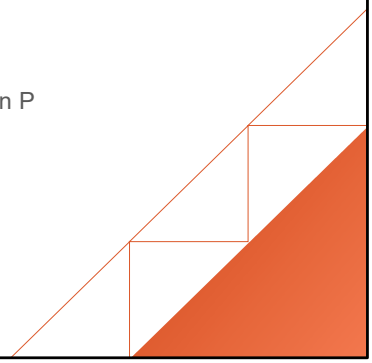
A case study to compare different indirect treatment comparison methods under varying access to individual patient data

ISPOR presentation

November 22nd, 2021

Authors: Julie E Park, Dieter Ayers, Shannon Cope, David M Phillippo, Jeroen P Jansen, Sisi Wang, Yong Yuan

Acknowledgement: Kevin M Towle



Hello everyone. Thank you for inviting me to present our work on a case study to compare different indirect treatment comparison methods under varying access to individual patient data

The need for unanchored indirect comparisons given evidence base

- In some cases, phase II single-arm studies are acceptable to support regulatory approval and reimbursement of new interventions while individual patient data (IPD) from real-world evidence studies may also be available regarding standard of care that could serve as historical control
- **Population-adjusted indirect comparisons methods are often used** for IPD-AD analysis to adjust for between-study differences in prognostic factors and treatment effect modifiers, based on recommendations by Decision Support Unit (DSU) Technical Support Document (TSD) related to National Institute for Health and Care Excellence (NICE)
- For IPD-IPD comparisons, DSU TSD 17 provides recommendations on alternative **methods for comparative IPD** from different studies for the intervention and comparator, which include, inverse probability weighting (IPW), regression adjustment (RA), and doubly robust (IPW+RA) methods
- There is interest in assessing alternative indirect comparison methods, depending on the data available from the comparator study

Abbreviations: *AD, aggregate data; IPD, individual-level patient data;* ***References:*** *1. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submission to NICE. 2016 2. Faria R, Alava MH, Manca A, Wailoo AJ. NICE DSU Technical Support Document 17: The use of observational data to inform estimates of treatment effectiveness in technology appraisal: methods for comparative individual patient data. 2015*

2

There is unmet need for unanchored indirect comparisons given evidence base

When a new drug is developed, individual patient data (IPD) is typically available from the manufacturer's clinical trial. However, in most cases, only aggregate (AD) are available for other interventions used in the standard of care.
Recently, IPD are collected for patients who receive SOC to serve as historical controls.

Therefore, unanchored indirect comparisons have been done for IPD-AD and there are recommended methods for IPD-IPD comparisons to adjust for between study differences, **which is crucial given the randomization design does not hold anymore.**

