean (see figure). For non-Gaussian distributions, the SD does not describe a known proportion of the observations.



2. Analysis of Continuous Measures

Analyses of continuous measures fall into two broad classifications of statistical procedures; parametric and non-parametric. Parametric analyses have the following assumptions about the underlying data: i.) the data were derived from a population in which the characteristic to be studied is normally distributed; ii.) the variances within the groups to be studied must be homogeneous; and iii.) the data are independent. These assumptions should be confirmed or assumed with good reason when using these tests. If these assumptions are violated, the resulting statistics and conclusions will not be valid, and the tests may lack power relative to alternative tests. Non-parametric tests are sometimes called **distribution-free tests** because they are based on fewer assumptions (e.g., they do not assume that the outcome is approximately normally distributed). These tests assume that the underlying distributions have the same

shape and spread. The cost of fewer assumptions is that non-parametric tests are generally less powerful than their parametric counterparts (i.e., when the alternative is true, they may be less likely to reject H_0).

The following table presents parametric tests and their non-parametric counterparts for specific types of analyses of continuous variable. For example: Say you were interested in evaluating whether a particular diet was effective, you could test the change in a person's weight from baseline to 6 months to determine if it was significantly different from 0 (i.e., no change) using either a paired t-test or Wilcoxon signed-rank test.

Analysis Type	Parametric Test	Non-parametric Test
Compare quantitative measure between two distinct/independent groups	Two-sample t-test	Wilcoxon rank-sum test
Compare two quantitative measurements taken from the same individual	Paired t-test	Wilcoxon signed-rank test
Compare quantitative measurements between three or more distinct/independent groups	Analysis of variance (ANOVA)	Kruskal-Wallis test
Estimate the degree of association between two quantitative variables	Pearson coefficient of correlation	Spearman's rank correlation

2.1. Linear Regression

In linear regression, both the dependent and independent measures are continuous. Linear regression attempts to model the relationship between the dependent and independent measures by fitting a linear equation to observed data. A linear regression line has an equation of the form $\boldsymbol{Y = \alpha + \beta X}$, where $\boldsymbol{X}$ is the independent variable and $\boldsymbol{Y}$ is the dependent variable. The slope of the line is $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ is the intercept (the value of $\boldsymbol{y}$ when $\boldsymbol{x} = 0$).

Linear Regression Example: Is there an association between total number of radium doses and prostate specific antigen (PSA), hemoglobin, and alkaline phosphatase (ALK)?

Measure	β	SE of β	95% LCI	95% UCI	p-value
Intercept	4.972	1.296	2.403	7.542	< 0.001
PSA	-0.000	0.000	-0.001	0.001	0.56
Hemoglobin	0.002	0.106	-0.208	0.212	0.99
ALK	-0.001	0.001	-0.002	0.001	0.45

Interpretation: In addition to the p-values being large, the slope, $\boldsymbol{\beta}$, values are close to 0, indicating that there's no association between total number of radium doses and the independent factors.

ANOVA Example: Is there an association between total number of radium doses and subjects with PSA > 10 ng/mL, hemoglobin < 10 g/dL, and ALK ≥ 146 U/l?

Measure	β	SE of β	95% LCI	95% UCI	p-value
Intercept	5.187	0.286	4.620	5.754	< 0.001
PSA > 10 ng/mL	0.047	0.331	-0.610	0.704	0.89
Hemoglobin < 10 g/dL	0.001	0.423	-0.837	0.840	1.00
ALK ≥ 146 U/l	-0.973	0.317	-1.601	-0.345	0.003

Measure	LS Means	SE
PSA		
> 10 ng/mL	4.75	0.21
≤ 10 ng/mL	4.70	0.32
Hemoglobin		
< 10 g/dL	4.73	0.38
≥ 10 g/dL	4.72	0.19
ALK		
≥ 146 U/l	4.24	0.26
< 146 U/l	5.21	0.27

Interpretation: When the ALK measure is divided into two clinically meaning groups, a significant association between total number of radium doses and ALK is observed. Subjects with ALK values ≥ 146 U/l received significantly fewer radium doses (least square [LS] means: 4.24) compared with those with ALK values < 146 U/l (LS means: 5.21). It may be more clinically meaningful to interpret these results using LS means (adjusted means) as opposed to slopes.

STATISTICAL METHODOLOGY

TIME-TO-EVENT/SURVIVAL MEASURES

1. Analysis of Time-to-Event/Survival Measures

In many medical studies the outcome of interest is the length of time until an event occurs, i.e., the time elapsed from one well-defined event, for example, start of treatment, to another well-defined event, for example, death. For convenience, we will refer to this time as "survival time" and to the outcome as "death", even though the outcome may be some other, possibly favorable, event such as disease remission. The distribution of survival times is most often described in terms of the survival function, S(t). This function is defined for each time t as the probability that an individual survives longer than time t. The graph of S(t) against time is called the survival curve. The distribution of survival times is also described in terms of the hazard function, h(t). This function is defined as the probability that an individual dies during a short interval of time given that the individual survived up to that interval.

