Logic respecting efficacy measures in the presence of prognostic or predictive biomarker subgroups

Yi Liu and the Oncology Estimand working group TF8

JSM 2021

Aug 12, 2021

Acknowledgement

Oncology Estimand TF8

- Yi Liu (Nektar, Lead)
- Yue Shentu (Merck)
- Miao Yang (Nektar)
- Shoubhik Mondal (BI)
- Hong Tian (Beigene)
- Liwei Wang (J&J)
- Siyoen Kil (LSK global PS)
- Jiang Li (Beigene)
- Godwin Yung (Genentech)
- Jonathan Siegel (Bayer)

Additional collaborators

- Jason Hsu
- Ying Ding
- Hui-Min Lin
- Szu-Yu Tang
- Bushi Wang
- Haiyan Xu

Outline

- Puzzling behavior of HR in real Clinical trials with subgroups
  - HR can make a purely prognostic biomarker seem predictive
- Two issues:
  - Efficacy measure such as HR and OR are not logic respecting and non-collapsible at the population level
  - Current computer software and common analysis methods help mask the problem
- Our proposal: Follow Subgroup Mixable Estimation (SME) in analyzing clinical trial results for logic respecting efficacy measures
  - Shiny app available to produce simultaneous CIs for subgroups and overall population

*https://jchsustatsci.shinyapps.io/Ratio_of_Median_survival_times

nature medicine

Article | Published: 06 August 2018

Blood-based tumor mutational burden as a predictor of clinical benefit in non-small-cell lung cancer patients treated with atezolizumab

David R. Gandara, Sarah M. Paul, Marcin Kowanetz, Erica Schleifman, Wei Zou, Yan Li, Achim Rittmeyer, Louis Fehrenbacher, Geoff Otto, Christine Malboeuf, Daniel S. Lieber, Doron Lipson, Jacob Silterra, Lukas Amler, Todd Riehl, Craig A. Cummings, Priti S. Hegde, Alan Sandler, Marcus Ballinger, David Fabrizio, Tony Mok & David S. Shames

Nature Medicine **24**, 1441–1448 (2018) | Download Citation

- POPLAR data demonstrated proof of principle for bTMB as a predictor of PFS clinical outcome

Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial

*Louis Fehrenbacher, Alexander Spira, Marcus Ballinger, Marcin Kowanetz, Johan Vansteenkiste, Julien Mazieres, Keunchil Park, David Smith, Angel Artal-Cortes, Conrad Lewanski, Fadi Braiteh, Daniel Waterkamp, Pei He, Wei Zou, Daniel S Chen, Jing Yi, Alan Sandler, Achim Rittmeyer, for the POPLAR Study Group**

Background Outcomes are poor for patients with previously treated, advanced or metastatic non-small-cell lung cancer *Lancet* 2016; 387: 1837-46

- OAK data confirm bTMB as a potential non-invasive biomarker of PD-L1-directed immunotherapy.

Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial

*Achim Rittmeyer, Fabrice Barlesi, Daniel Waterkamp, Keunchil Park, Fortunato Ciardiello, Joachim von Pawel, Shirish M Gadgeel, Toyoaki Hida, Dariusz M Kowalski, Manuel Cobo Dols, Diego L Cortinovis, Joseph Leach, Jonathan Polikoff, Carlos Barrios, Fairooz Kabbinavar, Osvaldo Arén Frontera, Filippo De Marinis, Hande Turna, Jong-Seok Lee, Marcus Ballinger, Marcin Kowanetz, Pei He, Daniel S Chen, Alan Sandler, David R Gandara, for the OAK Study Group**

Summary

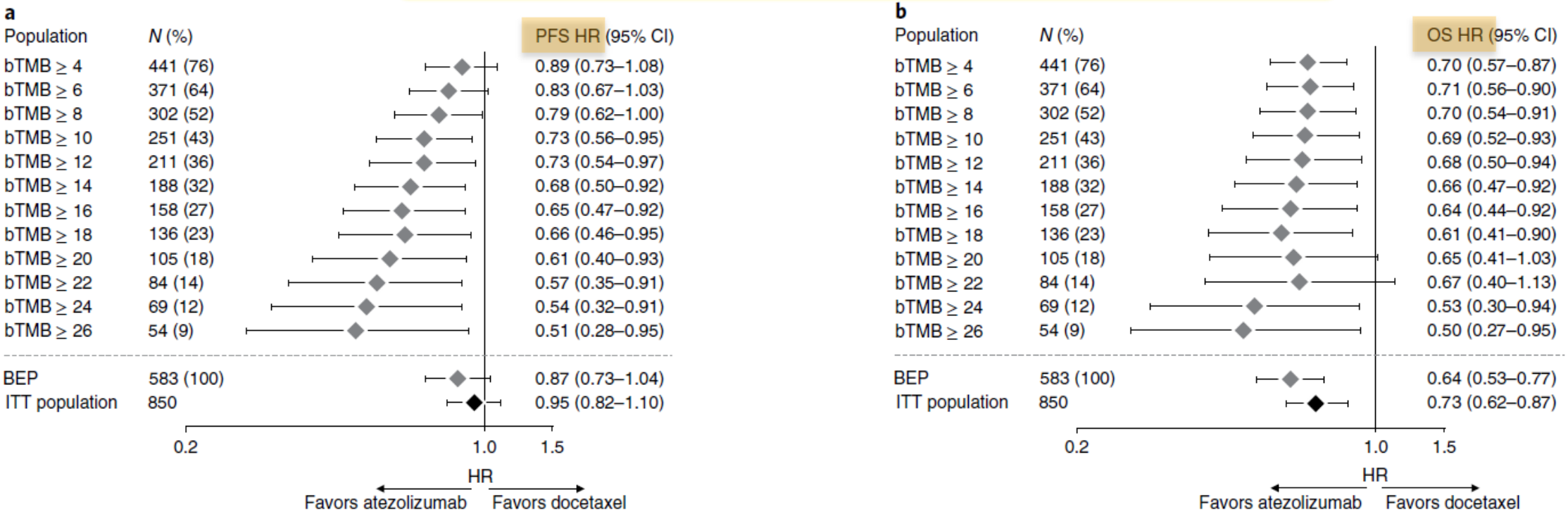
Background Atezolizumab is a humanised antiprogrammed death-ligand 1 (PD-L1) monoclonal antibody that *Lancet* 2017; 389: 255-65

