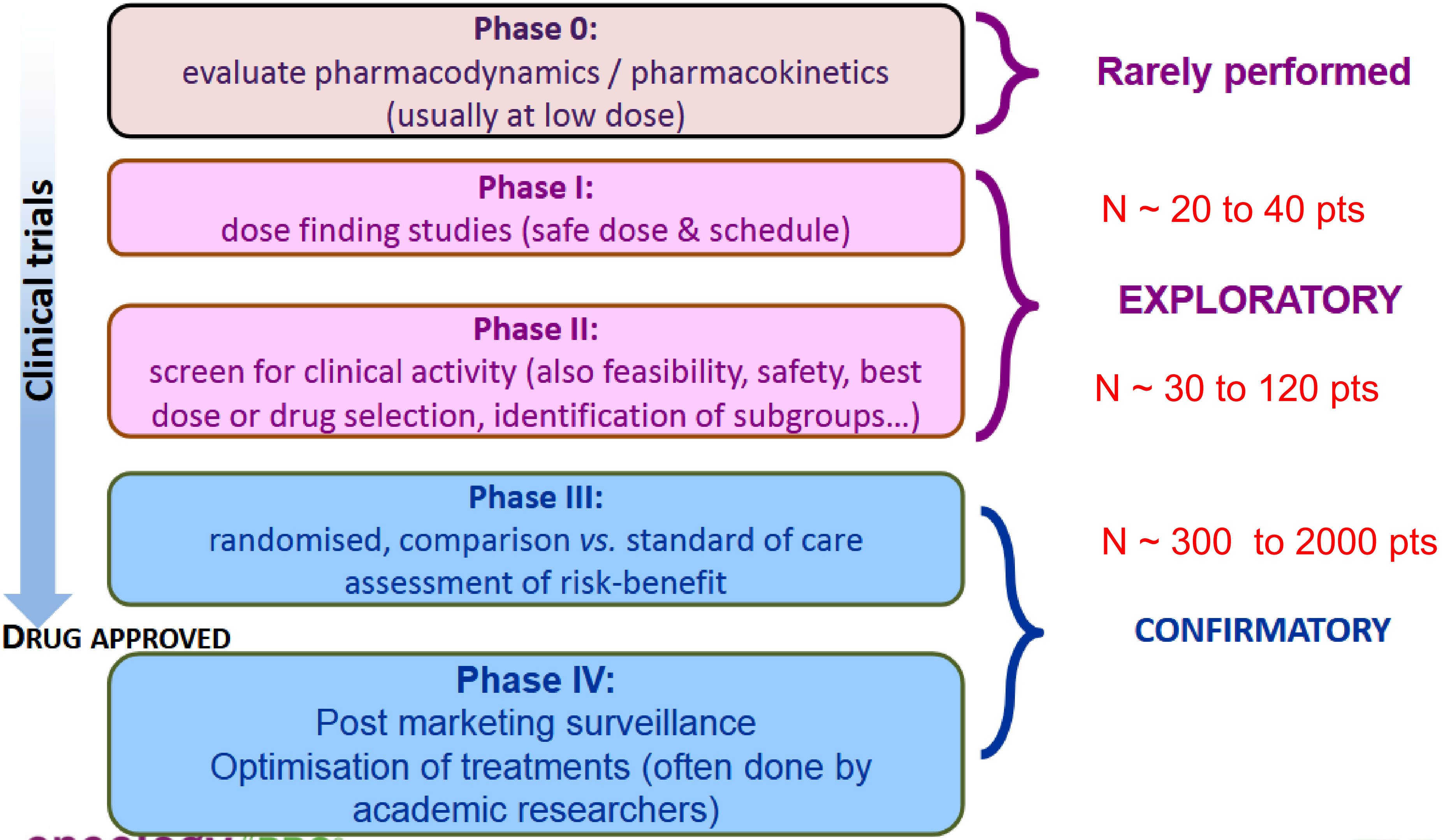
Design and Analysis of Randomized Phase III Trials

Marinela Capanu
Attending Biostatistician
Department of Epidemiology and Biostatistics
Memorial Sloan Kettering Cancer Center

capanum@mskcc.org

March 17th, 2026
Clinical Trial Design Course Lecture

STEPS IN CLINICAL DEVELOPMENT PROGRAMME



Clinical trials

Phase 0:
evaluate pharmacodynamics / pharmacokinetics
(usually at low dose)

Rarely performed

Phase I:
dose finding studies (safe dose & schedule)

N ~ 20 to 40 pts

EXPLORATORY

Phase II:
screen for clinical activity (also feasibility, safety, best dose or drug selection, identification of subgroups…)

N ~ 30 to 120 pts

Phase III:
randomised, comparison *vs.* standard of care
assessment of risk-benefit

N ~ 300 to 2000 pts

CONFIRMATORY

DRUG APPROVED

Phase IV:
Post marketing surveillance
Optimisation of treatments (often done by academic researchers)

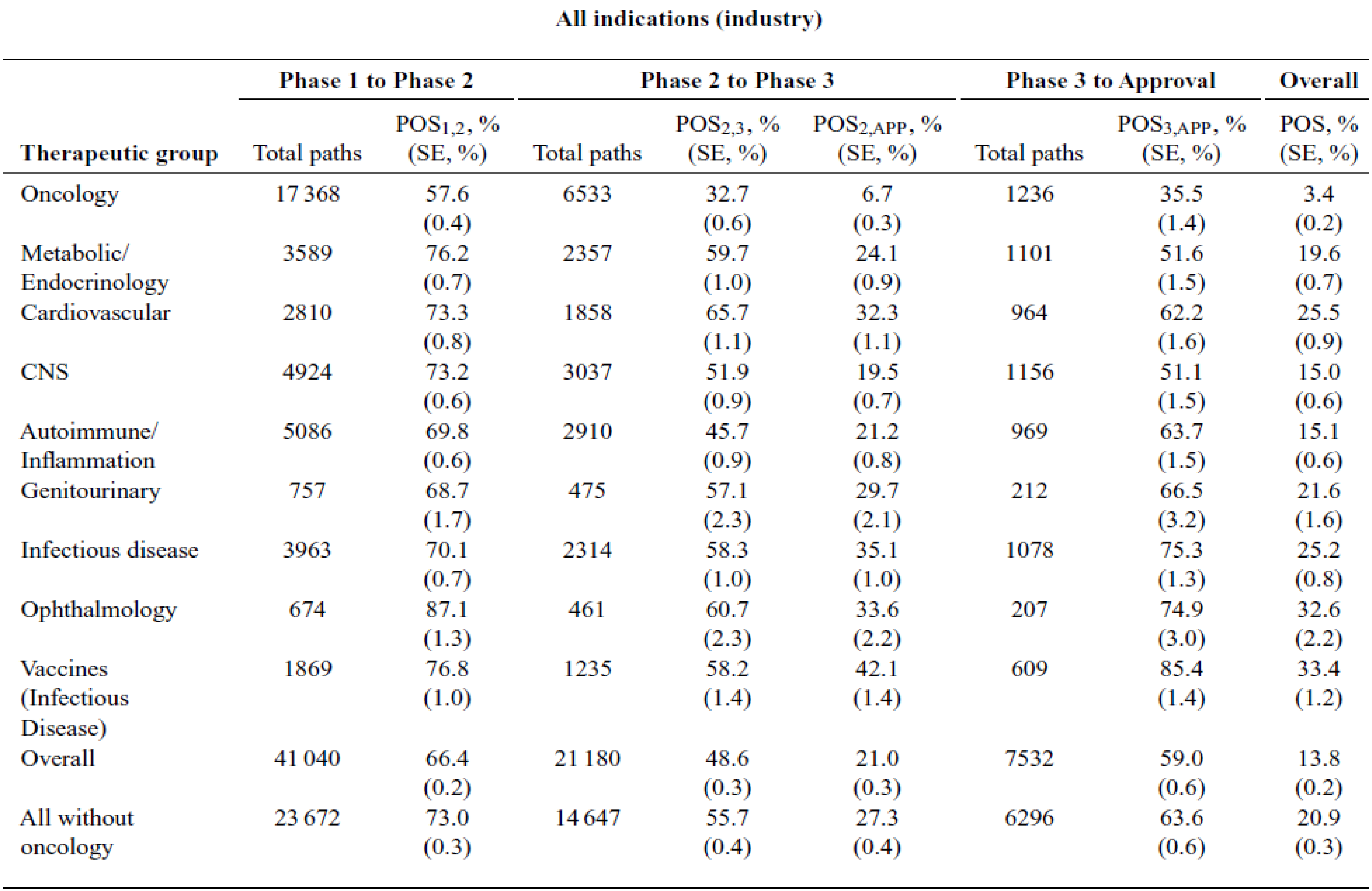
This lecture: Phase III trials

- Phase III trials provide a definitive assessment of how effective one treatment is compared to the standard treatment.

- Most rigorous and extensive clinical investigation of a new therapy.

→ Only ~3-11% of oncology compounds end up being successful during phase I, II and III