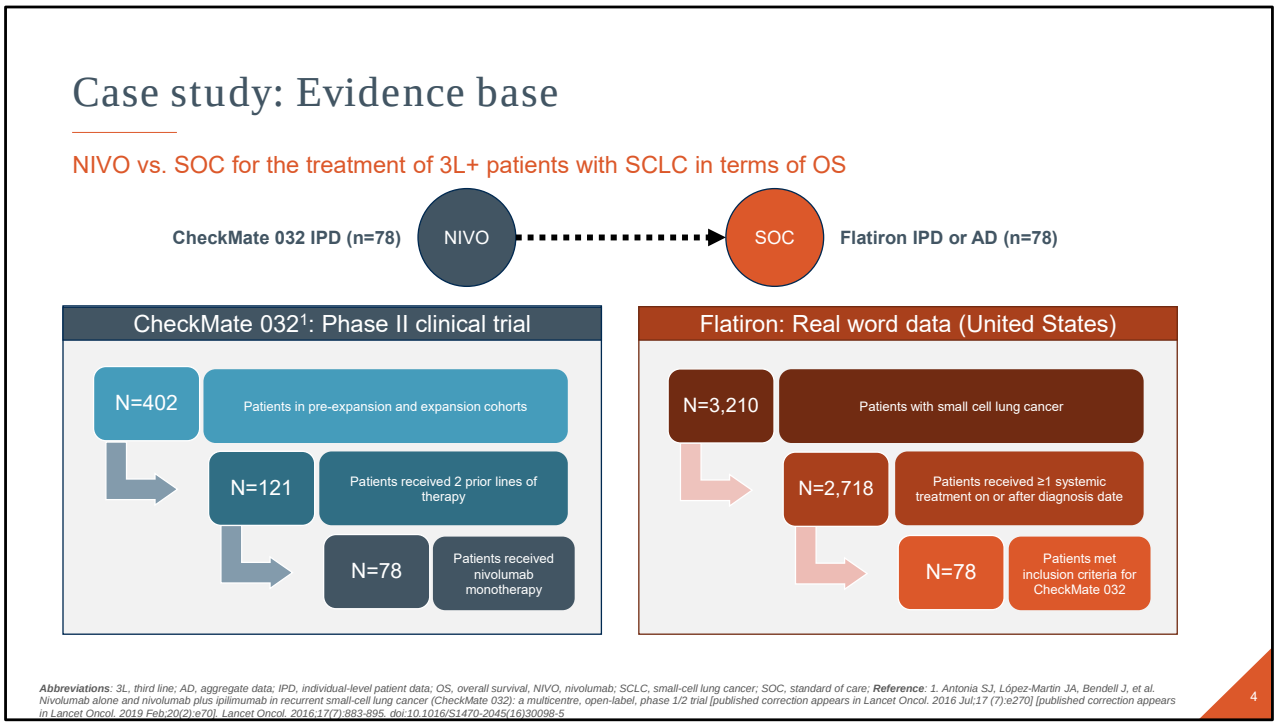
And there's interest in assessing the performance of these methods.

Objective

- To compare alternative methods for unanchored indirect comparisons of two interventions for overall survival (OS) when:
  - Individual patient data (IPD) are available for both the intervention and comparator studies (IPD-IPD analyses)
  - IPD are available for intervention study and aggregate data (AD) for the comparator study (IPD-AD analyses)

3

So our objective was to compare alternative methods for unanchored indirect comparisons of two interventions for overall survival (OS) when individual patient data (IPD) are available for both studies or a single study.



First, I want to talk about the data used for our case study analysis.
We evaluated the comparative efficacy of nivolumab versus standard of care for the treatment of 3rd line patients with small cell lung cancer.
CheckMate 032 is a Phase II clinical trial. We restricted the analysis to patients who treated in the third line with nivolumab monotherapy.
Flatrion data reflects real word data from the US, patient who met **inclusion/exclusion criteria of CheckMate 032 were selected.**
In total, there were 78 patients included from each study.

Case study: Patient characteristics

Slightly more extensive disease, platinum resistant, ECOG status = 0 in CheckMate 032

Characteristic (inclusion in model)	Measure	CheckMate 032 (N=78)	Flatiron (N=78)
Age (included)	Mean (SD)	63.1 (8.16)	64 (10)
	Median (range)	64 (45-81)	64 (34-84)
Sex (included)	Male	41 (52.6%)	38 (48.7%)
	Female	37 (47.4%)	40 (51.3%)
Race (included)	Caucasian	72 (92.3%)	54 (69.2%)
	Black or African American	4 (5.1%)	5 (6.4%)
	Asian	1 (1.3%)	2 (2.6%)
	American Indian or Alaska Native	0 (0%)	--
	Other	1 (1.3%)	6 (7.7%)
	Missing	0 (0%)	11 (14.1%)
Smoking status (excluded)	Current/former	73 (93.6%)	77 (98.7%)
	Never smoked	5 (6.4%)	1 (1.3%)
	Unknown	0 (0%)	0 (0%)
Disease stage (included)	Limited	21 (26.9%)	22 (28.2%)
	Extensive	57 (73.1%)	53 (67.9%)
	Unknown	0 (0%)	3 (3.8%)
Platinum sensitivity in first-line (included)	Platinum resistant	26 (33.3%)	20 (25.6%)
	Platinum sensitive	52 (66.7%)	54 (69.2%)
	Unknown	0 (0%)	4 (5.1%)
ECOG performance status (excluded)	0	25 (32.1%)	17 (21.8%)
	1	52 (66.7%)	28 (35.9%)
	≥2	1 (1.3%)	6 (7.7%)
	Not reported	0 (0%)	27 (34.6%)