Is bTMB a predictor of clinical benefit in NSCLC patients treated with atezolizumab in OAK study?

cut-points of bTMB in the OAK study. Overall, there was a clear monotonic relationship between an increasing bTMB score and PFS outcomes (Fig. 4a). A similar, although less compelling, monotonic trend was observed for OS (Fig. 4b). Unlike PFS, numerical

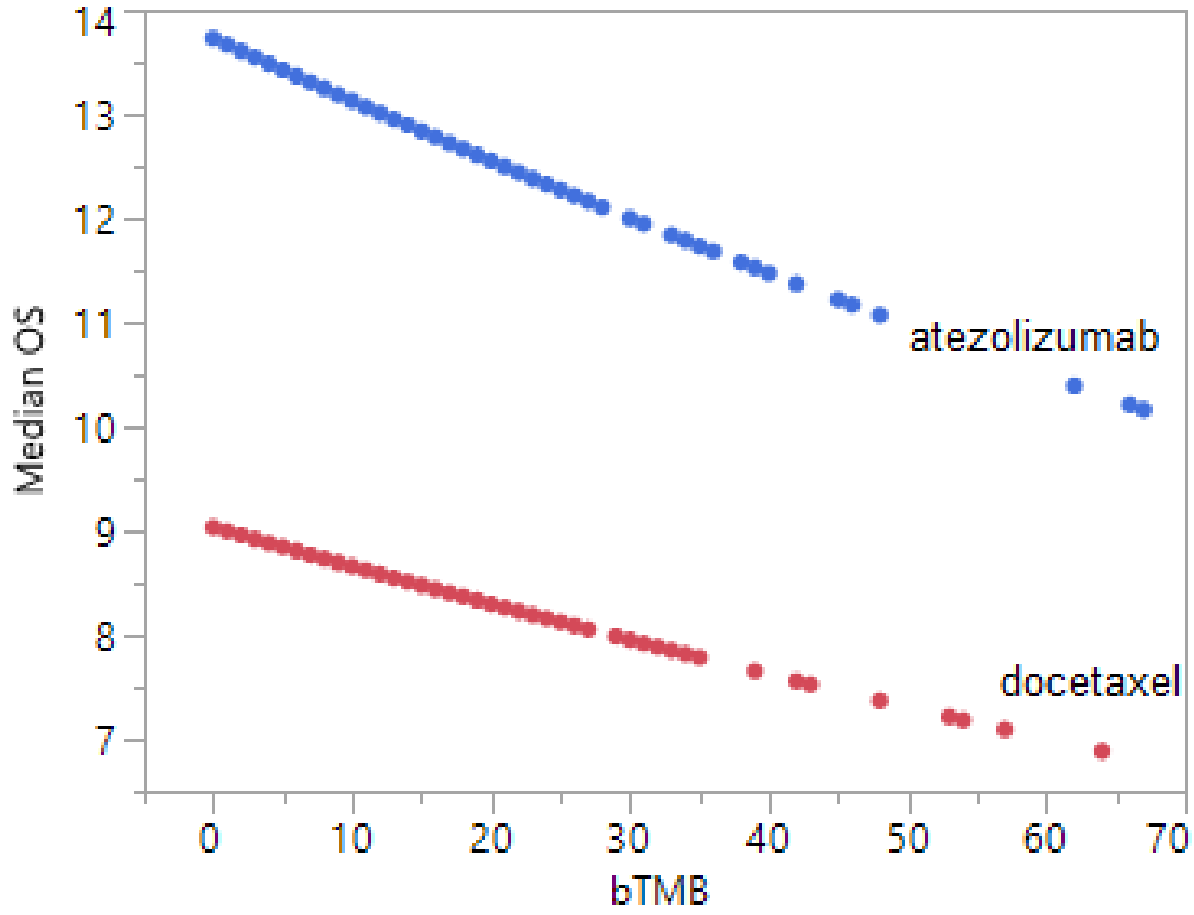
PFS

OS

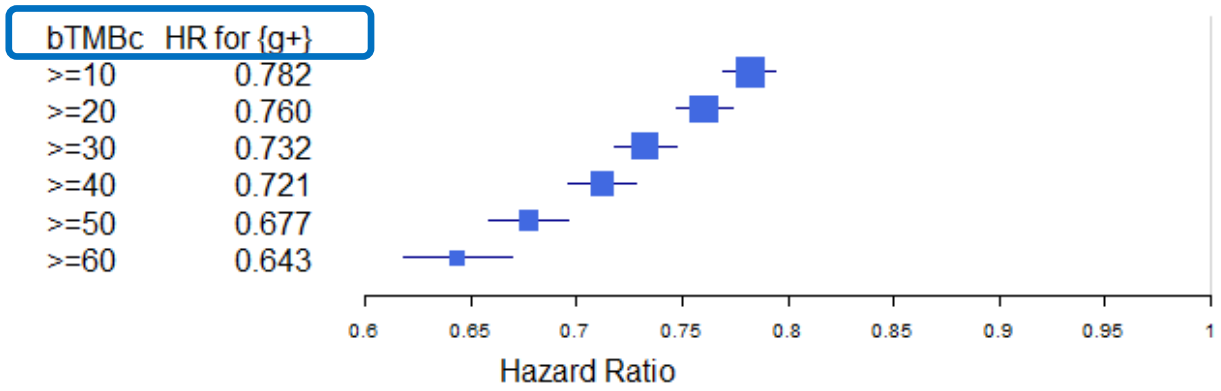


Rerun of the OAK trial data shows that bTMB is mostly a prognostic (instead of predictive) biomarker in terms of OS