Although survival time is a continuous variable, one cannot, in general, use standard analysis techniques for continuous measures with survival data because of the presence of censored observations. Censored observations arise in subjects for whom the critical event has not yet occurred at the time of the analysis. Censored observations can arise if a subject is known to be alive when the analysis is carried out or is lost to follow-up during the course of a study. The time to the last date the live subject was examined is known as the censored survival time. Thus, the relevant response data for a survival analysis consists of two components: (1) the subject's status at the last follow-up observation (e.g., dead or alive) and (2) the length of time the subject was followed. Special techniques have been developed to deal with censored survival data which take in account the information provided by censored observations.

1.1. Survival Function Estimation

The two most common methods for estimating the survival function in the presence of censored data are the Kaplan-Meier product-limit method and the Cutler-Ederer (actuarial) life-table method. These are both non-parametric methods which do not require specification of the functional form of the survival time distribution (which is often unknown).

In the life-table method, survival times are grouped into convenient intervals. The probability of dying during an interval is computed for each interval and the survival function is taken as the product of the survival probabilities for succeeding intervals. This method assumes that censored observations are uniformly distributed within each interval and that the risk of death is fairly constant within each interval. The product-limit can be considered as a special case of the life-table estimate where each interval contains only one observation. In the Kaplan-Meier approach, the intervals are determined by the data and thus the results are not dependent on the user's choice of time intervals.

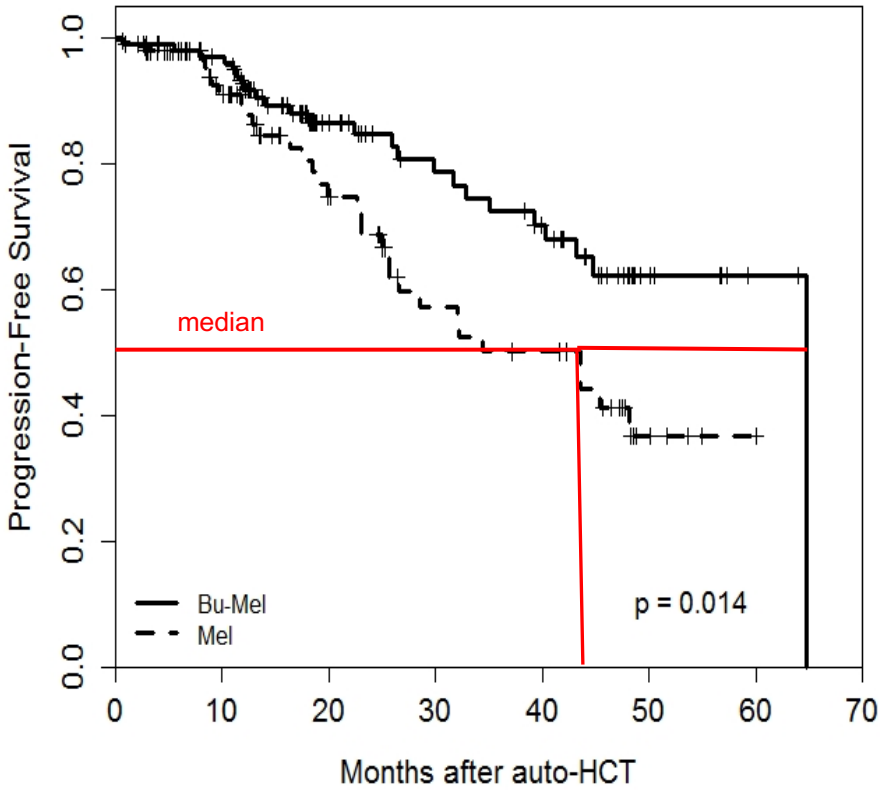
1.2. Comparing Survival Distributions

If subjects are divided into groups according to treatment or prognosis, then the survival function can be estimated for each group, and a test of equality of the survival functions across groups may be performed. Because survival times are positively valued and often highly skewed, nonparametric tests are most often used to make comparisons. In the absence of censoring, standard nonparametric tests could be used to compare survival distributions, however, for censored data, special tests are needed. The tests most often encountered are nonparametric linear rank tests: typically, the Mantel-Cox log-rank test, and the Breslow-Gehan generalized Kruskal-Wallis test. These tests accumulate weighted differences over time between what is observed and what would have been expected under the null hypothesis of equivalence. Different weights lead to different test statistics. Each test is sensitive to a characteristic pattern of difference between

survival distributions being compared. Hence one should decide beforehand what pattern of difference is most important clinically and select a significance test that has good statistical power for detecting differences of that type. The Mantel-Cox test gives equal weight to all observations. The Breslow-Gehan test gives greater weight to early observations, thus it is less sensitive than the Mantel-Cox test to late events when few subjects remain in the study.