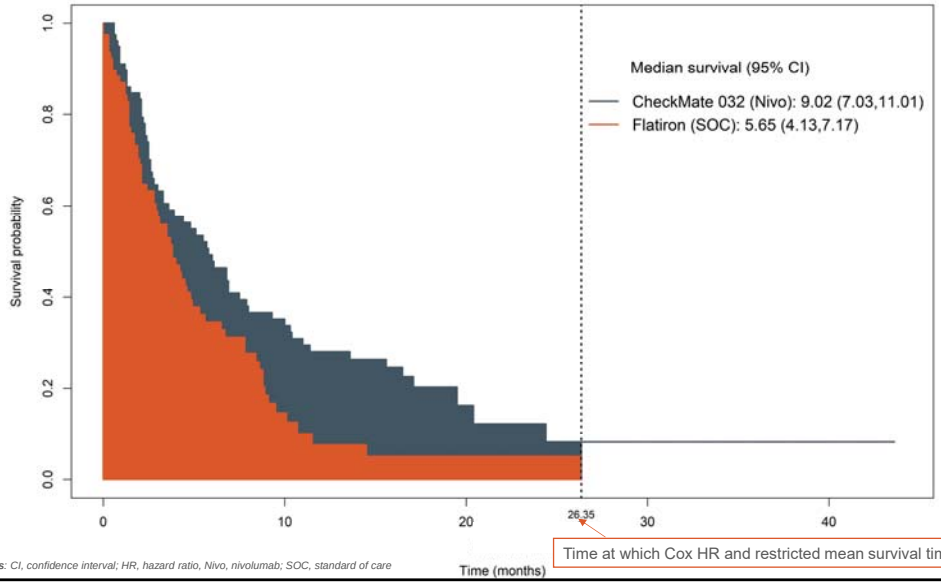
Abbreviations: *ECOG, Eastern Cooperative Oncology Group performance status; SCLC, small-cell lung cancer*

Notes: *In the Flatiron dataset, 11 values were missing for race, 3 values were missing for extended disease, and 4 values were missing for platinum sensitivity. Mean imputation was performed for these missing values (missing values were replaced with the average obtained from the non-missing covariate values)*

5

These are the prognostic factors identified by the targeted literature review. Given the selection process, the two studies were reasonably similar although there were some differences in extensive disease, platinum sensitivity and ECOG performance. Five patient characteristics were included in all models (age, sex, race, disease stage, and platinum sensitivity to first-line treatment); two variables were not used in our analysis due to Flatiron data availability.

Case study: Overall survival

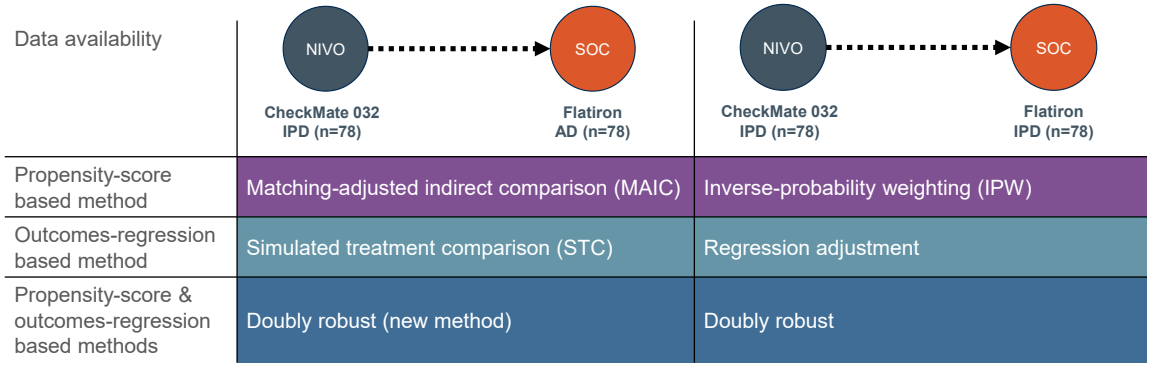


Abbreviations: CI, confidence interval; HR, hazard ratio, Nivo, nivolumab; SOC, standard of care