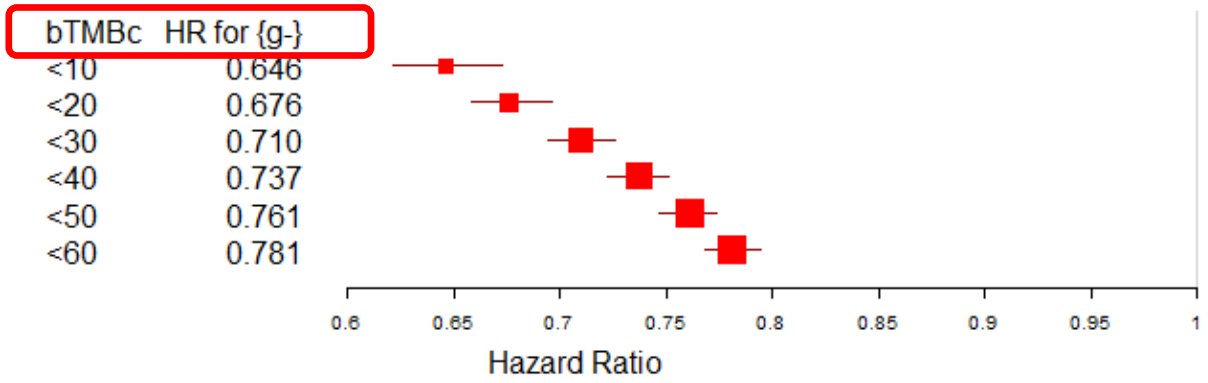
Estimated median OS from Weibull fit with bTMB, Trt and interaction terms



HR behavior for purely prognostic biomarker based on simulation



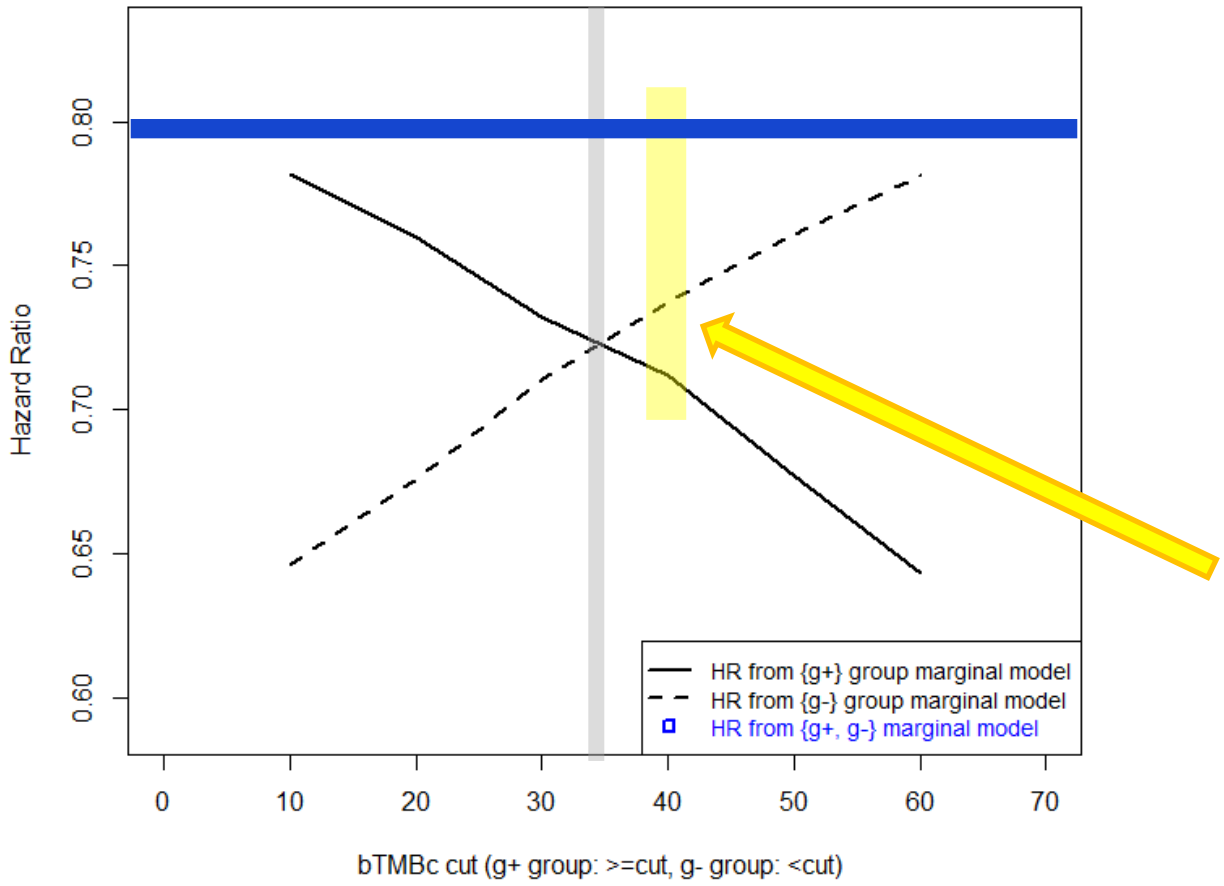
Replicated the pattern observed in OAK trial



Conflicting message in terms which pt subgroup benefits most

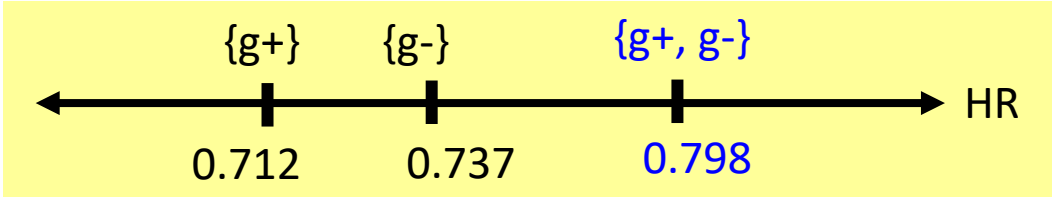
Per disjoint biomarker subgroup, generated 10,000 (total 70,000) time-to-event random variable that follows Weibull distribution. Simulated data present purely prognostic biomarker that is the treatment effects (effect size, HR) are same throughout the disjoint biomarker subgroup with increasing baseline hazard.

HR behavior for purely prognostic biomarker based on simulation



For any cut point of the bTMBc value, the marginal HR for whole data {g+, g-} is always outside range of the HRs of bTMBc subgroups.