When interpreting the results of the statistical tests for comparing survival time distributions, consideration should be given to the sample size of subjects used, and the pattern and amount of censoring. The number of subjects is important, since the distribution of the test statistics and reported p-values are based on asymptotic (large sample) statistical theory. Thus, ideally the test statistics are calculated using a large sample of subjects. When only a small sample of subjects is used, the test results should be interpreted with care.

Example: Is there a difference in progression-free survival between Bu-Mel and Mel subjects?



Stratum 1: Bu-Mel

Product-Limit Survival Estimates						
pfsmnths		**Survival**	**Failure**	**Survival Standard Error**	**Number Failed**	**Number Left**
0.0000		1.0000	0	0	0	104
0.6571		0.9904	0.00962	0.00957	1	103
0.9528	*	.	.	.	1	102
2.1355	*	.	.	.	1	101
2.9569	*	.	.	.	1	100
2.9897	*	.	.	.	1	99
4.0411	*	.	.	.	1	98
4.1068	*	.	.	.	1	97
5.5524		0.9802	0.0198	0.0139	2	96
7.8522		0.9700	0.0300	0.0171	3	95
9.0349	*	.	.	.	3	94
10.2834		0.9596	0.0404	0.0198	4	93
11.0390		0.9493	0.0507	0.0221	5	92
11.0390	*	.	.	.	5	91
11.1704	*	.	.	.	5	90
11.4990		0.9388	0.0612	0.0242	6	89
11.7618		0.9282	0.0718	0.0262	7	88
11.7618	*	.	.	.	7	87
11.9261	*	.	.	.	7	86
12.0246	*	.	.	.	7	85
12.0246	*	.	.	.	7	84
12.1889		0.9172	0.0828	0.0281	8	83
12.2218	*	.	.	.	8	82
12.4517	*	.	.	.	8	81
12.5503	*	.	.	.	8	80
12.6817	*	.	.	.	8	79
13.0431	*	.	.	.	8	78

Product-Limit Survival Estimates						
pfsmnths		**Survival**	**Failure**	**Survival Standard Error**	**Number Failed**	**Number Left**
13.1088		0.9054	0.0946	0.0301	9	77
13.4374	*	.	.	.	9	76
13.7988	*	.	.	.	9	75
13.7988	*	.	.	.	9	74
13.9959		0.8932	0.1068	0.0321	10	73
14.2916	*	.	.	.	10	72
14.3573	*	.	.	.	10	71
15.6057	*	.	.	.	10	70
15.8029	*	.	.	.	10	69
16.1314	*	.	.	.	10	68
16.4600		0.8801	0.1199	0.0342	11	67
16.6242	*	.	.	.	11	66
17.3470	*	.	.	.	11	65
17.4456	*	.	.	.	11	64
17.5441	*	.	.	.	11	63
17.8398	*	.	.	.	11	62
17.9713	*	.	.	.	11	61
18.0370		0.8656	0.1344	0.0365	12	60
18.0370	*	.	.	.	12	59
18.1684	*	.	.	.	12	58
18.3984	*	.	.	.	12	57
18.4641	*	.	.	.	12	56
18.5955	*	.	.	.	12	55
18.7269	*	.	.	.	12	54
18.9240	*	.	.	.	12	53
19.3511	*	.	.	.	12	52
20.0411	*	.	.	.	12	51
21.1253	*	.	.	.	12	50
21.2238	*	.	.	.	12	49

Product-Limit Survival Estimates						
pfsmnths		**Survival**	**Failure**	**Survival Standard Error**	**Number Failed**	**Number Left**
21.8809	*	.	.	.	12	48
22.4066		0.8476	0.1524	0.0400	13	47
22.8008	*	.	.	.	13	46
23.0308	*	.	.	.	13	45
23.4908	*	.	.	.	13	44
23.4908	*	.	.	.	13	43
24.1150	*	.	.	.	13	42
25.9548		0.8274	0.1726	0.0438	14	41
26.4805		0.8072	0.1928	0.0472	15	40
26.6776	*	.	.	.	15	39
29.8973		0.7865	0.2135	0.0503	16	38
31.6386		0.7658	0.2342	0.0531	17	37
32.9199		0.7451	0.2549	0.0555	18	36
35.1211		0.7244	0.2756	0.0577	19	35
38.3737	*	.	.	.	19	34
39.2936		0.7031	0.2969	0.0598	20	33
39.3265	*	.	.	.	20	32
39.9179	*	.	.	.	20	31
40.3450		0.6804	0.3196	0.0620	21	30
41.0678	*	.	.	.	21	29
41.7906	*	.	.	.	21	28
41.8563	*	.	.	.	21	27
41.9877	*	.	.	.	21	26
42.0534	*	.	.	.	21	25
43.1704		0.6532	0.3468	0.0653	22	24
43.9589	*	.	.	.	22	23
44.1232	*	.	.	.	22	22
44.7146		0.6235	0.3765	0.0687	23	21
45.3060	*	.	.	.	23	20