6

This figure summarizes the Kaplan Meier overall survival data from CheckMate 032 and Flatiron as well as the restricted mean survival time, which is the area under the survival curve, up to 26.35 months. This time point was selected as the shortest of the maximum follow-up time from either study.

Methods: Overview



- Treatment effects assessed with
  - **Hazard ratios (HRs) using Cox model**
  - Difference in restricted mean survival time (RMST) using generalized estimating equation (GEE)
- Target population is Flatiron

Abbreviations: *AD, aggregate data; IPD, individual-level patient data; NIVO, nivolumab; SOC, standard of care;*

7

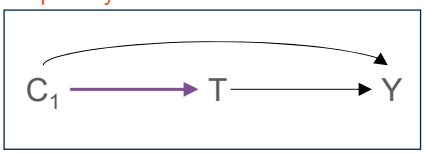
To give an overview of analysis performed,
first, we had two scenarios based on data availability of Flatiron population: either aggregate data (AD), or IPD.

We explored three different methods based on models that were: propensity-score based, outcome regression-based and doubly robust methods, **where the doubly robust method in IPD-AD setting is a new method.**
In terms of relative treatment effect estimated, we considered both the hazard ratio (HR) and difference in the restricted mean survival time, which is a good alternative to cox model especially when the proportional hazard assumption is questionable.

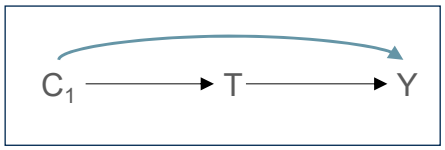
I want to point out that the Flatiron was the target population in MAIC, so we want to make it consistent across all methods. For weighting method, the average treatment in the control (ATC) weight was used. For regression-based methods, the simulated/observed Flatiron covariate data were used to the predicted outcomes which were then used to marginalize the relative treatment effect.

Methods: Motivation for doubly-robust models

Propensity score-based method



Regression adjustment



Y=Outcome (i.e. overall survival);
T=Treatment (i.e. NIVO vs. SOC);
C1=Binary patient characteristic (i.e. sensitive vs. resistant)

- Indirect comparisons are often based on propensity-score based methods *or* regression models
- Doubly robust (DR) estimators 'aim to reduce the impact of model misspecification by incorporating **both** outcome regression and propensity score models into one estimator, which is consistent if at least one of the constituent models is correct'[1]
- DSU TSD 17 IPD-IPD: Recommends DR methods if good overlap between two studies
- DSU TSD 18 IPD-AD: Recommends research regarding DR methods[2]
- DR studies focus on continuous outcomes, rather than time-to-event outcomes, partly because it is necessary to marginalize the HRs from regression adjustment, which produces conditional estimates

Abbreviations: AD, aggregate data; DSU, decision supporting unit; IPD, individual-level patient data; NIVO, nivolumab; SOC, standard of care; TSD, technical supporting document
References: 1. Faria R, Alava MH, Manca A, Wailoo AJ. NICE DSU Technical Support Document 17: The use of observational data to inform estimates of treatment effectiveness in technology appraisal: methods for comparative individual patient data. 2015 2. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submission to NICE. 2016

8

Before jumping into results, I'd like to dive a bit more into doubly robust method, a new method we explored in IPD-AD setting

Propensity score methods use weighted data to remove difference in covariate distributions between two studies and then assess the treatment effect. So they give marginal treatment effects.
And typical outcome regression models assess the treatment effect conditional on covariates.
And doubly robust method is designed to combines these two methods to reduce the model misspecification because it gives a correct estimate as long as either model is correctly specified.
Given this advantage, doubly robust models have been recommended for IPD-IPD analysis and also for IPD-AD analysis;
however, it hasn't been widely used in survival analysis because of an extra step involved: the conditional hazard ratio from regression model needs to be marginalized.