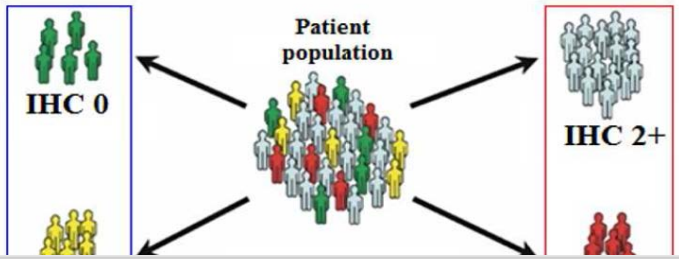
Ex) bTMBc cut = 40



Marginal HR: HR for the overall population using Trt as only covariate in the cox model

Clinical Trials with two subgroups where HR is not logic respecting

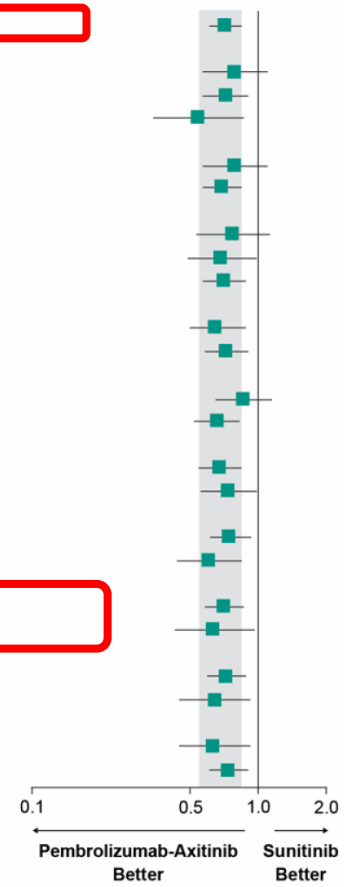
MET study: Ph2 NSCLC[1]



HR=0.47

KN-426: Ph3 RCC PFS[3]

	Pembrolizumab + Axitinib N/No. Events	Sunitinib N/No. Events	HR (95% CI)
Overall	**432/264**	**429/281**	0.71 (0.62-0.84)
IMDC category 1			
Favorable	138/77	131/75	0.79 (0.57-1.09)
Intermediate	238/145	246/163	0.72 (0.57-0.90)
Poor	56/42	52/43	0.54 (0.34-0.86)
IMDC risk category 2			
Favorable	138/77	131/75	0.79 (0.57-1.09)
Intermediate/Poor	294/187	298/206	0.69 (0.56-0.84)
			0.77 (0.53-1.13)
			0.68 (0.48-0.97)
			0.70 (0.56-0.87)
			0.65 (0.49-0.87)
			0.72 (0.58-0.89)
			0.86 (0.64-1.15)
			0.66 (0.52-0.82)
			0.67 (0.54-0.84)
			0.74 (0.55-0.99)
			0.74 (0.60-0.91)
			0.60 (0.43-0.84)
			0.70 (0.58-0.85)
			0.63 (0.42-0.95)
			0.72 (0.59-0.87)
			70.64 (0.44-0.92)
			0.63 (0.44-0.91)
			0.74 (0.61-0.90)



	Nivolumab plus ipilimumab with chemotherapy (two cycles)		Chemotherapy		Unstratified hazard ratio for death (95% CI)
	Events/ patients	Median overall survival, months (95% CI)	Events/ patients	Median overall survival, months (95% CI)	
PD-L1 expression subgroups					
<1%	69/135	16·8 (13·7-NR)	89/129	9·8 (7·7-13·7)	0·62 (0·45-0·85)
≥1%	105/203	15·8 (13·8-NR)	139/204	10·9 (9·5-13·2)	0·64 (0·50-0·82)
1-49%	68/127	15·4 (12·6-NR)	78/106	10·4 (8·7-12·4)	0·61 (0·44-0·84)
≥50%	37/76	18·0 (13·1-NR)	61/98	12·6 (9·4-16·9)	0·66 (0·44-0·99)
All randomly assigned patients	**190/361**	**15·6 (13·9-20·0)**	**242/358**	**10·9 (9·5-12·6)**	**0·66 (0·55-0·80)***

CM-9LA: Ph3 NSCLC OS[2]

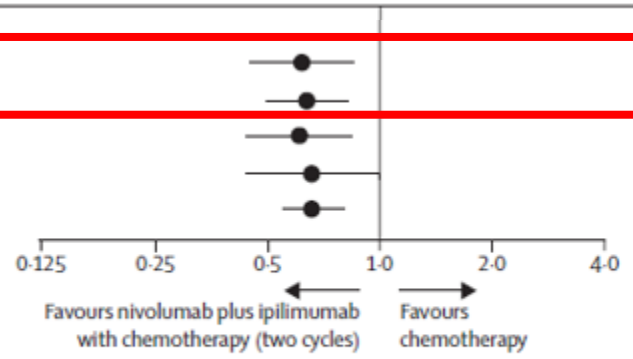
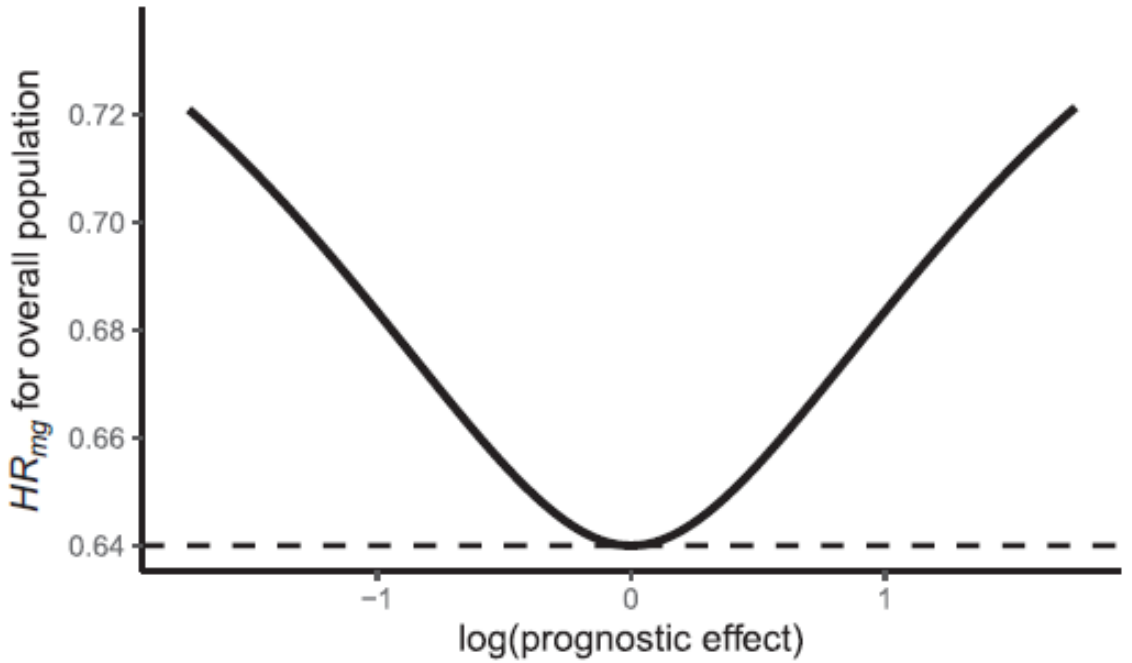


Figure 3: **Forest plot of overall survival based on longer follow-up in predefined patient subgroups**
ECOG=Eastern Cooperative Oncology Group. NR=not reached. *Stratified hazard ratio. Unstratified hazard ratio was 0·67 (95% CI 0·55–0·81).