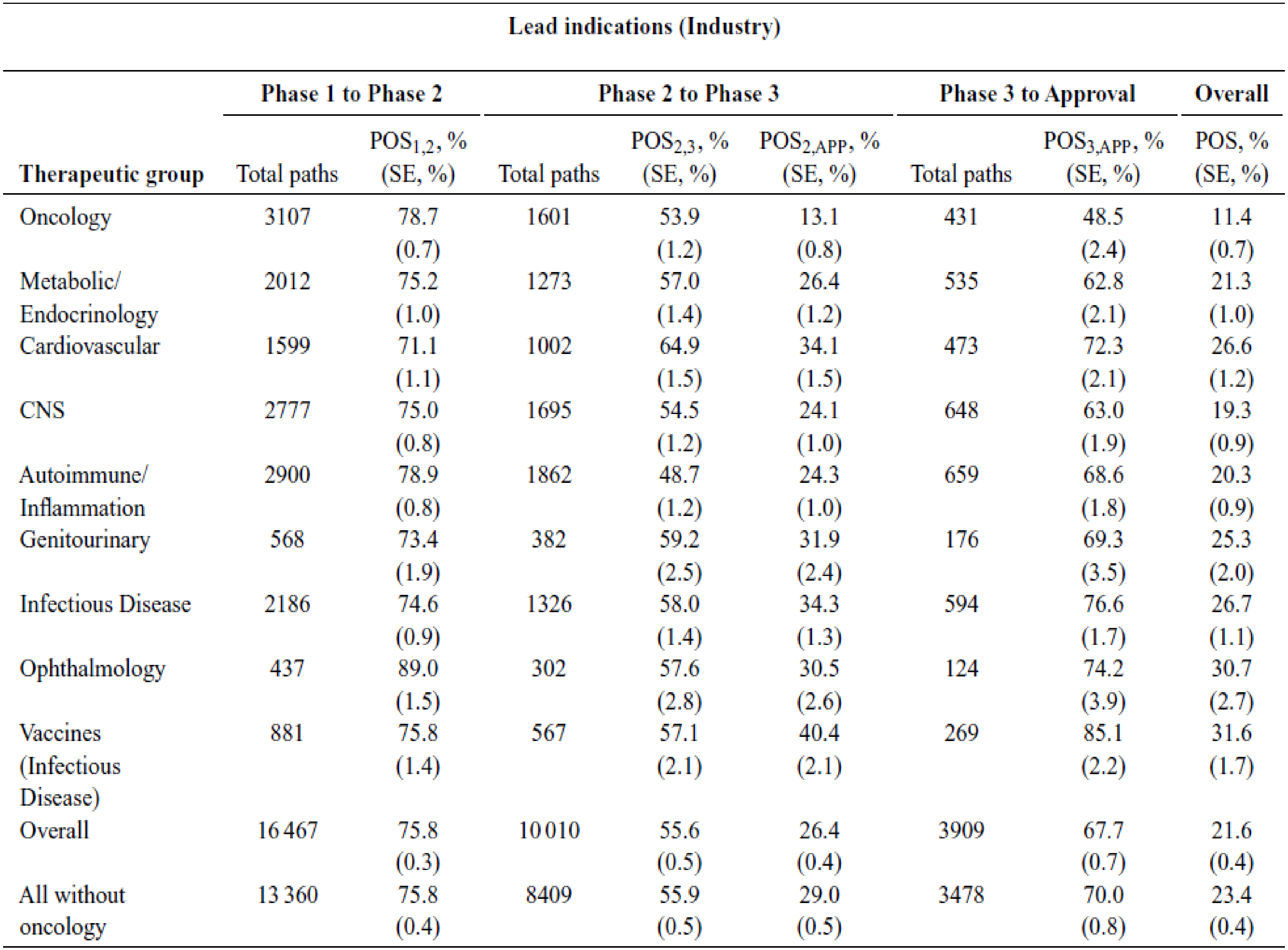
Wong and Siah (2019) Biostatistics



All indications (industry)

	Phase 1 to Phase 2		Phase 2 to Phase 3			Phase 3 to Approval		Overall
Therapeutic group	Total paths	$POS_{1,2}$, % (SE, %)	Total paths	$POS_{2,3}$, % (SE, %)	$POS_{2,APP}$, % (SE, %)	Total paths	$POS_{3,APP}$, % (SE, %)	POS, % (SE, %)
Oncology	17 368	57.6 (0.4)	6533	32.7 (0.6)	6.7 (0.3)	1236	35.5 (1.4)	3.4 (0.2)
Metabolic/ Endocrinology	3589	76.2 (0.7)	2357	59.7 (1.0)	24.1 (0.9)	1101	51.6 (1.5)	19.6 (0.7)
Cardiovascular	2810	73.3 (0.8)	1858	65.7 (1.1)	32.3 (1.1)	964	62.2 (1.6)	25.5 (0.9)
CNS	4924	73.2 (0.6)	3037	51.9 (0.9)	19.5 (0.7)	1156	51.1 (1.5)	15.0 (0.6)
Autoimmune/ Inflammation	5086	69.8 (0.6)	2910	45.7 (0.9)	21.2 (0.8)	969	63.7 (1.5)	15.1 (0.6)
Genitourinary	757	68.7 (1.7)	475	57.1 (2.3)	29.7 (2.1)	212	66.5 (3.2)	21.6 (1.6)
Infectious disease	3963	70.1 (0.7)	2314	58.3 (1.0)	35.1 (1.0)	1078	75.3 (1.3)	25.2 (0.8)
Ophthalmology	674	87.1 (1.3)	461	60.7 (2.3)	33.6 (2.2)	207	74.9 (3.0)	32.6 (2.2)
Vaccines (Infectious Disease)	1869	76.8 (1.0)	1235	58.2 (1.4)	42.1 (1.4)	609	85.4 (1.4)	33.4 (1.2)
Overall	41 040	66.4 (0.2)	21 180	48.6 (0.3)	21.0 (0.3)	7532	59.0 (0.6)	13.8 (0.2)
All without oncology	23 672	73.0 (0.3)	14 647	55.7 (0.4)	27.3 (0.4)	6296	63.6 (0.6)	20.9 (0.3)

Wong and Siah (2019) Biostatistics



Lead indications (Industry)

	Phase 1 to Phase 2		Phase 2 to Phase 3			Phase 3 to Approval		Overall
Therapeutic group	Total paths	$POS_{1,2}$, % (SE, %)	Total paths	$POS_{2,3}$, % (SE, %)	$POS_{2,APP}$, % (SE, %)	Total paths	$POS_{3,APP}$, % (SE, %)	POS, % (SE, %)
Oncology	3107	78.7 (0.7)	1601	53.9 (1.2)	13.1 (0.8)	431	48.5 (2.4)	11.4 (0.7)
Metabolic/ Endocrinology	2012	75.2 (1.0)	1273	57.0 (1.4)	26.4 (1.2)	535	62.8 (2.1)	21.3 (1.0)
Cardiovascular	1599	71.1 (1.1)	1002	64.9 (1.5)	34.1 (1.5)	473	72.3 (2.1)	26.6 (1.2)
CNS	2777	75.0 (0.8)	1695	54.5 (1.2)	24.1 (1.0)	648	63.0 (1.9)	19.3 (0.9)
Autoimmune/ Inflammation	2900	78.9 (0.8)	1862	48.7 (1.2)	24.3 (1.0)	659	68.6 (1.8)	20.3 (0.9)
Genitourinary	568	73.4 (1.9)	382	59.2 (2.5)	31.9 (2.4)	176	69.3 (3.5)	25.3 (2.0)
Infectious Disease	2186	74.6 (0.9)	1326	58.0 (1.4)	34.3 (1.3)	594	76.6 (1.7)	26.7 (1.1)
Ophthalmology	437	89.0 (1.5)	302	57.6 (2.8)	30.5 (2.6)	124	74.2 (3.9)	30.7 (2.7)
Vaccines (Infectious Disease)	881	75.8 (1.4)	567	57.1 (2.1)	40.4 (2.1)	269	85.1 (2.2)	31.6 (1.7)
Overall	16 467	75.8 (0.3)	10 010	55.6 (0.5)	26.4 (0.4)	3909	67.7 (0.7)	21.6 (0.4)
All without oncology	13 360	75.8 (0.4)	8409	55.9 (0.5)	29.0 (0.5)	3478	70.0 (0.8)	23.4 (0.4)

Not all drugs are a success story

- What can we learn from a negative trial to inform future trials /hypotheses?

- Did we have sufficient power and reliable data (rigorous) to answer the questions:
  - Why did the drug/combination fail?
    - Wrong schedule /dose?
    - Too small sample size?
    - Did we choose the wrong patient population?
    - Is there efficacy in some subpopulation?
    - Was our historical control or estimate off?

RCT Phase III trials

- In cancer (and in medicine generally), randomized controlled trial (RCT) is the most common design

  -Objective is to *directly* compare two or more treatments

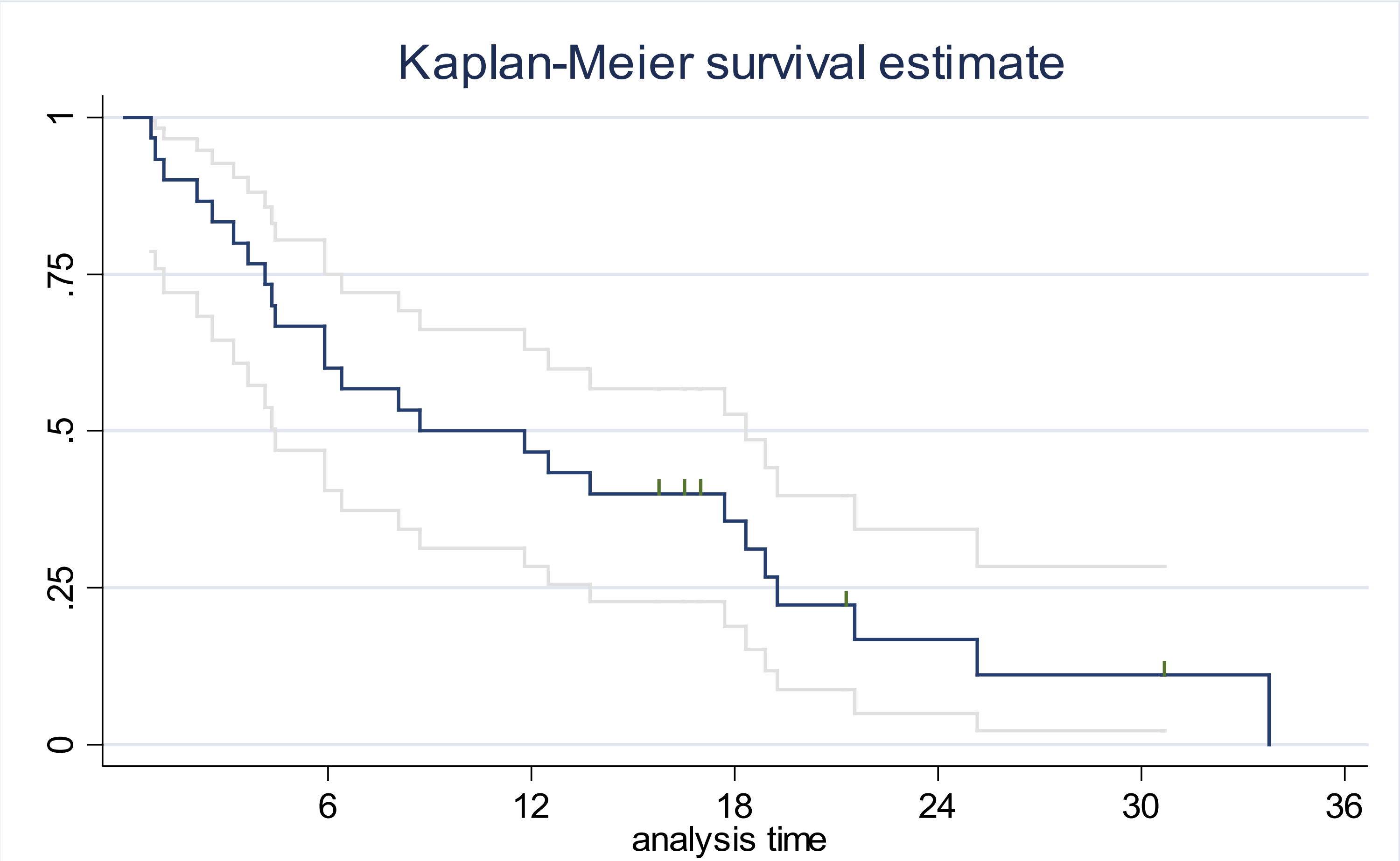
  -Randomization eliminates selection bias and balances the arms with respect to known and unknown confounders