Methods: Estimation of variance

Cox Hazard Ratio (HR)

- Propensity score-based method
  - Observed data can be fit using Cox model to directly obtain HR and estimate of variance
  - Sandwich variance estimators were used to account for correlation introduced by weighting the data
- Regression adjustment
  - Regression models predict survival outcomes from which HRs are estimated
  - Bootstrap samples were used to estimate variance of HRs in order to marginalize HRs
- Doubly robust method
  - To align with regression adjustment, bootstrap samples were used to estimate variance of HRs
  - Therefore, we also used bootstrapping for propensity-based methods for comparative purposes

Abbreviations: *AD, aggregate data; IPD, individual-level patient data; HR, hazard ratio*

9

Another thing to consider is the estimation of variance

For propensity score based method, we can fit the model and directly obtain hazard ratio and its variance. Sandwich variance estimator is used to account for correlation introduced by weights

However, what about outcome regression models? A typical regression model is fitted to predict survival outcomes from which hazard ratios are estimated. Therefore, bootstrap samples are used to estimate variance.

For doubly robust method, bootstrap samples were used as well. However, unlike regression adjustment, weighted data were used in analysis, like propensity score based method

Results: Cox hazard ratio

Based on sandwich variance

IPD-AD

Analysis	Cox HR (95% CI)	Forest plots
Naive	0.63 (0.44, 0.90)	
MAIC	0.64 (0.45, 0.93)	

0 0.5 1 1.5

IPD-IPD

Analysis	Cox HR (95% CI)	Forest plots
Naive	0.63 (0.44, 0.90)	
IPW	0.65 (0.45, 0.94)	

0 0.5 1 1.5

Based on bootstrap variance

IPD-AD

Analysis	Cox HR (95% CI)	Forest plots
Naive	0.62 (0.43, 0.89)	
MAIC	0.63 (0.42, 0.93)	
STC	0.61 (0.38, 0.97)	
Doubly robust	0.61 (0.38, 0.98)	

0 0.5 1 1.5

IPD-IPD

Analysis	Cox HR (95% CI)	Forest plots
Naive	0.62 (0.43, 0.89)	
IPW	0.64 (0.42, 0.99)	
Regression adjustment	0.68 (0.35, 1.29)	
Doubly robust	0.69 (0.36, 1.32)	

0 0.5 1 1.5

Abbreviations: AD, aggregate data; CI, confidence interval; HR, hazard ratio; IPD, individual-level patient data; IPW, inverse probability weighting; MAIC, matching adjusted indirect comparison; STC, simulated treatment comparison

10

Moving onto the results section, first Cox HR
On the left-hand side, original data were fitted, and sandwich variance estimators were used to get 95% CI for propensity score based models. On the right hand side, bootstrap samples were used to estimate the hazard ratios and 95% CIs

HRs ranged from 0.61 to 0.69, meaning Nivolumab had longer survival than SOC.

PRECISIONheor

Results: Cox hazard ratio

Based on sandwich variance

IPD-AD

Analysis	Cox HR (95% CI)	Forest plots
Naive	0.63 (0.44, 0.90)	
MAIC	0.6	

IPD-IPD

Analysis	C	Forest plots
Naive	0	
IPW	0.6	

0 0.5 1 1.5

Based on bootstrap variance

IPD-AD

Analysis	Cox HR (95% CI)	Forest plots
Naive	0.62 (0.43, 0.89)	

0 0.5 1 1.5

Analysis		Forest plots
Regression adjustment	0.68 (0.35, 1.29)	
Doubly robust	0.69 (0.36, 1.32)	

0 0.5 1 1.5

- Similar point estimates across various analysis methods regardless of data availability from Flatiron (AD versus IPD)
- Smaller variance estimates for IPD-AD analyses compared to IPD-IPD analyses

Abbreviations: AD, aggregate data; CI, confidence interval; HR, hazard ratio; IPD, individual-level patient data; IPW, inverse probability weighting; MAIC, matching adjusted indirect comparison; STC, simulated treatment comparison

11

The point estimates are similar across various analysis method; and across Flatiron AD vs. IPD availability.