1.Spigel et. al. (2013). 2. Paz-Ares et. Al. (2021); 3. Powles et. Al. (2020)

HR for overall population trending toward 1 as prognostic effect increases



HR for two subgroups are both set at 0.64 with 50% prevalence; prognostic effect is the HR between g+ and g-; HR_mg is calculated as HR from the cox model with Trt as the only covariate – even though the theoretical HR for overall pop depends on time when prognostic effect is present; HR_mg is viewed as average HR (Xu and O'Quigley 2000)

Our proposal

- Current literature* still focusing on how to “fix” HR
  - Not to compare marginal vs conditional HRs** as they are like apple and oranges
  - Advocating the use of conditional HR over marginal HR
  - how to derive a more efficient marginal HR based on conditional models

We propose to replace HR with alternative efficacy measures that follow:

Efficacy measure for the overall population should always be in between the efficacy in the subgroups at both population and sample level

- For efficacy measures that respects this logic in the population space, they are called *logic respecting efficacy measure*
- But even for logic respecting efficacy measure, if incorrect analysis methods are used, illogical behavior can still be observed in the sample space
  - Solution: follow SME (Ding et al 2016; Lin et al 2019)

*Daniel et al (2021)

**Conditional HR: subgroup HR, i.e. HR conditioning on the subgroup

With SME** one does not have to choose

Apples grow on apple trees

Conditioning doesn't make apples oranges

Subgroup Mixable Estimation makes marginal and conditional logical***

* For logic respecting efficacy measures

Assessing apples as oranges makes no sense

**Ding et al (2016); Lin et al (2019)

Logic respecting efficacy measures for all endpoint types

In population space:

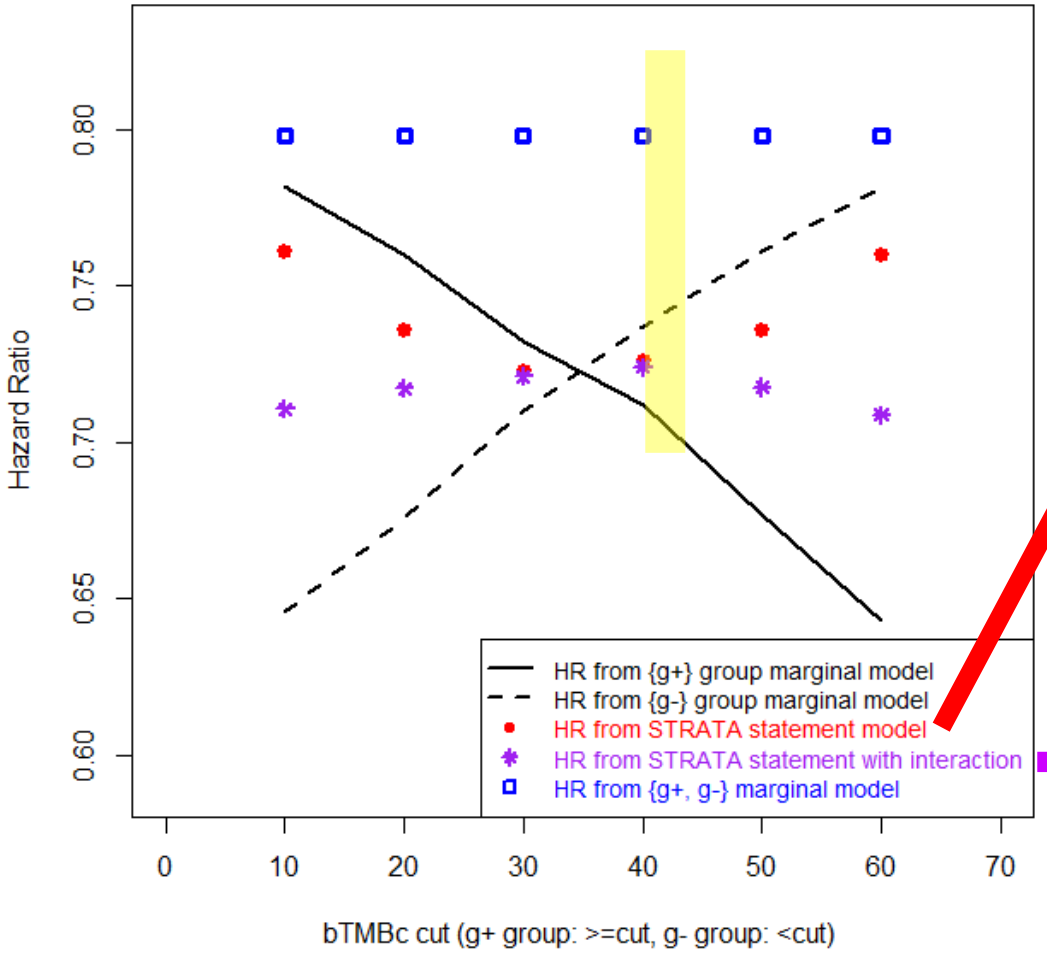
Endpoint type	Efficacy measure	Logic-respecting efficacy measure?
Continuous	Diff of means	Yes
Binary	Diff of props	Yes
	Relative risk (RR)	Yes
	Odds ratio (OR)	**No**
Time-to-event (TTE)	**HR**	**No**
	Diff of medians	No
	Ratio of medians (RoM)	Yes*
	Diff of RMST/milestone prob	Yes
	Ratio of RMST/milestone prob	Yes