  -Usually double-blinded

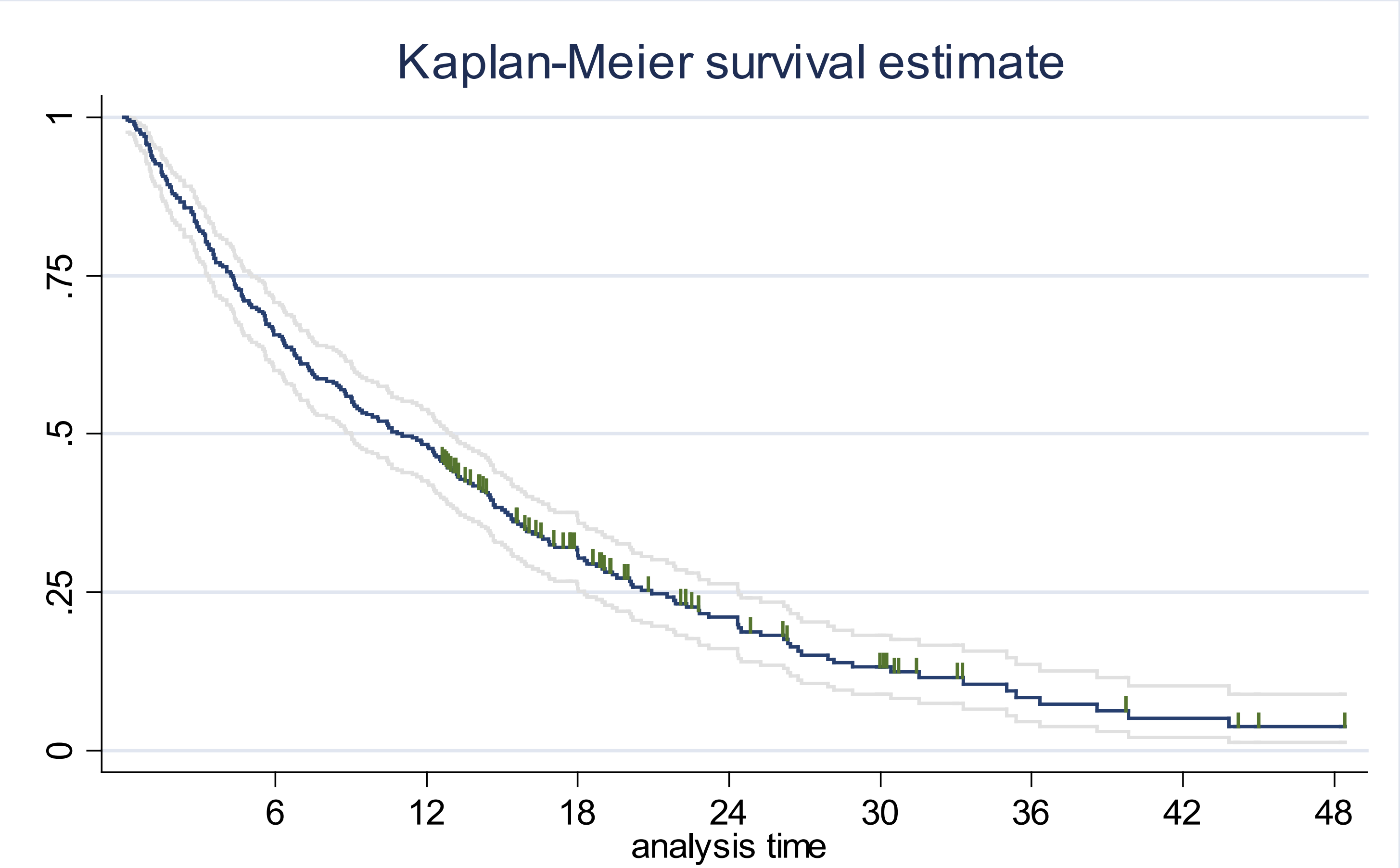
  -These are large, multicenter trials, can be lengthy, and costly

Essential for licensing of new drugs, FDA approval

KM curve from a Phase II with 95% CI



KM curve from a Phase III



Elements of a Phase III trial

-What is the *hypothesis*?

- -Compare two or more treatments for:
    - -Superiority
    - -Equivalence
    - -Noninferiority

-What is the *primary endpoint(s)*?

- -In cancer trials, common endpoints include
    - -Overall survival (OS)
    - -Progression-free survival (PFS)
    - -Others could include RR, QoL

-How about *interim analyses*?

-What are the *secondary endpoints*?

- -Quality of life, safety, test other ‘subhypotheses’, generate new hypotheses
- -Subset analyses?

At the conclusion of a RCT

-P-value comparing two or more treatment groups in terms of OS/PFS

-Hazard ratio estimate with confidence interval

-Median OS/PFS or x-year OS/PFS with confidence interval

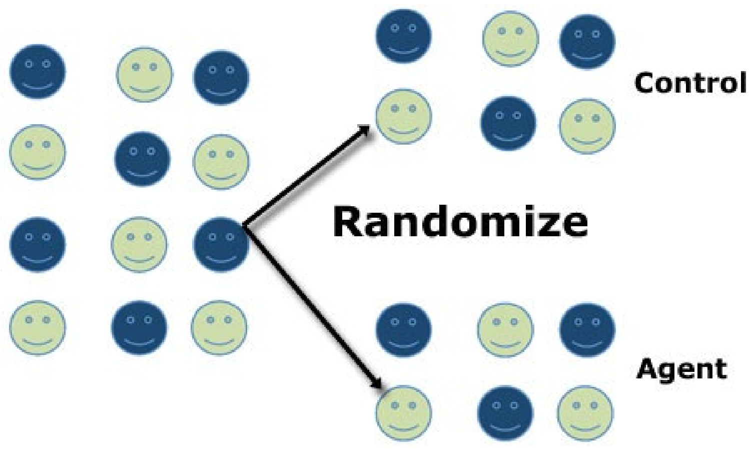
-Analysis of secondary endpoints, correlative and subset studies

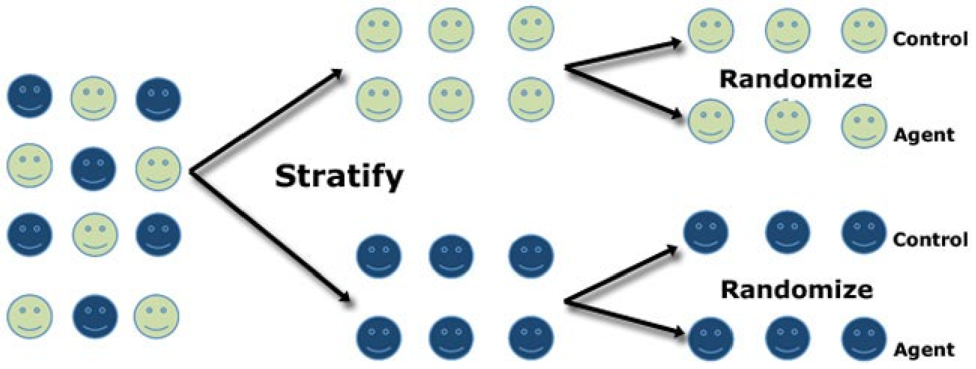
Elements of a Phase III trial

-Should the trial have *stratified randomization*?

-Randomization ensures balance in known and unknown patient attributes

-*Stratified* randomization ensures balance for patient characteristics known to be related with the primary endpoint





Stratified randomization reduces imbalance in prognostic factors between arms

Elements of a Phase III trial

-Blinding

-Double-blinding (patient nor investigator knows treatment assignment) reduces bias in the final analysis.

-Used to prevent influence of placebo effect or observer bias on the final analysis

-An external group (eg DSMB) will regularly monitor the trial for toxicity and benefit.

Other Topics

- Analyses are usually conducted using the Intent-to-Treat (ITT) principle: all patients should be analyzed in the groups they were originally randomized to, irrespective of the treatment they actually received

    -Can define per-protocol analysis as needed: includes only patients fully compliant with the clinical trial protocol

- Cross-overs: With the PFS endpoint, patients who progress are often allowed to go on to the other arm or on to other treatments. Analysis can be tricky to interpret (e.g. OS)

Concepts

- Null hypothesis
    - assertion that one wants to reject, e.g. no effect
- Alternative hypothesis
    - hypothesis one wants to prove, e.g. trt has an effect
- α-level or Type I error
    - False positive probability (usually set at 5%)
- Power=1-Type II error
    - Probability of seeing a true positive given that the alternative hypothesis is true.
    - Type II error: false negative probability
    - For phase III studies, power is often set at 80 to 90%

Choice of endpoint

- *Overall survival* or *progression-free survival* are common endpoints in Phase III cancer trials
  - -OS is the preferred endpoint for many reasons, but in some diseases not feasible to use.