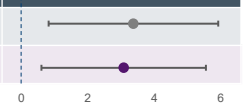
However, the variance estimators differed slightly. CIs were slightly narrower for IPD-AD analysis compared to IPD-IPD analysis.

Results: Difference in restricted mean survival

Based on sandwich variance

IPD-AD

Analysis	DRMST (months) with 95% CI	Forest plots
Naive	3.37 (0.82, 5.92)	
MAIC	3.08 (0.60, 5.55)	



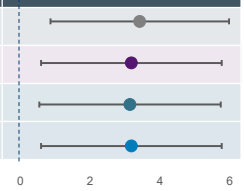
IPD-IPD

Analysis	DRMST (months) with 95% CI	Forest plots
Naive	3.37 (0.82, 5.92)	
IPW	2.92 (0.50, 5.34)	

Based on bootstrap variance

IPD-AD

Analysis	DRMST (months) with 95% CI	Forest plots
Naive	3.42 (0.86, 5.98)	
MAIC	3.18 (0.59, 5.77)	
STC	3.14 (0.54, 5.74)	
Doubly robust	3.18 (0.59, 5.77)	



IPD-IPD

Analysis	DRMST (months) with 95% CI	Forest plots
Naive	3.42 (0.86, 5.98)	
IPW	2.98 (0.21, 5.74)	
Regression adjustment	3.13 (0.56, 5.69)	
Doubly robust	3.15 (0.59, 5.71)	

Note: RMST was calculated up to 26.35 months; Abbreviations: AD, aggregate data; CI, confidence interval; DRM, doubly robust model; DRMST, difference in restricted mean survival time; IPD, individual-level patient data; IPW, inverse probability weighting; MAIC, matching adjusted indirect comparison; STC, simulated treatment comparison

In terms of difference in restricted mean survival, estimates ranged from 2.9 to 3.2 months, meaning Nivolumab had on average 3 months longer mean survival than SOC.

Results: Difference in restricted mean survival

Based on sandwich variance

IPD-AD

Analysis	DRMST (months) with 95% CI	Forest plots
Naive	3.37 (0.82, 5.92)	
MAIC	3.0	

IPD-IPD

Analysis	DRM	Forest plots
Naive	3	
IPW	2.5	

Based on bootstrap variance

IPD-AD

Analysis	DRMST (months) with 95% CI	Forest plots
Naive	3.42 (0.86, 5.98)	
Regression adjustment	3.13 (0.56, 5.69)	
Doubly robust	3.15 (0.59, 5.71)	

- Similar point estimates and variance estimates across various analysis methods regardless of data availability from Flatiron (AD versus IPD)

*Note: RMST was calculated up to 26.35 months; **Abbreviations**: AD, aggregate data; CI, confidence interval; DRM, doubly robust model; DRMST, difference in restricted mean survival time; IPD, individual-level patient data; IPW, inverse probability weighting; MAIC, matching adjusted indirect comparison; STC, simulated treatment comparison*

13

This time, both point estimates and variance estimates were similar across analysis models and data availability.

Strength and limitations

- Strengths
  - Collapsibility issue (marginal versus conditional effect) was resolved such that Cox HR could be compared across various analysis method
  - Non-parametric restricted mean survival time (RMST) extended to parametric models using pseudo-RMST
- Limitations
  - Differences between populations were limited due to selection of patients in Flatiron to align with CheckMate 032
    - Small sample sizes
    - Excluded variables where some differences due to missing data
    - Covariates dichotomized so did not assess matching on standard deviations
  - Case study may not be generalizable; cannot determine degree of bias or true coverage

14

First couple of strengths of our study

1. We dealt with non-collapsibility issue of HR by marginalizing the conditional HR from regression models.
2. RMST analysis was extended parametric models using pseudo-RMST

However, there were couple of limitations in this analysis

1. We selected Flatiron pop to be similar to CheckMate 032, which may not be feasible to do so in a typical IPD-AD analysis
2. A variable with a high missing rate was excluded, which could be an important prognostic factor
3. All covariates were binarized such that variance was not accounted for in MAIC
4. Therefore, this case study results may not be generalizable as we can't determine the bias and true variance, which is the disadvantage of all case studies.