In sample space:

- Incorrect analysis methods are currently implemented that can lead to
  - the masking of illogical behavior for non-logic-respecting efficacy measures
  - illogical behavior for logic-respecting efficacy measures - marginal analysis

* When there is proportional hazards within each subgroup under Weibull model

Incorrect estimate of marginal HR in SAS LSMEANS that masks illogical behavior of HR




```
PROC PHREG DATA=DA2;
CLASS TRT01P(REF="CTL") BTMB40(REF="g-") /PARAM=GLM;
MODEL OS*OSCNSR(1)=TRT01P;
STRATA BTMB40;
HAZARDRATIO 'H1' TRT01P/DIFF=ALL CL=BOTH;
LSMEANS TRT01P;
RUN;
```

$$\sim \gamma^{+}(HR_{+}) + \gamma^{-}(HR_{-})$$

```
PROC PHREG DATA=DA2;
CLASS TRT01P(REF="CTL") BTMB40(REF="g-")
/PARAM=GLM;
MODEL OS*OSCNSR(1)=TRT01P BTMB40
TRT01P*BTMB40;
STRATA BTMB40;
HAZARDRATIO 'H1' TRT01P/DIFF=ALL CL=BOTH;
LSMEANS TRT01P;
RUN;
```

$$exp\left\{\frac{1}{2}(logHR_{+}) + \frac{1}{2}(logHR_{-})\right\}$$

Marginal model estimates can lead to illogical behavior even for logic respecting efficacy measure

Consider following two models to estimate difference of means (DoM): $\theta = E(Y_i|T_i = Rx) - E(Y_i|T_i = C)$

Conditional model : $Y_i = \mu + \alpha T_i + \beta G_i + \delta T_i G_i + \varepsilon_i$ **with** $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Marginal model : $Y_i = \mu^* + \alpha^* T_i + \varepsilon_i$ **with** $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Mix within each Rx and C first

- DoM estimator from conditional model is obtained by the following
  - $\hat{\theta}_c = \left[\hat{E}(Y_i|T_i = Rx, G_i = g^+)\gamma^+ + \hat{E}(Y_i|T_i = Rx, G_i = g^-)\gamma^-\right] - \left[\hat{E}(Y_i|T_i = C, G_i = g^+)\gamma^+ + \hat{E}(Y_i|T_i = C, G_i = g^-)\gamma^-\right]$
  - $= [\hat{E}(Y_i|T_i = Rx, G_i = g^+) - \hat{E}(Y_i|T_i = C, G_i = g^+)]\gamma^+ + [\hat{E}(Y_i|T_i = Rx, G_i = g^-) - \hat{E}(Y_i|T_i = C, G_i = g^-)]\gamma^-$

Special for DoM

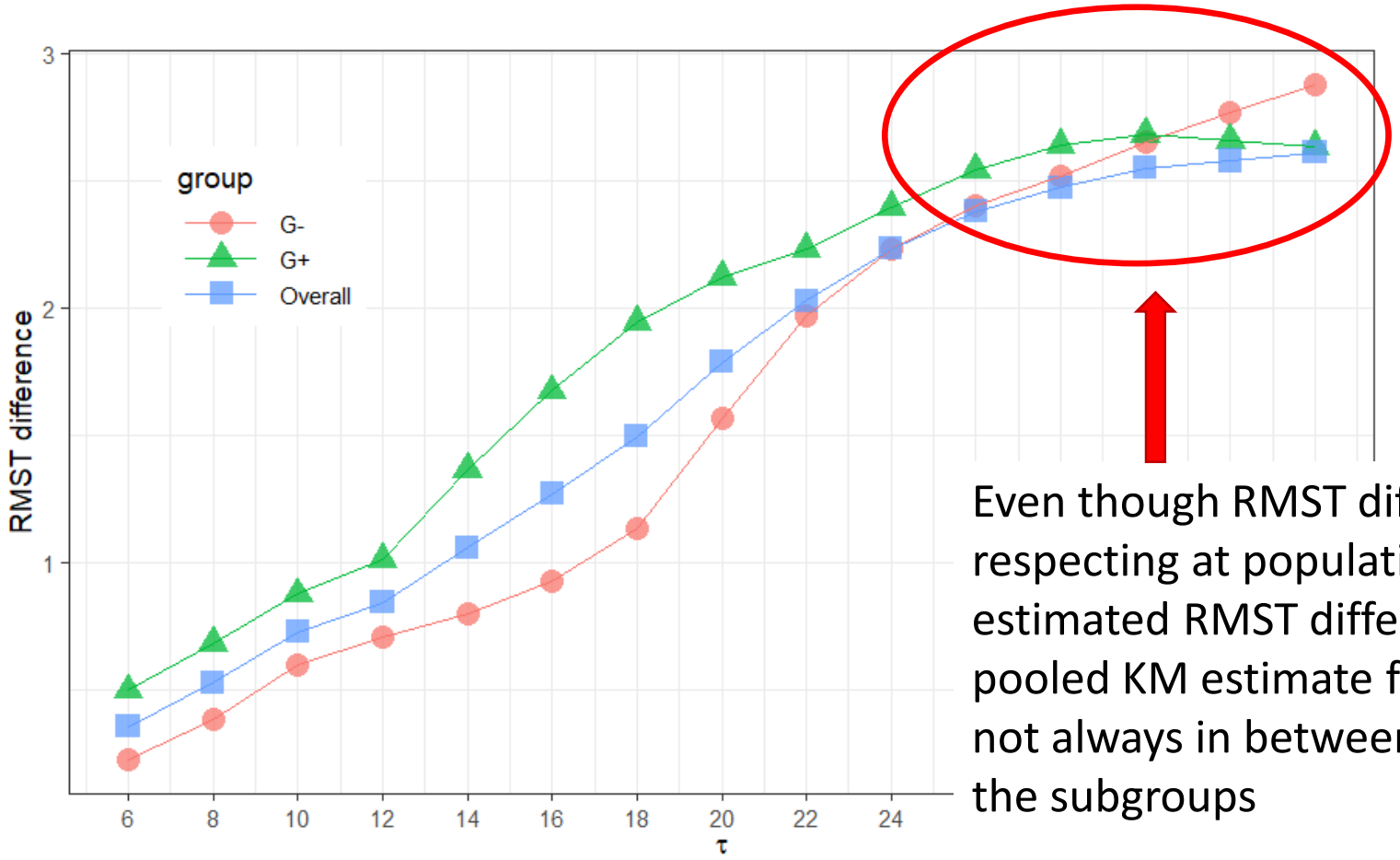
$\hat{\theta}_{g^+}$ $\hat{\theta}_{g^-}$