- Time-to-event endpoints

Example: Breast Cancer Endpoints

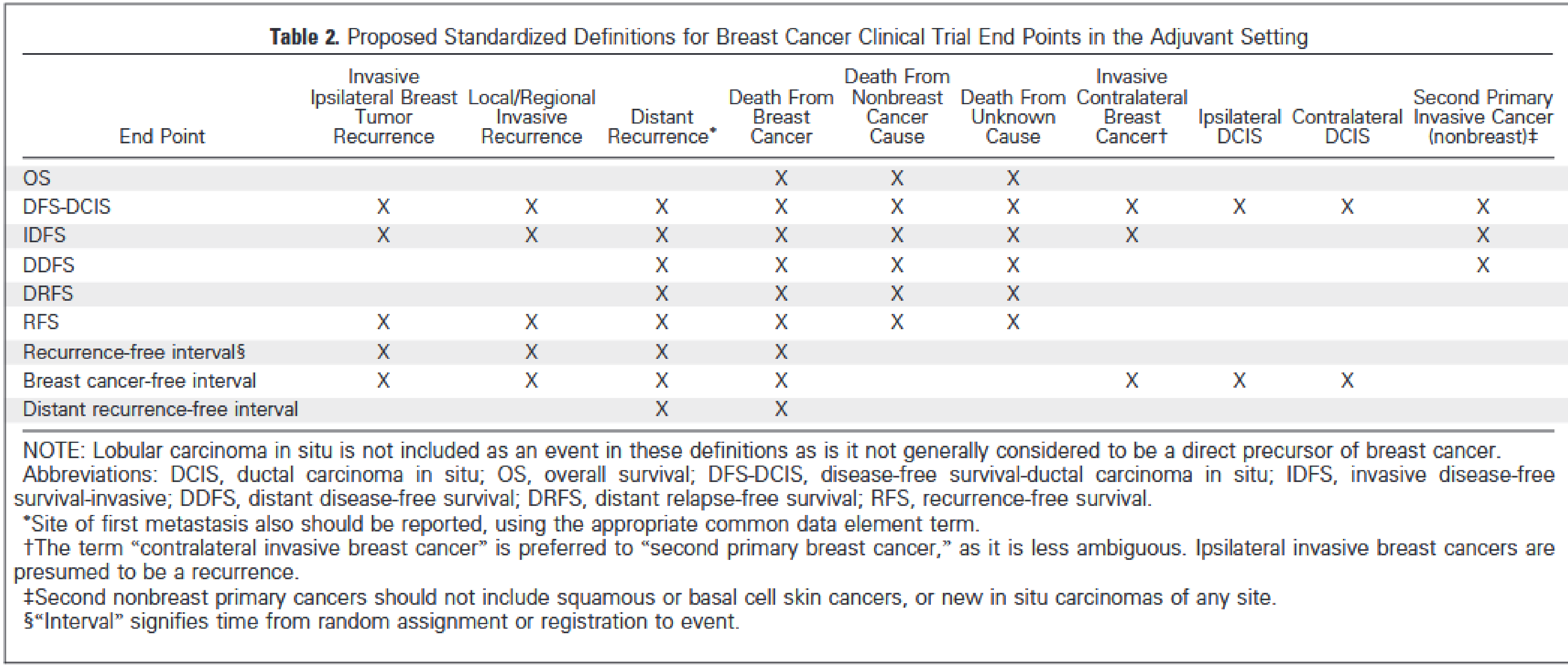


Table 2. Proposed Standardized Definitions for Breast Cancer Clinical Trial End Points in the Adjuvant Setting

End Point	Invasive Ipsilateral Breast Tumor Recurrence	Local/Regional Invasive Recurrence	Distant Recurrence*	Death From Breast Cancer	Death From Nonbreast Cancer Cause	Death From Unknown Cause	Invasive Contralateral Breast Cancer†	Ipsilateral DCIS	Contralateral DCIS	Second Primary Invasive Cancer (nonbreast)‡
OS				X	X	X				
DFS-DCIS	X	X	X	X	X	X	X	X	X	X
IDFS	X	X	X	X	X	X	X			X
DDFS			X	X	X	X				X
DRFS			X	X	X	X				
RFS	X	X	X	X	X	X				
Recurrence-free interval§	X	X	X	X						
Breast cancer-free interval	X	X	X	X			X	X	X	
Distant recurrence-free interval			X	X						

NOTE: Lobular carcinoma in situ is not included as an event in these definitions as is it not generally considered to be a direct precursor of breast cancer.

Abbreviations: DCIS, ductal carcinoma in situ; OS, overall survival; DFS-DCIS, disease-free survival-ductal carcinoma in situ; IDFS, invasive disease-free survival-invasive; DDFS, distant disease-free survival; DRFS, distant relapse-free survival; RFS, recurrence-free survival.

*Site of first metastasis also should be reported, using the appropriate common data element term.

†The term "contralateral invasive breast cancer" is preferred to "second primary breast cancer," as it is less ambiguous. Ipsilateral invasive breast cancers are presumed to be a recurrence.

‡Second nonbreast primary cancers should not include squamous or basal cell skin cancers, or new in situ carcinomas of any site.

§"Interval" signifies time from random assignment or registration to event.

STEEP system, Hudis et al JCO 2007

Time-to-event endpoints
Censoring

-Need to account for censoring

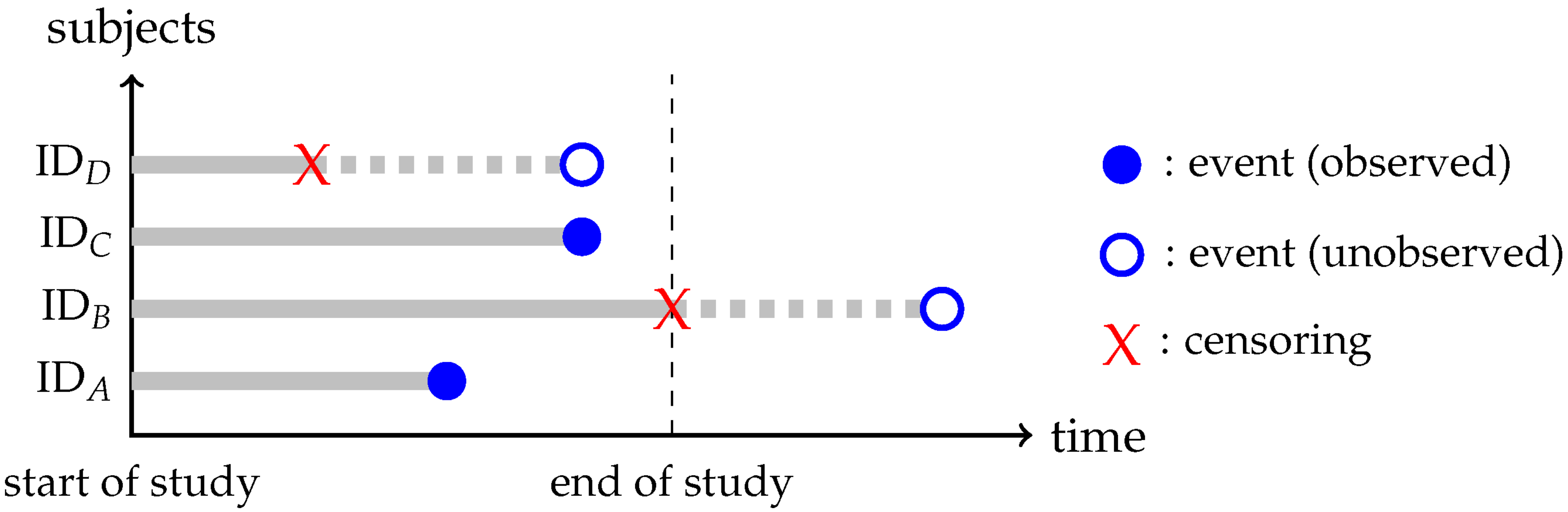
-Patient does not experience event of interest by end of study

-Patient is lost to followup (withdrew consent, moved)

-> Commonly, we know a patient has lived longer than a certain number of days, but not how much longer

-> Less information than if someone experiences the event of interest, but it is still information that you want to use in the analysis

Right-Censoring



Note: Where you censor matters

https://www.mdpi.com/2504-4990/1/3/58/htm

Time-to-Event Endpoints
Median Survival

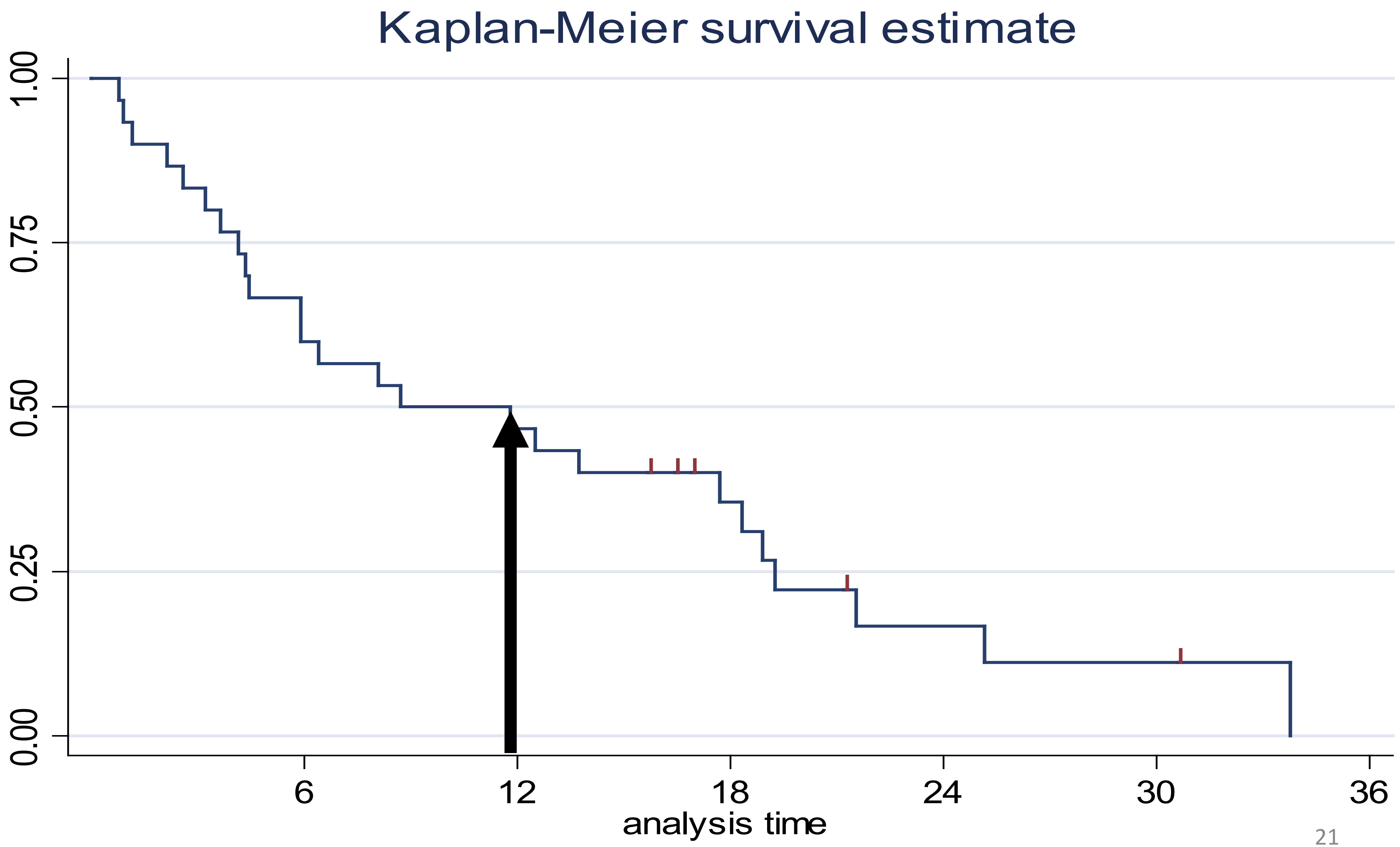
- Use life-table methods (most commonly used: Kaplan-Meier method) to estimate probabilities of survival at each timepoint following the start of treatment.