Conclusion

- Unanchored indirect comparisons of two interventions for overall survival (OS) :
  - IPD-AD analysis was comparable to IPD-IPD analysis in terms of point estimates
  - Variance estimate of relative treatment effects tended to be larger for IPD-IPD analyses compared to IPD-AD analyses, especially for Cox models
- Doubly robust methods were consistent with other methods, and can be applied to IPD-AD scenarios, using either Cox model or restricted mean survival time analysis
- However, all unanchored indirect comparisons rely on strong assumptions that
  - No unmeasured confounders
  - Sufficient overlap between two studies

Abbreviations: *AD, aggregate data; IPD, individual-level patient data*

15

The take home message is

1. Our unanchored ITC for overall survival showed that IPD-AD analysis performed as well as IPD-IPD analysis in terms of point estimate of relative treatment effect estimation
2. while the variance estimate was smaller for IPD-AD analysis compared to IPD-IPD analysis, especially for Cox models
3. Doubly robust methods had consistent estimates and can be used in IPD-AD setting

4. However, we need to be mindful of that all unanchored ITC relies on very strong assumptions that there's no unmeasured confounders and this method works only when there's good overlap between two studies.

Which makes us wonder when are these ITC results valid? We were able to select Flatiron population to have sufficient overlap with CheckMate 032 but there's no way of knowing if all confounders were adjusted for.

Next steps: Simulation study

- Variances of estimates with IPD-AD and IPD-IPD:
  - IPD-AD confidence intervals underestimated?
  - Are IPD-IPD confidence intervals for Cox HR appropriate or over-estimated, especially for regression based method?
  - Do doubly-robust confidence intervals have better coverage than other methods?
- Validity of estimates across varying scenarios:
  - Which methods are least biased?
  - Which models perform best when the they are mis-specified with respect to the true prognostic factors and effect modifiers?
    - Doubly robust models are truly doubly robust?
  - Do we need to account for uncertainty of prognostic factors in AD when MAIC is performed, in addition to matching mean values?

Abbreviations: *AD, aggregate data; HR, hazard ratio; IPD, individual-level patient data; MAIC, matching adjusted indirect comparison*

16

To answer that question, we are performing simulation studies
1. In terms of variance estimates, we want to explore which analysis methods estimate variance properly given different data availability.

2. In terms of point estimates, we want to assess which method has least bias given various scenarios, and when models are mis-specified.

THANK YOU!

we look forward to staying in touch

Appendix

Cox model: Marginal versus conditional effect

Propensity score-based method

Targets **marginal effects**

(i.e. average effect over population of prognostic factors)

Regression model

Targets **conditional effects**

(i.e. effect at specific values of prognostic factors)

Y=Outcome (i.e. overall survival)

T=Treatment (i.e. NIVO vs. SOC)

C1=Binary patient characteristic (i.e. sensitive vs. resistant)

Abbreviations: *SOC, Standard of care*

19

The doubly robust method is supposed to combine propensity score method and outcome regression method. The problem is the collapsibility issue, in Cox model.
PS methods give the marginal effect. Which means the HR is the average relative treatment effect over the target population.
But regression methods give the conditional effect. For example, if a regression model was adjusted for gender, then that HR is for males or among females, not considering the distribution of gender in the target population.