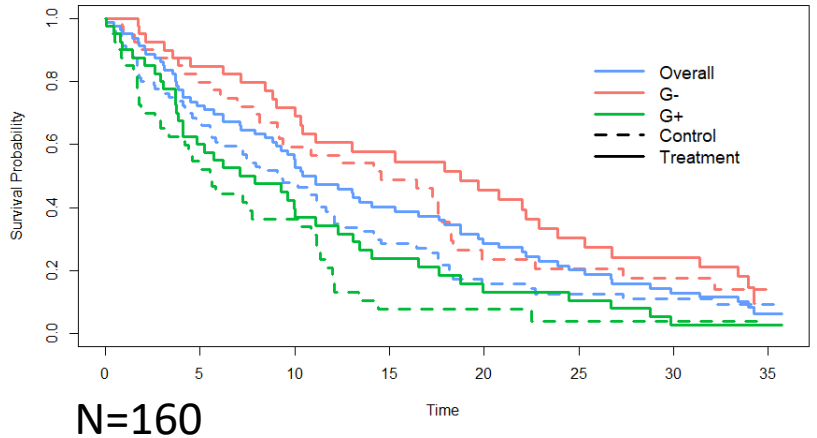
- DoM estimator from marginal model is $\hat{\theta}_m = \hat{\alpha}^*$
- Known in the literature: $\hat{\theta}_c$ is more efficient than $\hat{\theta}_m$ regardless of the underlying true model
  - $Var(\hat{\theta}_c) \leq Var(\hat{\theta}_m)$ under conditional model
  - $Var(\hat{\theta}_c) \approx Var(\hat{\theta}_m)$ under marginal model
- Additionally: $\hat{\theta}_c$ is always logical while illogical behavior exists for $\hat{\theta}_m$ regardless of which model is true

Under $\gamma^+ = 1/3$, 1:1 allocation, 10,000 simulations,
% illogical behavior $\hat{\theta}_m \notin [\hat{\theta}_{g^-}, \hat{\theta}_{g^+}]$

True model	N=24	N=120
Conditional model $\mu = 0, \alpha = 1, \beta = 2, \delta = 3, \sigma = 1$	8.8%	0.1%
Marginal model $\mu^* = 0, \alpha^* = 1, \sigma = 1$	9.8%	5.3%

RMST difference based on marginal KM curves may disrespect logic





N=160

Even though RMST difference is logic respecting at population level, estimated RMST difference by the pooled KM estimate for Rx and C is not always in between those from the subgroups

Correct analysis methods for logic respecting efficacy measures for all endpoint types
Principle of Subgroup Mixable Estimation (SME)

1. Fit a model (e.g. linear, logistic, log-linear, or weibull) to get LS estimates of main effects + interactions and associated variance-covariance matrix estimates
2. Convert to estimates of Rx and C effect for g+ and g- and overall pop and estimate the corresponding var-cov matrix using δ-method
   - To get estimates of Rx and C effect for overall pop: mix within *Rx* and *C* on the probability scale by population or pooled sample prevalence
3. Calculate estimates of efficacy (Rx vs C) in g+ and g- and overall pop and the corresponding var-cov matrix using δ-method
4. Calculate simultaneous CIs for efficacy in subgroups and overall pop based on Normal approximation based on δ-method

Applying SME to Checkmate-9LA OS

HR for overall: 0.67

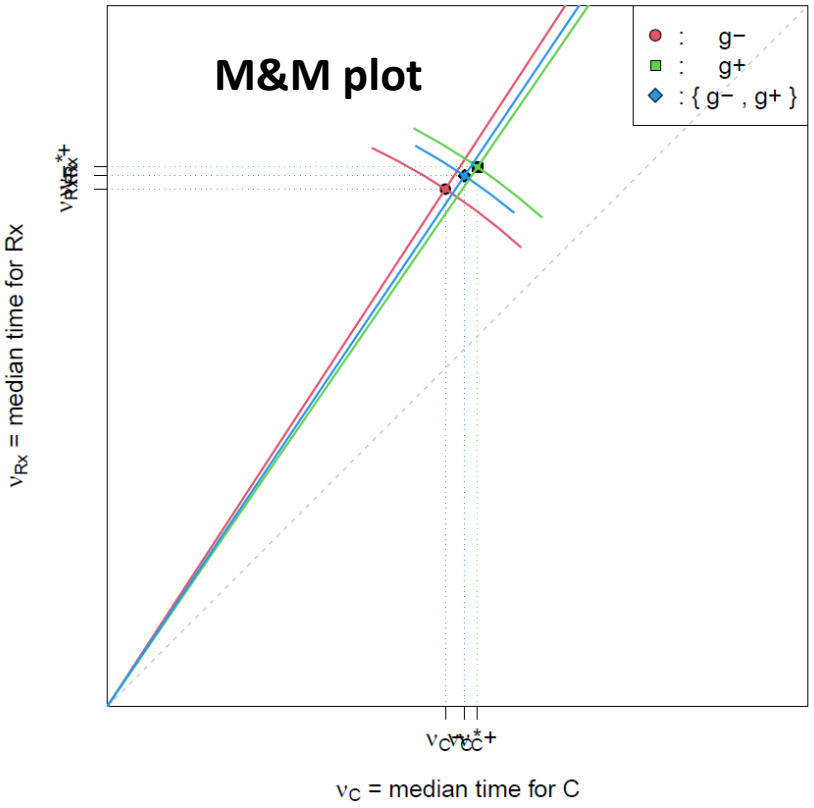
Fit digitized data to the following Weibull model:

$$h(t|arm, group) = h_0(t)exp\{\beta_1 arm + \beta_2 group + \beta_3 arm \times group\}$$