    -This takes into account censoring

    -Censoring is assumed noninformative: censoring time is independent of the survival time; for example, drop outs should be unrelated to the endpoint of interest

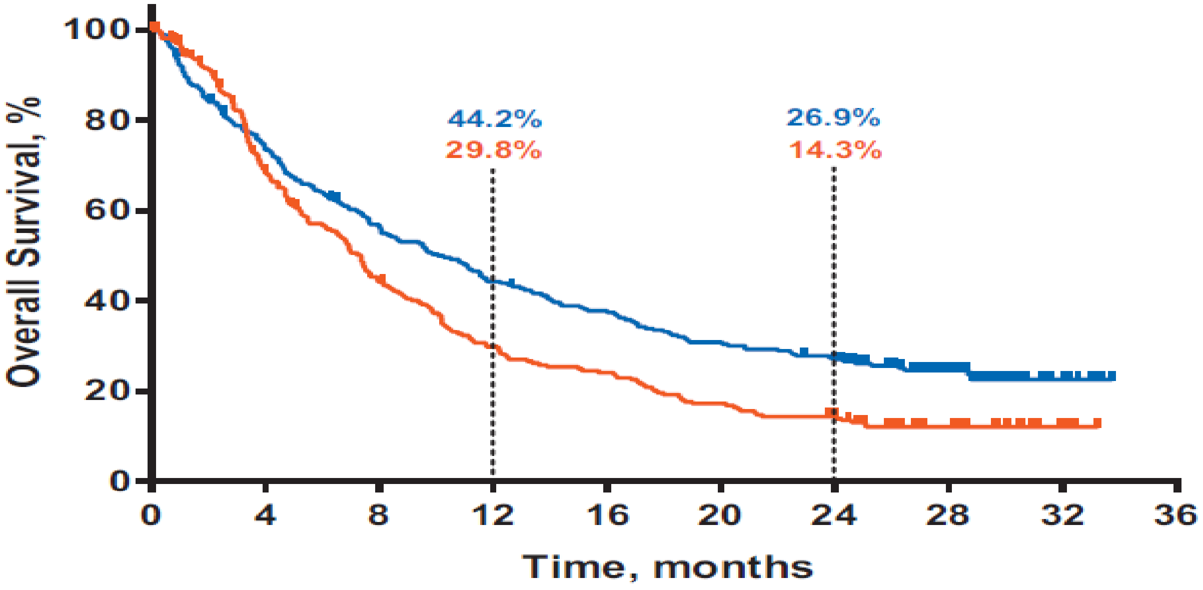
-Median survival: time at which the survival probability first drops below 50%.

Median Survival



KEYNOTE-045: pembrolizumab vs chemotherapy in recurrent advanced urothelial cancer

Median OS		
Pembrolizumab	10.1 months (95% CI, 8.0–12.3 months)	HR = 0.70, 95% CI = 0.57–0.85
Chemotherapy	7.3 months (95% CI, 6.1–8.1 months)	*P* = 0.00015

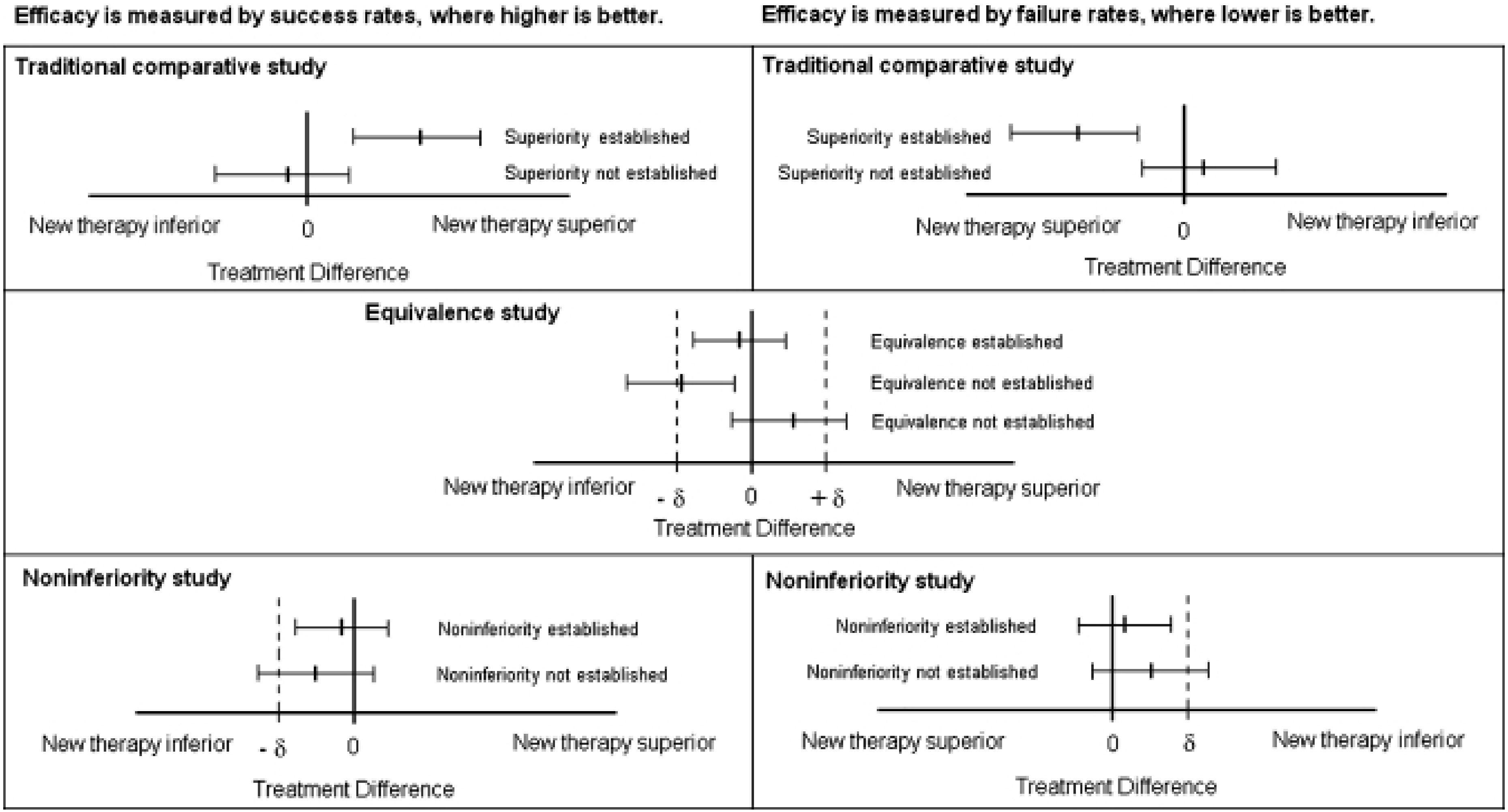


n at risk										
Pembrolizumab	270	195	148	116	98	80	67	33	7	0
Chemotherapy	272	173	109	73	59	42	34	18	4	0

Overview of Common Hypotheses

	Superiority	Equivalence	Non-Inferiority
Purpose	detect a difference btwn treatments	confirm absence of meaningful difference, δ, btwn treatments	new trtm is no less effective than an existing trtm by some specific amount, δ
Null	$X = Y$ No difference	$-\delta \geq (X - Y)$ or $(X - Y) \geq \delta$ Not equivalent	$(X - Y) \geq \delta$ Y inferior to X
Alternative	$X < Y$ or $X > Y$ There is difference	$-\delta < (X - Y) < \delta$ Equivalent	$(X - Y) < \delta$ Y non-inferior to X
	Usually 2-sided	Usually 2-sided	1-sided

X = Effect of Treatment X (standard) Y = Effect of Treatment Y (experimental)



https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=3019319_11606_2010_1513_Fig1_HTML.jpg

Superiority

-1:1 randomization common

-Power (80%), type I error (5%), clinically meaningful difference, accrual per unit of time, committed length of followup, expected number of events are all interconnected in these sample size calculations.

- Can get more complicated if we know and can expect certain patterns of dropout.

Superiority—Effect Sizes

- What is the sample size required to detect a difference in 5-year survival from 75% to 80% in a 1:1 two-arm randomized trial?

    - Corresponds to a hazard ratio of .78
    - Corresponds to change in median survival from 12 to 15 yrs
    - Sample size for 80% power: 1030 pts, 490 events, 18 yrs

-What is the sample size required to detect a difference in 5-year survival from 75% to 85% in a 1:1 two-arm randomized trial?

    - Corresponds to a hazard ratio of .57
    - Corresponds to change in median survival from 12 to 21 yrs
    - Sample size for 80% power: 405 pts, 102 events, 9 years

-> smaller differences require larger sample sizes

Superiority—Low vs High Risk

- There will be more censoring when studying a low risk population than when studying a high risk population over a similar followup time period.

-Suppose we are looking for a 50% increase in median survival. Assume it takes three years to accrue patients with one year of followup

-If median survival is 9 months: n=315

-If median survival is 5 years: n=1050

-If median survival is 8 years: n=1575

-> Longer survival means fewer events. This provides less information and requires more patients.

Equivalence

-Designed to confirm the *absence* of a meaningful difference, δ, between treatments

- Why? New treatment may be less expensive, or have less toxicity.

-Generally requires large sample sizes. Difficult to determine what δ should be.

➢ How do we define the margin, δ, for equivalence?

Noninferiority

- Can be thought of as a 1-sided equivalence trial

- Increasingly being used, requires smaller sample sizes compared to equivalence trials

- Designed to show that new treatment is no less effective than an existing treatment by some specific amount, δ

-Very difficult to design. Need a very good estimate of δ

Choice of δ

- Most critical step in designing equivalence/non-inferiority studies

- A small δ determines a narrower equivalence region, and makes it more difficult to establish equivalence/noninferiority.

- δ not only determines the result of the test, but also gives scientific credibility to a study: how well the equivalence margin can be justified in terms of relevant evidence and sound clinical considerations?

-Choose δ based on the margin of superiority of the current therapy against the placebo (estimated from previous studies)

- In noninferiority testing: set δ to a fraction, *f*, of the lower limit of a confidence interval of the difference between the current therapy and the placebo

-The smaller *f*, the more difficult to establish equivalence/noninferiority

Interim Looks

- Interim looks at the data can be incorporated into the design
    - Stop for futility
    - Stop for early efficacy
    - Stop for either early efficacy or futility
- These looks (single or multiple) are usually incorporated into the study design by controlling the overall type I error, must be done apriori
- Has an effect on the trial design parameters, particularly the type 1 error.