Product-Limit Survival Estimates						
pfsmnths		Survival	Failure	Survival Standard Error	Number Failed	Number Left
45.5359	*	.	.	.	23	19
45.6016	*	.	.	.	23	18
46.0287	*	.	.	.	23	17
47.1129	*	.	.	.	23	16
47.1129	*	.	.	.	23	15
47.6715	*	.	.	.	23	14
48.0000	*	.	.	.	23	13
48.1643	*	.	.	.	23	12
48.3943	*	.	.	.	23	11
48.5914	*	.	.	.	23	10
48.7228	*	.	.	.	23	9
49.2156	*	.	.	.	23	8
50.1027	*	.	.	.	23	7
50.5298	*	.	.	.	23	6
56.6407	*	.	.	.	23	5
56.8378	*	.	.	.	23	4
57.2977	*	.	.	.	23	3
59.2690	*	.	.	.	23	2
63.9671	*	.	.	.	23	1
64.6899		0	1.0000	.	24	0

Note: The marked survival times are censored observations.

Quartile Estimates				
Percent	Point Estimate	95% Confidence Interval		
		Transform	[Lower	Upper)
75	64.6899	LOGLOG	.	.
50	64.6899	LOGLOG	44.7146	64.6899
25	32.9199	LOGLOG	22.4066	44.7146

Stratum 2: Mel

Quartile Estimates				
Percent	Point Estimate	95% Confidence Interval		
		Transform	[Lower	Upper)
75	.	LOGLOG	48.1971	.
50	43.5318	LOGLOG	25.7248	.
25	19.8768	LOGLOG	13.4374	25.7248

Interpretation: Bu-Mel subjects experience significantly longer PFS median [95% CI] compared with their Mel counterparts (64.7 [44.7, 64.7] vs. 43.5 [25.7, not estimated]; p=0.014).

Note: The difference between the median follow-up time and median OS results from subjects who were alive at their last follow-up visit and how censoring is handled in computing Kaplan-Meier estimates. For example, the median follow-up time for the Bu-Mel group was 22.6 months and for the Mel group was 20.2 months. If all subjects had died, then the median OS times would be the same as the median follow-up times.

1.3. Cox Proportional Hazards Regression Modeling

Survival analysis methods can also be extended to assess several risk factors simultaneously, similar to multiple linear and multiple logistic regression analysis. One of the most popular regression techniques for survival analysis is Cox proportional hazards regression, which is used to relate one or several independent factors, considered simultaneously, to survival time. In a Cox proportional hazards regression model, the measure of effect is the **hazard rate**, which is the risk of failure (i.e., the risk or probability of suffering the event of interest), given that the participant has survived up to a specific time. There are several important assumptions for appropriate use of the Cox proportional hazards regression model, including: i.) independence of survival times between distinct individuals in the sample; ii.) a multiplicative relationship between the independent factors and the hazard (as opposed to a linear one as was the case with multiple linear regression analysis); and iii.) a constant hazard ratio over time (i.e., proportional hazards). There are many advantages to the Cox model, one is its ability to include time-dependent covariates, specifically those factors that can change after the start time (e.g., experiencing acute GVHD after transplantation).

Example: Is there an association between treatment group and progression-free survival adjusting for age, ethnicity, cytogenetic risk, ISS, response to induction therapy, and randomization algorithm?

Measure	Hazard Ratio (95% CI)	p-value
Bu-Mel vs. Mel	0.57 (0.30, 1.10)	0.09
> 65 vs. ≤ 65 years of age	1.06 (0.48, 2.32)	0.89
High vs. Standard Cytogenetic Risk	1.29 (0.60, 2.78)	0.52
R-ISS Stage III vs. other	1.24 (0.58, 2.64)	0.58
PR or better to induction therapy vs. worse than PR	0.58 (0.18, 1.91)	0.37
2^{nd} vs. 1^{st} Randomization Algorithm	2.33 (0.45, 12.19)	0.32

Interpretation: Adjusting for covariates, Bu-Mel subjects experienced a decreased risk of progression/death compared with Mel subjects, however, this difference was not statistically significant at the 5% α-level.