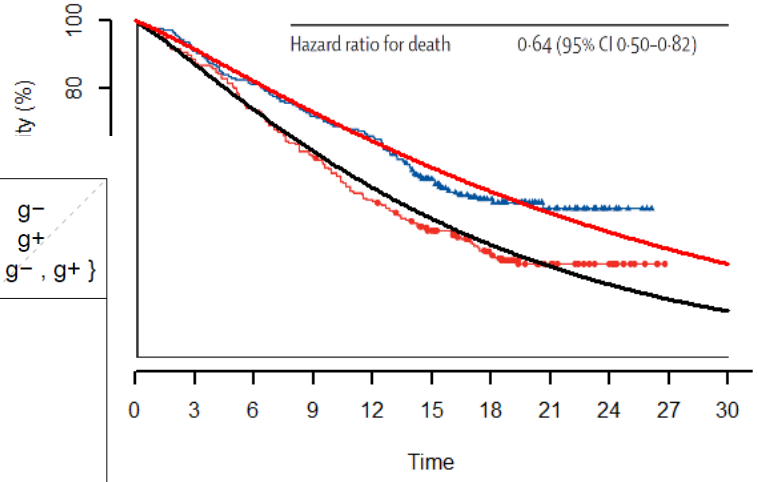
where $h_0(t) = \kappa\lambda^{\kappa}t^{\kappa-1}$

95% simultaneous CIs for RoM (right) and ratio/difference of RMST and 1-year OS rate (below)

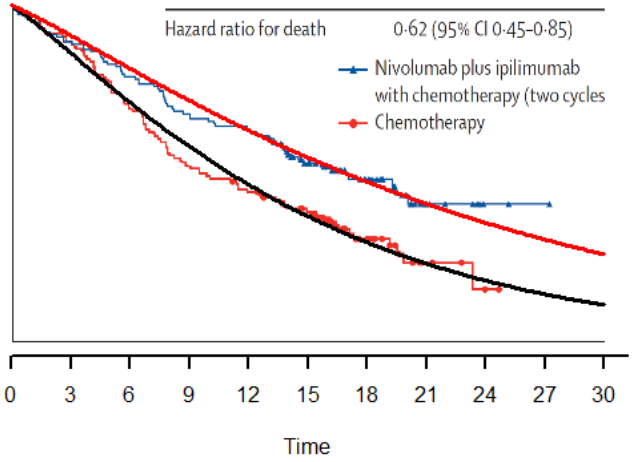
Efficacy Measure	Group	Weibull model	
		Ratio	Difference
RMST	PD-L1-	1.264 (1.06,1.508)	3.276 (0.849,5.703)
	PD-L1+	1.217 (1.064,1.391)	2.842 (0.917,4.767)
	Overall	**1.235 (1.11,1.374)**	**3.013 (1.504,4.521)**
1-year survival rate	PD-L1-	1.344 (1.071,1.686)	0.161 (0.042,0.281)
	PD-L1+	1.273 (1.077,1.505)	0.138 (0.044,0.232)
	Overall	**1.3 (1.136,1.488)**	**0.147 (0.073,0.221)**



PD-L1+ (203 vs. 204)



PD-L1- (135 vs. 129)



Summary

- Using non-logic respecting efficacy measures such as HR can potentially harm patients due to incorrect treatment benefit assessment
- Explaining to clinicians that "*HR in the overall pop and HR in the subgroups are apples and oranges and should not be compared*" is not the right message

Our recommendation:

- Summarize clinical trial results with logic respecting efficacy measure such as RoM as first step
- Then use SME to correctly analyze clinical trial results for logic respecting efficacy measures to guarantee logical behavior even with limited sample size

References

- Liu, Y, Wang, B, Yang, M, Hui, J, Xu, H, Kil, S, Hsu, JC. Correct and logical causal inference for binary and time-to-event outcomes in randomized controlled trials. Biometrical Journal. 2021; 1– 27. https://doi.org/10.1002/bimj.202000202
- Rubin, D. B. (1978). *Annals* of *Statistics* 6, 34–58.
- Holland, P. (1986). *J*. *Amer*. *Statist*. *Assoc*. 81, 945–970.
- Huitfeldt, A., Stensrud, M.J. & Suzuki, E. On the collapsibility of measures of effect in the counterfactual causal framework. Emerg Themes Epidemiol 16, 1 (2019). https://doi.org/10.1186/s12982-018-0083-9
- Greenland, Sander, James M. Robins, and Judea Pearl (1999). Confounding and Collapsibility in Causal Inference. *Statistical Science* 14, 29–46.
- Ding, Ying, Hui-Min Lin, and Jason C. Hsu (2016). Subgroup Mixable Inference on treatment efficacy in mixture populations, with an application to time-to-event outcomes. *Statistics in Medicine* 35, 1580–1594.
- Lin, Hui-Min, Haiyan Xu, Ying Ding, and Jason C. Hsu (2019). Correct and Logical Inference on efficacy in subgroups and their mixture for binary outcomes. *Biometrical Journal* 61, 8–26.
- Ding, Peng and Fan Li (2018). *Statistical Science* 33, 214–237.
- Gandara et al (2018). Blood-based tumor mutational burden as predictor … NSCLC … atezolizumab. *Nature Medicine* 24, 1441–1448.
- Spigel et. al. (2013). Randomized phase II trial of onartuzumab in combination with erlotinib in patients with advanced non-small-cell lung cancer. *Journal of Clinical Oncology:* 31(32): 4105-4114.
- Paz-Ares, L., Ciuleanu, T.E., Cobo, M., Schenker, M., Zurawski, B., Menezes, J., Richardet, E., Bennouna, J., Felip, E., Juan-Vidal, O. and Alexandru, A., 2021. First-line nivolumab plus ipilimumab combined with two cycles of chemotherapy in patients with non-small-cell lung cancer (CheckMate 9LA): an international, randomised, open-label, phase 3 trial. The Lancet Oncology, 22(2), pp.198-211.
- Powles, T., Plimack, E.R., Soulières, D., Waddell, T., Stus, V., Gafanov, R., Nosov, D., Pouliot, F., Melichar, B., Vynnychenko, I. and Azevedo, S.J., 2020. Pembrolizumab plus axitinib versus sunitinib monotherapy as first-line treatment of advanced renal cell carcinoma (KEYNOTE-426): extended follow-up from a randomised, open-label, phase 3 trial. *The Lancet Oncology*, *21*(12), pp.1563-1573.
- Daniel, R, Zhang, J, Farewell, D. Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal*. 2021; 63: 528– 557. https://doi.org/10.1002/bimj.201900297