Multiple Endpoints

- Example: PFS and OS

- Need to control Type I error for multiplicity
  - Eg Bonferroni Adjustments
  - Eg Gatekeeping procedures or hierarchical testing: two families of outcome variables, e.g., primary variables (gatekeeper) and secondary variables. The gatekeeper family is tested without an adjustment for the other family, and the second family is examined only if the gatekeeper has been successfully passed;

https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm536750.pdf

Innovations

- Integrated Phase II/III design
  - Phase II portion uses PFS (or some other molecular/imaging intermediate variable) as endpoint
  - If p-value testing difference in PFS is above a threshold after a pre-defined number of patients, accrual will terminate.
  - Otherwise, accrual continues to final sample size and a test comparing OS is conducted (Hunsberger, Zhao, Simon)
  - Advantages: when appropriate, it can result in smaller sample size, time and resources savings, and shorter study duration

Integrated Phase II/III design

- Appropriate when
  - ➢ a rapidly obtained Phase II endpoint, correlated with the primary endpoint in the Phase III component, is available;
  - ➢ positive Phase II results are sufficient motivation for the launch of the Phase III component, and other needs such as budgetary support, patient accrual, and drug distribution are in place.
- Inappropriate when:
  - ➢ insufficient evidence to warrant the implementation of Phase III component after the positive Phase II results
  - ➢ a Phase II endpoint that requires a lengthy period of time to observe failures with the Phase III endpoint observed soon after the Phase II endpoint, and thus, little savings in patient numbers or duration of study would be observed
  - ➢ insufficient resources to implement a Phase III trial after positive Phase II results

Innovations

- Using genomic predictive biomarkers in Phase III design
  - Use validated predictive marker
    - Enrichment design: Include only those patients with the marker
    - Stratification design: Include patients with and without marker, randomization is stratified so that both arms are balanced
- Group sequential designs
- Other adaptive designs
  - FDA guidance: https://www.fda.gov/downloads/drugs/guidances/ucm201790.pdf

Basket, umbrella, platform trials

Woodcock J, and LaVange L 2017 NEJM

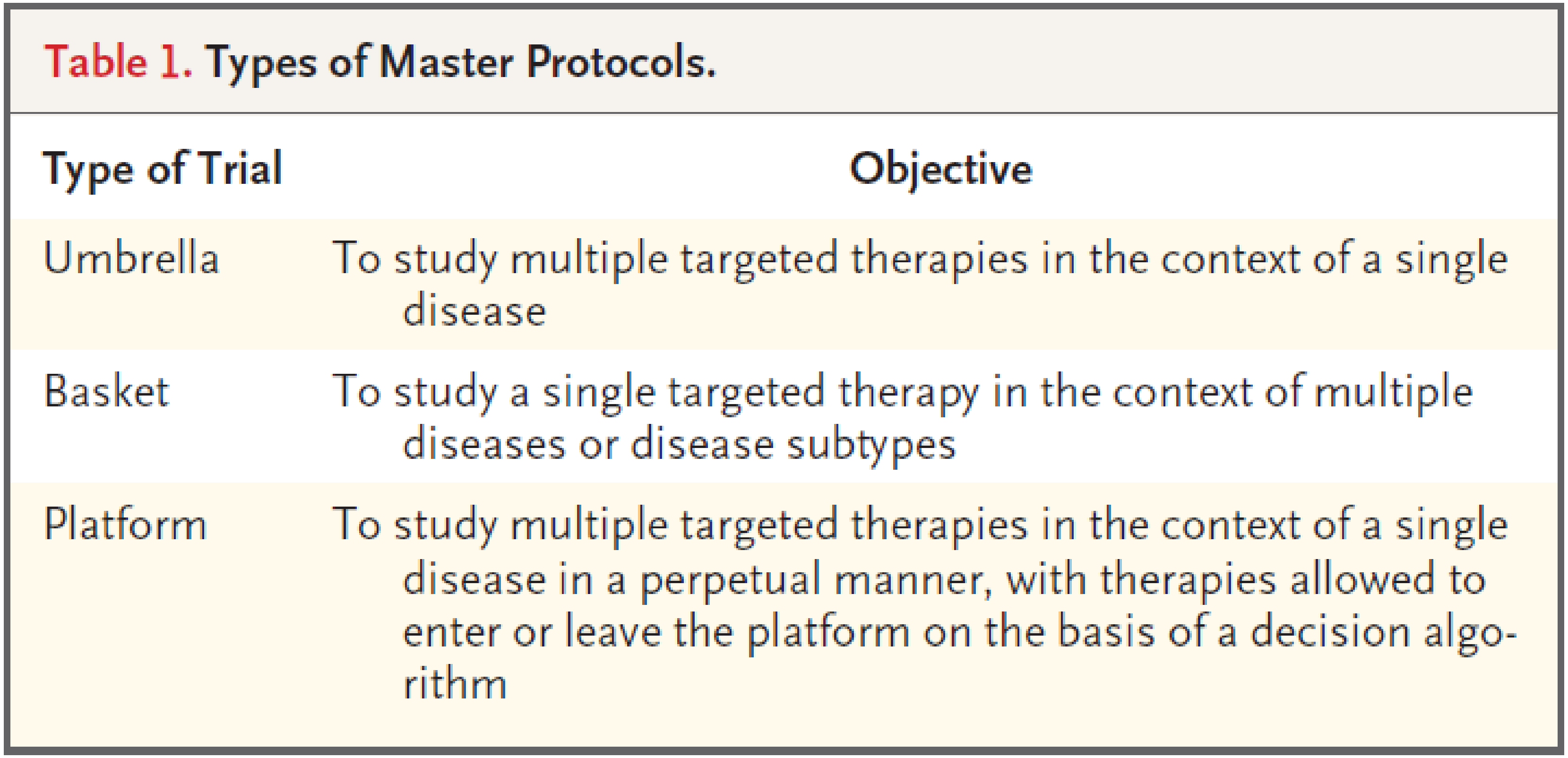
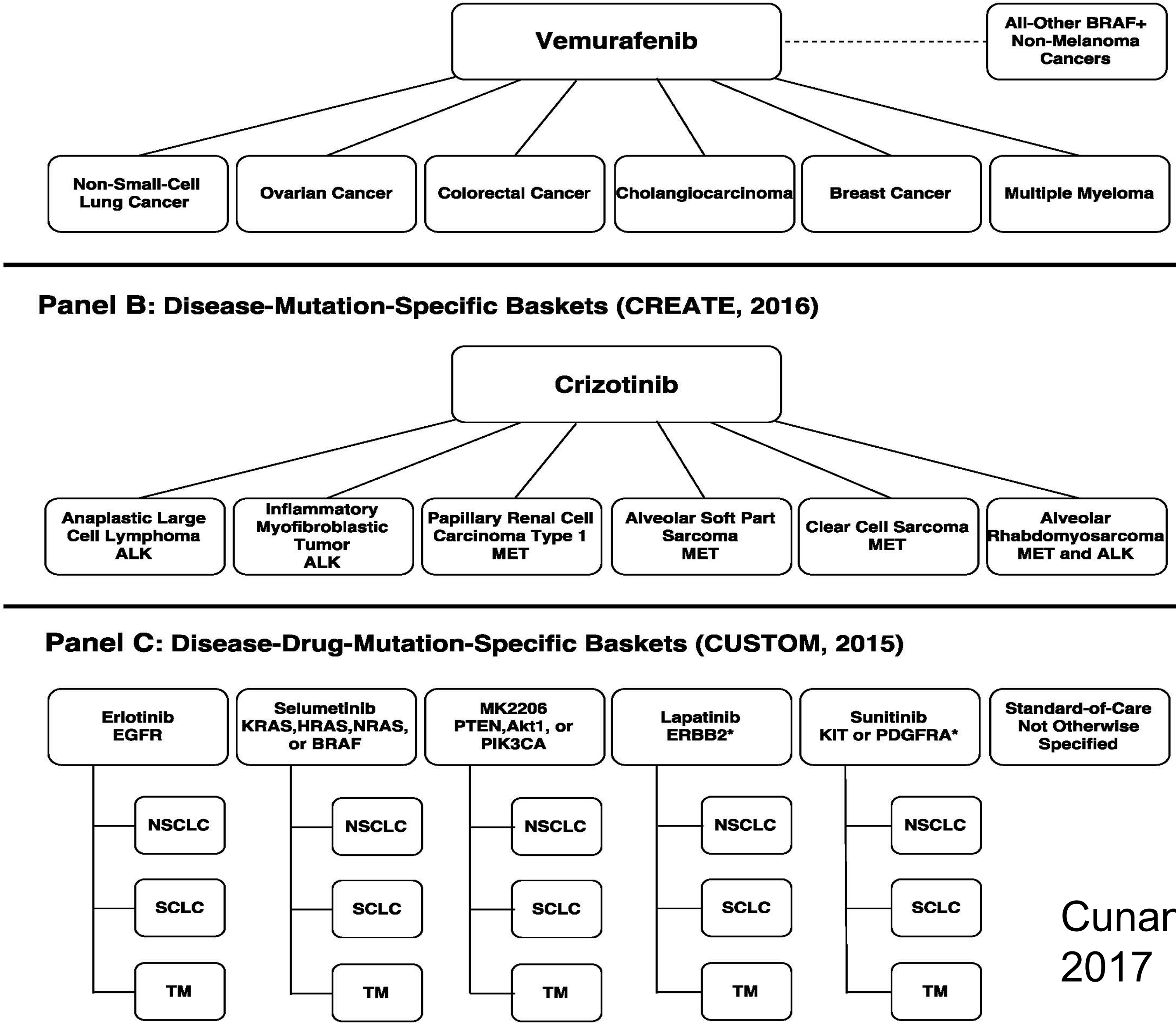


Table 1. Types of Master Protocols.

Type of Trial	Objective
Umbrella	To study multiple targeted therapies in the context of a single disease
Basket	To study a single targeted therapy in the context of multiple diseases or disease subtypes
Platform	To study multiple targeted therapies in the context of a single disease in a perpetual manner, with therapies allowed to enter or leave the platform on the basis of a decision algorithm



Panel A: Disease-Specific Baskets (Hyman et al., 2015)

Vemurafenib

All-Other BRAF+ Non-Melanoma Cancers

Non-Small-Cell Lung Cancer

Ovarian Cancer

Colorectal Cancer

Cholangiocarcinoma

Breast Cancer

Multiple Myeloma

Panel B: Disease-Mutation-Specific Baskets (CREATE, 2016)

Crizotinib

Anaplastic Large Cell Lymphoma ALK

Inflammatory Myofibroblastic Tumor ALK

Papillary Renal Cell Carcinoma Type 1 MET

Alveolar Soft Part Sarcoma MET

Clear Cell Sarcoma MET

Alveolar Rhabdomyosarcoma MET and ALK

Panel C: Disease-Drug-Mutation-Specific Baskets (CUSTOM, 2015)

Erlotinib EGFR

Selumetinib KRAS,HRAS,NRAS, or BRAF

MK2206 PTEN,Akt1, or PIK3CA

Lapatinib ERBB2*

Sunitinib KIT or PDGFRA*

Standard-of-Care Not Otherwise Specified

NSCLC

SCLC

TM

Cunanan et al JCO 2017

What is I-SPY?