1.4. Competing Risks Analysis

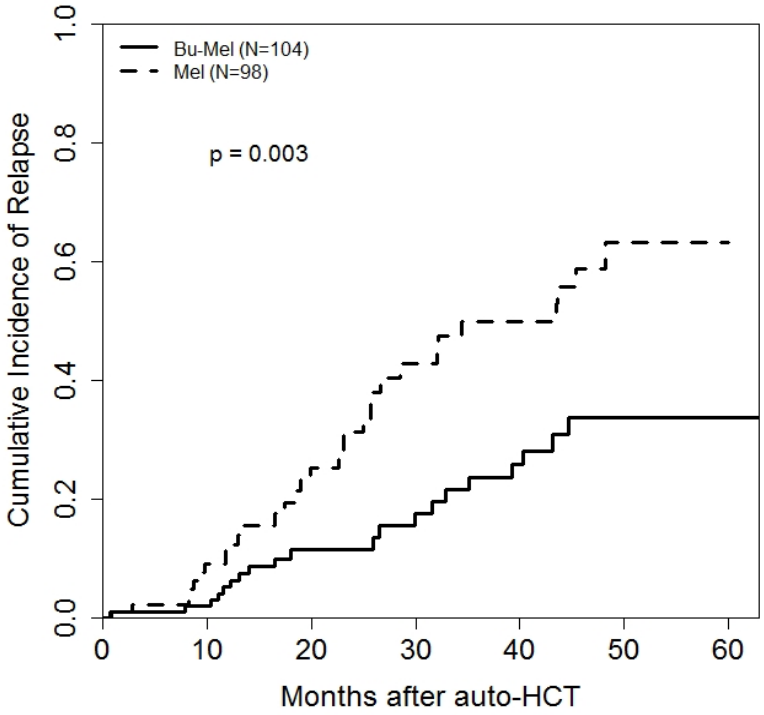
Competing risks occur frequently in the analysis of survival data. A competing risk is an event whose occurrence precludes the occurrence of the primary event of interest. For instance, in a study in which the primary outcome is time to disease progression, death without disease progression would serve as a competing event. A subject who dies is no longer at risk of progression. Regardless of how long the duration of follow-up is extended, a subject will not be observed to progress once he or she has died. Conventional statistical methods for the analysis of survival data assume that competing risks are absent. Estimating the incidence of an event as a function of follow-up time provides important information on the absolute risk of an event. In the absence of competing risks, the Kaplan-Meier estimate of the survival function is frequently used for estimating the incidence function. One minus the Kaplan-Meier estimate of the survival function provides an estimate of the cumulative incidence of events over time. However, using the Kaplan-Meier estimate of the survival function to estimate the incidence function in the presence of competing risks generally results in upward biases in the estimation of the incidence function. The problem here is that the Kaplan-Meier estimator estimates the probability of the event of interest in the absence of competing risks, which is generally larger than that in the presence of competing risks.

The Cumulative Incidence Function (CIF), as distinct from 1 – $S(t)$, allows for estimation of the incidence of the occurrence of an event while taking competing risks into account. This allows one to estimate incidence in a population where all competing events must be accounted for in clinical decision making. The cumulative incidence function for the kth cause is defined as: $CIF_k(t) = Pr(T \leq t, D = k)$, where D is a variable denoting the type of event that occurred. A key point is that, in the competing risks setting, only 1 event type can occur, such that the occurrence of 1 event precludes the subsequent occurrence of other event types. The function $CIF_k(t)$ denotes the probability of experiencing the kth event before time t and before the occurrence of a different type of event.

The CIF has the desirable property that the sum of the CIF estimates of the incidence of each of the individual outcomes will equal the CIF estimates of the incidence of the composite outcome consisting of all of the competing events. Unlike the survival function in the absence of competing risks, $CIF_k(t)$ will not

necessarily approach unity as time becomes large, because of the occurrence of competing events that preclude the occurrence of events of type *k*.

Example: Is there a difference in cumulative incidence of relapse, where death is a competing risk, between Bu-Mel and Mel subjects?



Interpretation: The cumulative incidence relapse rate was significantly lower for Bu-Mel subjects compared with Mel subjects (34% vs. 63%; p=0.003).

References

1. Allison, Paul D., *Survival Analysis Using the SAS® System: A Practical Guide*, Cary, NC: SAS Institute Inc., 1995, 292 pp.
2. Austin, PC, Lee, DS, and Fine, JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. Circulation. 2016 Feb 9; 133(6): 601-609.

STATISTICAL METHODOLOGY
BAYESIAN METHODS

1. Bayesian Methods

The field of statistics exists because it is usually impossible to collect data from all individuals of interest (population). Our only solution is to collect data from a subset (sample) of the individuals of interest, but our real desire is to know the "truth" about the population. Quantities such as means, standard deviations and proportions are all important values and are called "parameters" when we are talking about a population. Since we usually cannot get data from the whole population, we cannot know the values of the parameters for that population. We can, however, calculate estimates of these quantities for our sample. When they are calculated from sample data, these quantities are called "statistics." A statistic estimates a parameter.

There are two schools of thought when it comes to statistical methodology; i.) frequentist view and ii.) Bayesian view. The analysis methods described thus far represent the frequentist approach. When carrying out statistical inference, that is, inferring statistical information from probabilistic systems, the two approaches have very different philosophies.

Frequentist statistics assumes that probabilities are the *frequency* of particular random events occurring in a *long run* of *repeated trials*. For example, as we roll a *fair* (i.e. unweighted) six-sided die repeatedly, we would see that each number on the die tends to come up 1/6 of the time. Whereas, Bayesian inference interprets *probability* as a measure of *believability* or *confidence* that an *individual* may possess about the occurrence of a particular event. For example, we may have a *prior* belief about an event, but our beliefs are likely to change when new evidence is brought to light. Bayesian statistics gives us a solid mathematical means of incorporating our prior beliefs, and evidence, to produce new *posterior* beliefs. Frequentist statistics tries to *eliminate* uncertainty by providing *estimates*. Bayesian statistics tries to *preserve* and *refine* uncertainty by adjusting *individual* beliefs in light of new evidence.