Cox model: Marginal versus conditional effect

Propensity score-based method

Targets **marginal effects**

(i.e. average effect over population of prognostic factors)

Regression model

Targets **conditional effects**

(i.e. effect at specific values of prognostic factors)

We can **marginalize** the log hazards to align with propensity score methods by averaging predicted survival over covariates from target population using either observed or simulated IPD

Y=Outcome (i.e. overall survival)

T=Treatment (i.e. NIVO vs. SOC)

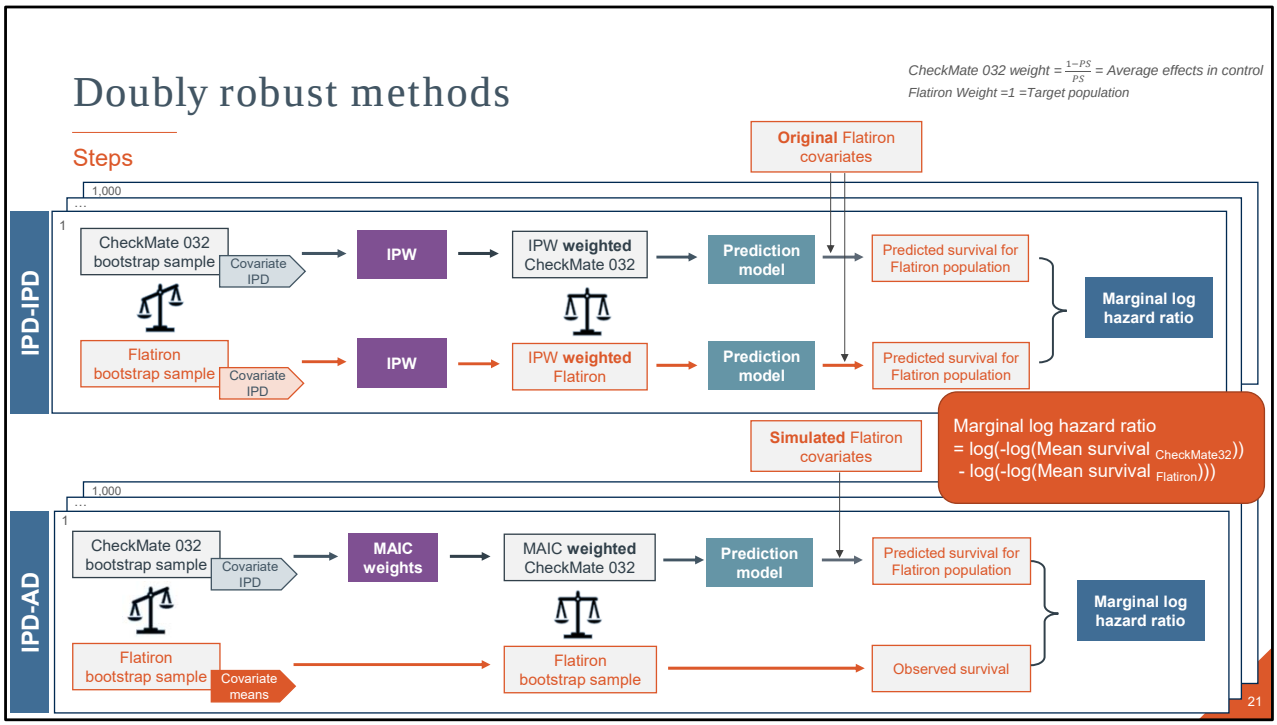
C1=Binary patient characteristic (i.e. sensitive vs. resistant)

Abbreviations: *IPD, individual patient data; SOC, Standard of care*

20

So we need to marginalize the log HR. Instead of fitting a regular regression model.

Such that both methods have marginal HRs so we have a consistent estimate for the doubly robust model. Then the question is how can you marginalize the conditional effect? We need to use bootstrap samples. (Go to next slide)



Here let me focus on IPD-AD setting, the new method we explored. So we bootstrap data 1000 times.

Let's look at bootstrap sample 1, we apply MAIC weights and build a typical regression model. Since only AD are available for flatiron, 1000 simulated flatiron covariates were plugged into this model to get 1000 predicted survivals at 26.35 months. Take mean of it, plug into the equation above.
For Flatiron, bootstrap 78 observed survivals, take the mean of them, plug it into the equation above.

Once you have 1000 marginal log HR, you can estimate the point estimate and variance.

26.35 months was made because Cox HR uses data up to the largest event time across two studies.