The I-SPY series of trials are changing the way new treatments are developed for breast cancer, helping make available new, better and more personalized treatments, faster. At the heart of the I-SPY program is the ground-breaking I-SPY 2 platform trial for neoadjuvant treatment of locally advanced breast cancer.

1400

patients enrolled

16

agents completed evaluation since 2010

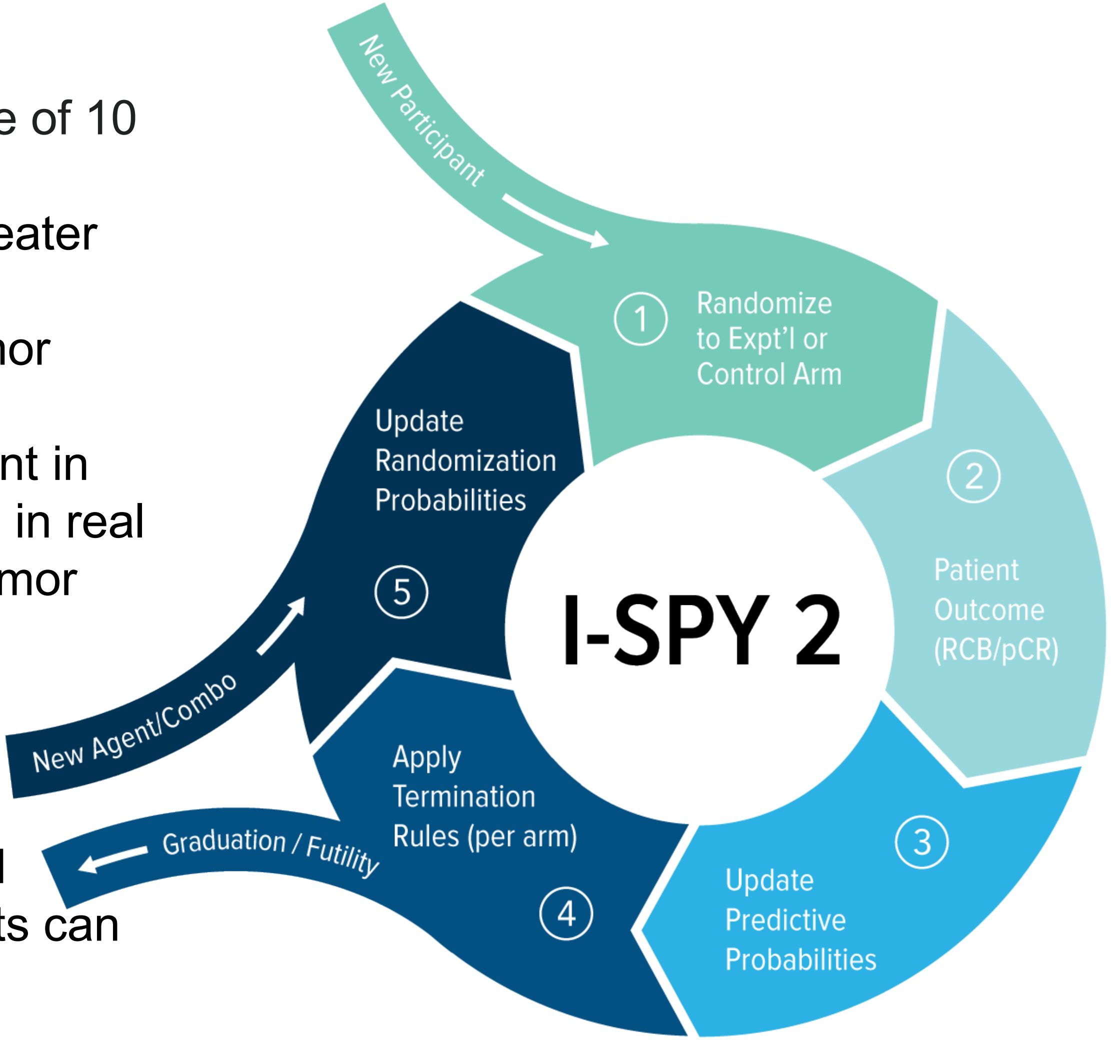
3

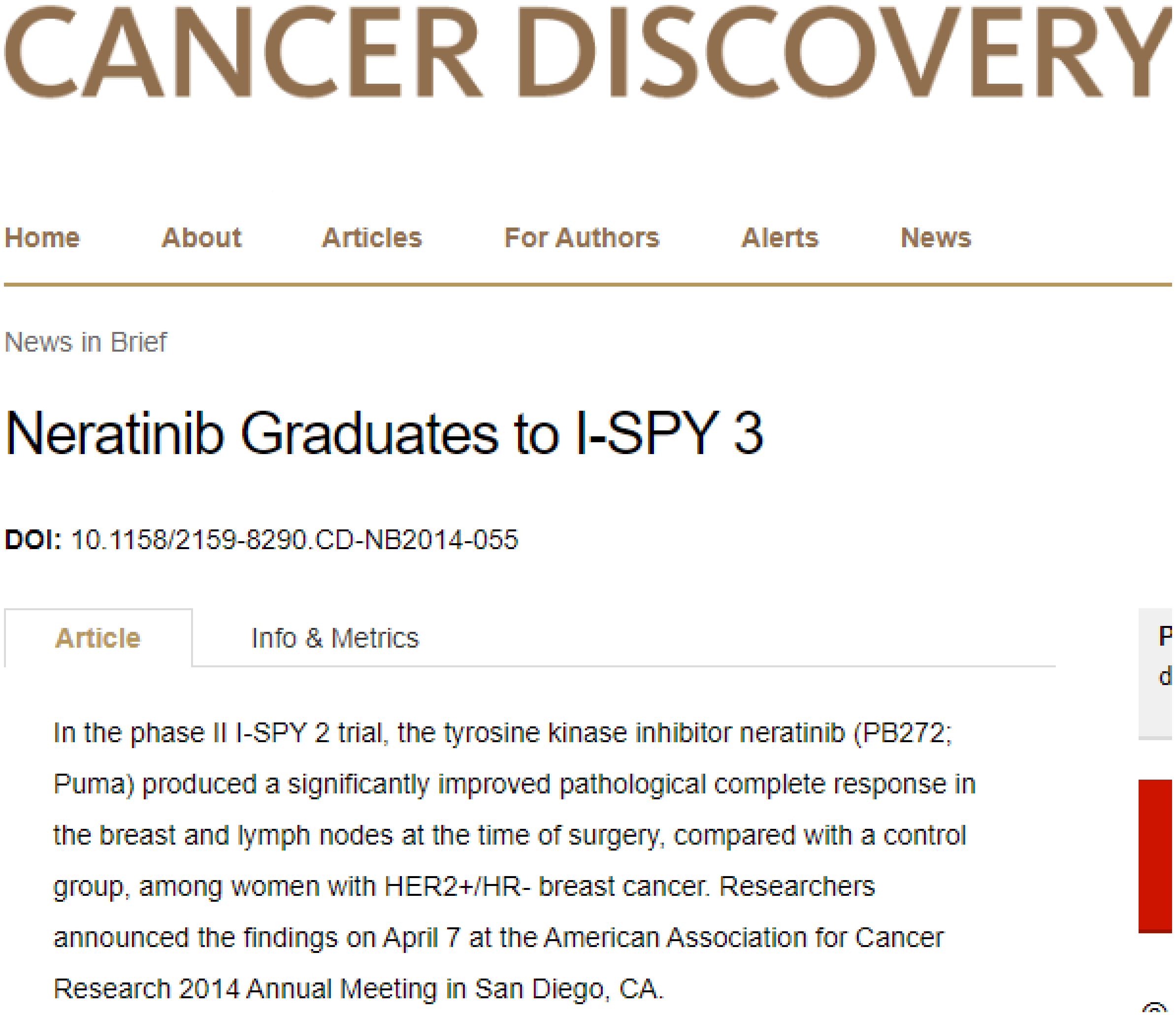
agents received accelerated approval

How I-SPY 2 Works

- I-SPY 2 breaks from the traditional randomized clinical trial model, employing an 'adaptive' model designed to increase trial efficiency by minimizing the number of participants and time required to evaluate an experimental agent.

- New participant classified into one of 10 molecular subtypes
- Adaptive randomization; gives greater weight to arms that have been successful in the participant's tumor subtype
- Predictive probabilities of the agent in the various subtypes are updated in real time based on the participant's tumor subtype, outcome and treatment received
- If pre-determined level of efficacy reached: declared a success ("graduates"). If not, then stopped for futility. At any point new agents can enter the trial through a protocol amendment.
- The participant's serial MRI measures, outcome and tumor subtype are used to update the prior probabilities of the randomization engine -- over time this refines the targeting of subsequent participants.


New Participant
1
Randomize to Expt'l or Control Arm
2
Patient Outcome (RCB/pCR)
3
Update Predictive Probabilities
4
Apply Termination Rules (per arm)
Graduation / Futility
5
Update Randomization Probabilities
New Agent/Combo
I-SPY 2



News in Brief

Neratinib Graduates to I-SPY 3

DOI: 10.1158/2159-8290.CD-NB2014-055

In the phase II I-SPY 2 trial, the tyrosine kinase inhibitor neratinib (PB272; Puma) produced a significantly improved pathological complete response in the breast and lymph nodes at the time of surgery, compared with a control group, among women with HER2+/HR- breast cancer. Researchers announced the findings on April 7 at the American Association for Cancer Research 2014 Annual Meeting in San Diego, CA.