The following table summarizes the differences between the frequentist and Bayesian approaches:

Frequentist	Bayesian
Parameters are fixed, but unknown	Parameters are unknown; therefore they have a subjective probability distribution
Data are random, until collected	Data are fixed once they are observed
After data are collected, the only thing that is random is potential future data based on repeated sampling	Parameters are random
Inferences are made conditional on future, unobserved data	Inference are made conditional on the current data and come from the posterior

Bayes Theorem:

$$Pr(A \mid B) = \frac{Pr(B \mid A)\, Pr(A)}{Pr(B)}$$

Bayes Theorem: Fun Example

Scenario: You go to a friend's party. 30% of his friends are Statisticians. You know that 70% of Statisticians are Geeks and 10% of Non-Statisticians are Geeks. You meet a person who is *clearly* a Geek. What is the probability s/he is a Statistician?

$$Pr(\text{Statistician}|\text{Geek}) = \frac{Pr(\text{Geek}|\text{Statistician}) \times Pr(\text{Statistician})}{Pr(\text{Geek})}$$

$$Pr(S|G) = \frac{Pr(G|S) \times Pr(S)}{Pr(G|S) \times Pr(S) + Pr(G|\neg S) \times Pr(\neg S)}$$

$$= \frac{.70 \times .30}{.70 \times .30 + .10 \times .70}$$

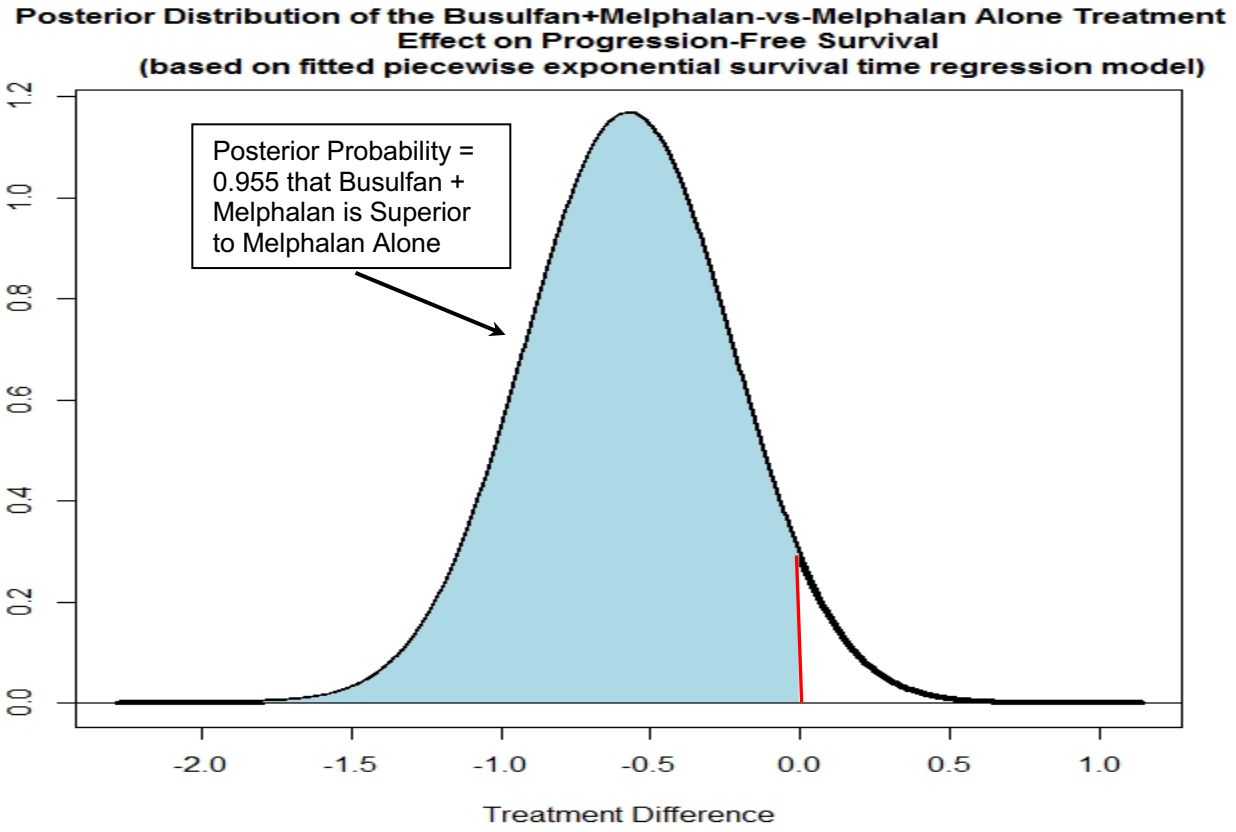
$$= .75$$

1.1. Bayesian Survival Analysis Example

Is there an association between treatment group and progression-free survival adjusting for age, ethnicity, cytogenetic risk, ISS, response to induction therapy, and randomization algorithm?