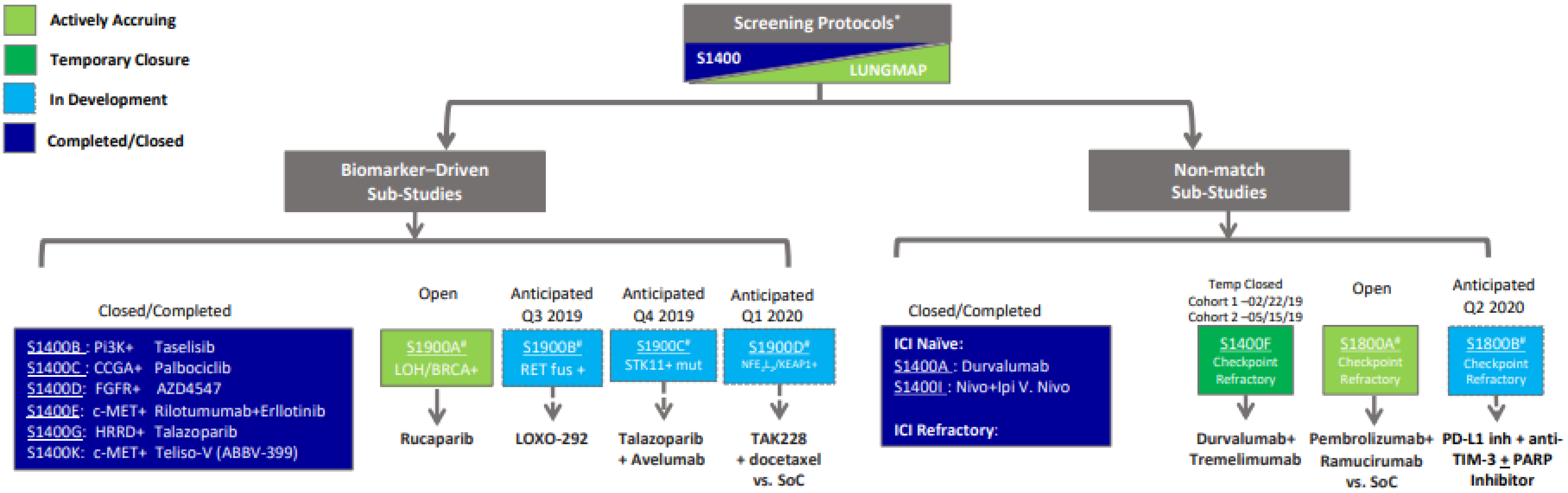
Innovations: Lung-MAP, Umbrella

- Lung-MAP is designed to quickly and efficiently test new treatments for advanced non-small cell lung cancer

- Lung-MAP is the first precision medicine trial in lung cancer supported by NCI and is the first major NCI trial to test multiple treatments, simultaneously, under one “umbrella” design

- Patients who enroll in Lung-MAP get a state-of-the-art genomic profile to determine the genomic alterations, or mutations, which may drive the growth of their cancer. Based on those results, patients are matched to a treatment being tested on Lung-MAP. If there isn’t a genomic “match,” patients have an option of receiving immunotherapy treatments used in the trial.

- Launched in 2014, Lung-MAP has registered nearly 2,000 patients, and some have received new treatments that have extended or improved their lives.



🟩 Actively Accruing

🟩 Temporary Closure

🟦 In Development

🟦 Completed/Closed

Screening Protocols*

S1400 LUNGMAP

Biomarker–Driven Sub-Studies

Closed/Completed

S1400B : Pi3K+ Taselisib
S1400C : CCGA+ Palbociclib
S1400D: FGFR+ AZD4547
S1400E: c-MET+ Rilotumumab+Erllotinib
S1400G: HRRD+ Talazoparib
S1400K: c-MET+ Teliso-V (ABBV-399)

Open
S1900A#
LOH/BRCA+
Rucaparib

Anticipated Q3 2019
S1900B#
RET fus +
LOXO-292

Anticipated Q4 2019
S1900C#
STK11+ mut
Talazoparib + Avelumab

Anticipated Q1 2020
S1900D#
NFE$_2$L$_2$/KEAP1+
TAK228 + docetaxel vs. SoC

Non-match Sub-Studies

Closed/Completed

ICI Naïve:
S1400A : Durvalumab
S1400I : Nivo+Ipi V. Nivo

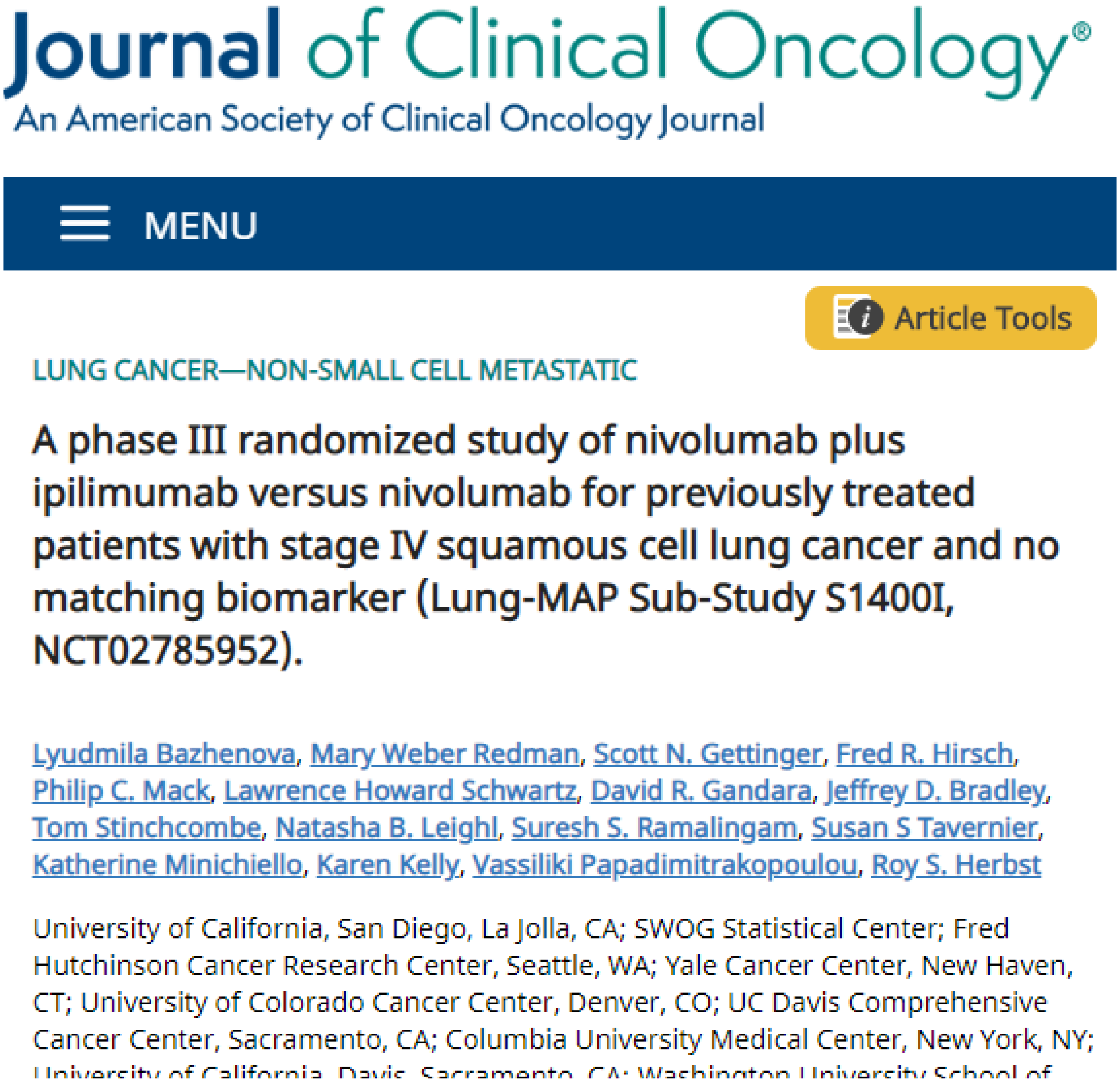
ICI Refractory:

Temp Closed
Cohort 1 –02/22/19
Cohort 2 –05/15/19
S1400F
Checkpoint Refractory
Durvalumab+ Tremelimumab

Open
S1800A#
Checkpoint Refractory
Pembrolizumab+ Ramucirumab vs. SoC

Anticipated Q2 2020
S1800B#
Checkpoint Refractory
PD-L1 inh + anti-TIM-3 ± PARP Inhibitor

A therapy found to be effective in phase II will move directly into the phase III registration setting, incorporating the patients from phase II.



Journal of Clinical Oncology®
An American Society of Clinical Oncology Journal

LUNG CANCER—NON-SMALL CELL METASTATIC

A phase III randomized study of nivolumab plus ipilimumab versus nivolumab for previously treated patients with stage IV squamous cell lung cancer and no matching biomarker (Lung-MAP Sub-Study S1400I, NCT02785952).

Lyudmila Bazhenova, Mary Weber Redman, Scott N. Gettinger, Fred R. Hirsch, Philip C. Mack, Lawrence Howard Schwartz, David R. Gandara, Jeffrey D. Bradley, Tom Stinchcombe, Natasha B. Leighl, Suresh S. Ramalingam, Susan S Tavernier, Katherine Minichiello, Karen Kelly, Vassiliki Papadimitrakopoulou, Roy S. Herbst

University of California, San Diego, La Jolla, CA; SWOG Statistical Center; Fred Hutchinson Cancer Research Center, Seattle, WA; Yale Cancer Center, New Haven, CT; University of Colorado Cancer Center, Denver, CO; UC Davis Comprehensive Cancer Center, Sacramento, CA; Columbia University Medical Center, New York, NY; University of California, Davis, Sacramento, CA; Washington University School of

Final Remarks

- Gan et al (JNCI 2012) report that 62% of Phase III trials do not achieve statistical significance
  - Need more realistic estimates of hypothesized difference
  - More interim analyses
  - Adaptive trials
  - Better biological characterization of patients (biomarkers)
- As efficacy gains become smaller, toxicity becomes more important. Incorporating these events into the conduct and interpretation of Phase II and III trial results important (Amiri-Kordestani, Fojo JNCI 2012)

Concepts
What is a p-value?

- p-value: the probability that an observed result is due to chance alone if the null hypothesis is true.

  - If *p-value* is less than the α-level (typically 0.05) chosen prior to the study, then the null hypothesis is rejected.
  - Commonly misinterpreted as the probability that the null hypothesis is true (can never accept the null)

Concepts
What is a confidence interval?

- confidence interval: 95% confident that the interval from *lower* to *upper* actually contains the true population value.
  - A narrow confidence interval implies high precision
  - A wide interval implies poor precision
  - The degree of confidence is typically set at 90% or 95%

A Note on Hazard Ratios

- Hazard: risk of experiencing the event of interest at any time point

- Hazard ratio (HR): ratio of the hazard in one treatment group divided by the hazard in the second treatment group.

- An HR of 1 means that the risk of experiencing the event is similar for both treatment groups.
  -If the confidence interval for the HR crosses 1, no difference in hazards was observed.

- Calculating the HR is appropriate when the hazards are proportional (e.g. two curves that do not cross).