Fitted Bayesian piecewise exponential survival time regression model for progression-free survival time (N=157, number of events=39)

	Posterior Quantities				
Measure Comparison	**Mean of β**	**SD of β**	**Posterior 95% Credible Interval**		**Pr(β > 0 | Data)**
Bu-Mel vs. Mel	-0.572	0.342	-1.258	0.077	0.045
> 65 vs. ≤ 65 years of age	-0.007	0.418	-0.826	0.795	0.513
High vs. Standard Cytogenetic Risk	0.200	0.395	-0.555	0.983	0.697
R-ISS Stage III vs. other	0.235	0.389	-0.563	0.957	0.730
PR or better to induction therapy vs. worse than PR	-0.375	0.639	-1.552	0.893	0.256
2nd vs. 1st Randomization Algorithm	-0.296	0.862	-2.055	1.289	0.397
Bu-Mel vs. Mel Mean HR (95% HPD Interval) = 0.60 (0.25, 1.01).					



Interpretation: The posterior probability that Bu-Mel is superior to Mel was 0.95 for PFS, adjusting for age, ethnicity, cytogenetic risk, ISS, response to induction therapy, and randomization algorithm. The mean hazard ratio (95% credible interval) was 0.60 (0.25, 1.01). [Reminder: The HR (95% CI) for Bu-Mel vs. Mel from Cox model was 0.57 (0.30, 1.10).]

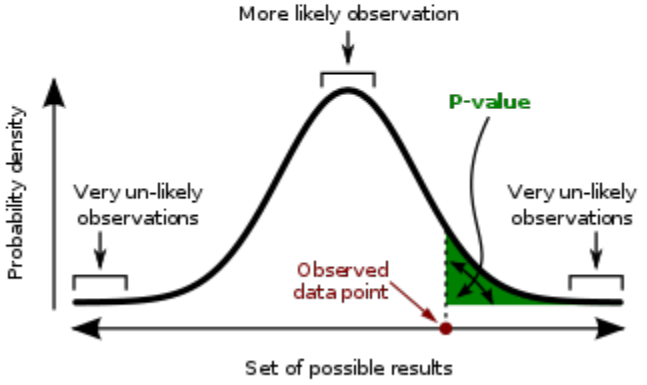
Reference

Bayesian Statistics: A Beginner's Guide. QuantStart.Tutorial.

MISCELLANEOUS

1. What exactly is a p-value?

The p-value, or calculated probability, is the probability of finding the observed, or more extreme, results when the **null hypothesis (*H_0*)** of a study question is true – the definition of 'extreme' depends on how the hypothesis is being tested. The null hypothesis is usually a hypothesis of "no difference" e.g. no difference between two treatment groups. The null hypothesis for each study question should be clearly defined before the start of your study. The **alternative hypothesis (*H_1*)** is the opposite of the null hypothesis; this is the hypothesis you set out to investigate. For example, question is "is there a significant (not due to chance) difference in response between Bu-Mel and Mel alone?" and alternative hypothesis is "there is a difference in response between Bu-Mel and Mel alone."



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

The p-value computed can be from either a one-tailed statistical test (one-sided p-value) or a two-tailed statistical test (two-sided p-value). If you are using a significance level of 0.05, a one-tailed test allots all of the alpha to testing the statistical significance in the one direction of interest. This means that 0.05 is in one tail of the distribution of your test statistic (see figure). When using a one-tailed test, you are testing for the possibility of the relationship in one direction and completely disregarding the possibility of a relationship in the other direction. Using the same significance level of 0.05, a two-tailed test allots half of your alpha to testing the statistical significance in one direction and half of your alpha to testing statistical significance in the other direction. This means that 0.025 is in each tail of the distribution of your test statistic. When using a two-tailed test, regardless of the direction of the relationship you hypothesize, you are testing for the possibility of the relationship in both directions. The only situation in which you should use a one-sided p-value is when a large change in an unexpected direction would have absolutely no relevance to your study. This situation is unusual; if you are in any doubt then use a two-sided p-value.

The significance level (alpha [α]) is the probability of incorrectly rejecting the null hypothesis (type I error or "false positive") (see table below). The significance level (α) is used to refer to a pre-chosen probability and the term "p-value" is used to indicate a probability that is computed after a given study. If the pre-specified p-value is less than the chosen significance level (α), then you reject the null hypothesis (i.e. accept that your sample gives reasonable evidence to support the alternative hypothesis). It does **NOT** imply a "meaningful" or "important" difference; that is for you to decide when considering the real-world relevance of your result. The choice of significance level at which you reject H_0 is arbitrary.

Conventionally, significance levels (α) of 5% (less than 1 in 20 chance of being wrong), 10% and 1% have been used.

	DECISION	
TRUTH	Do Not Reject H_0:	Reject H_0:
H_0 is true:	**correct decision P** *1- α*	**"false positive"** *α (significance)*
H_0 is false:	**"false negative"** *β*	**correct decision P** *1- β (power)*
P = probability		

The American Statistical Association's statement on statistical significance and p-values:

1. *P*-values can indicate how incompatible the data are with a specified statistical model.
2. *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning.

1.1. Multiple Comparisons

A scientific conclusion is always drawn from the statistical testing of hypothesis, in which the chosen significance level (α), is used for decision-making. However, the probability of committing false statistical inferences is considerably increased when more than one hypothesis is simultaneously tested (namely the multiple comparisons), which therefore requires proper adjustment. In statistical inference, a p-value is directly or indirectly computed for each hypothesis and then compared with the pre-specified significance level (α) for determining if the H_0 should be rejected or not. Therefore, there are two ways for adjusting the statistical inference of multiple comparisons. First, it could directly adjust the observed p-value for each hypothesis and keep the pre-specified significance level (α) unchanging; and this is referred to as the adjusted p-value (e.g., analysis). Second, an adjusted cut-off corresponding to the

initially pre-specified (α) could be computationally determined and then compared with the observed p-value for statistical inference (e.g., sample size/power computation).

In the ideal world, we would be able to define a "perfectly" random sample, the most appropriate test and one definitive conclusion. We simply cannot. What we can do is try to optimize all stages of our research to minimize sources of uncertainty.

2. Confidence Intervals

The p-value, which is the final common pathway for nearly all statistical tests, conveys no information about the extent to which two groups differ or two variables are associated. P-values, therefore, are not good measures of the strength of the relation between study variables. By choosing a measure that quantifies the degree of association or effect in the data and then calculating a confidence interval, researchers can summarize the strength of association in their data and allow for random variation in a simple and unambiguous way.

The statement that "the difference between treatments is not statistically significant ($p>0.05$)" amounts to a statement that the trial results are consistent with there being no difference between treatments and is not at all the same as saying that there is actually no difference. Confidence limits can advance our understanding; the width of the interval is a guide to how precisely or sensitively a parameter of interest can be estimated.

In statistical terms, the confidence interval means that if a series of identical studies were carried out repeatedly on different samples from the same populations and a 95% confidence interval was calculated in each study, then, in the long run, 95% of these confidence intervals would include the population value, thus, (in simpler and less exact terms) "there is a 95% chance that the indicated range includes the true population value"].

The general form for a confidence interval is:

$$\text{estimate} \pm (\text{factor related to confidence level}) \times (\text{standard error of the estimate}).$$

A single study usually gives an imprecise sample estimate of the overall population value of interest. This imprecision is indicated by the width of the confidence interval: the wider the interval the less precision. The width depends essentially on three factors. Firstly, the sample size: large sample sizes will give more precise results with narrower confidence intervals. In particular, wide confidence intervals emphasize the unreliability of conclusions based on small samples. Secondly, the variability of the characteristic being studied: the less variable it is, the more precise the sample estimate and the narrower the confidence interval. Thirdly, the degree of confidence required: the more confidence the wider the interval.

It should clearly be understood that a difference which is statistically significant may or may not be clinically relevant and a difference which is statistically non-significant does not necessarily imply a clinically unimportant finding. Calculating confidence intervals is an effective tool in providing information which can readily be interpreted clinically since it quantifies the magnitude of the differences directly on the scale in which the data were measured or determined.

3. Univariate vs. Multivariate vs. Multivariable

Univariate analysis refers to statistical methods that explore the relationship between one dependent variable and one independent variable:

Example Model: Response = Treatment
(dependent variable) (independent variable)

Multivariable analysis (or multiple regression) refers to statistical methods that explore the relationship between one dependent variable and more than one independent variable:

Example Model: Response = Treatment + Gender + Age + Disease Status
(dependent variable) (independent variables)

Multivariate analysis refers to a broad category of statistical methods used when more than one dependent variable at a time is analyzed for a subject. In multivariate analysis, models that predict a vector of responses for each observation can be fit simultaneously.

Example Model: Response, Toxicity = Treatment + Gender + Age + Disease Status
(dependent variables) (independent variables)

References

1. Chen SY, Feng Z, and Yi X (2017) *A general introduction to adjustment for multiple comparisons.* Journal of Thoracic Disease 9(6): 1725-1729.
2. Rothman KJ. (1978) *A Show of Confidence*. The New England Journal of Medicine 299:1362-1363.
3. Gore SM. (1981) *Assessing Methods - Confidence Intervals*. British Medical Journal 283:660-662.
4. Gardner MJ, Altman DG. (1986) *Confidence Intervals Rather than P Values: Estimation Rather than Hypothesis Testing*. British Medical Journal 292:746-750.
5. Gardner MJ, Altman DG. (1988) *Estimating with Confidence*. British Medical Journal 296:1210-1211.
6. Luus HG, Muller FO, Meyer BH. (1989) *Statistical Significance Versus Clinical Relevance - Part II. The Use and Interpretation of Confidence Intervals*. South African Medical Journal 76:626-629.