# External controls

## Comparing quality of life - while alive - between treatment and (external) controls: methods for real world analysis in clinical trials

Goetghebeur, Reynders, Thomassen, le Cessie, WP3

ISCB Milan, August 31, 2023

# The EU IMI-SISAQOL project


www.imi.europa.eu/projects-results/project-factsheets/sisaqol-imi

Setting International Standards in Analysing Patient-Reported Outcomes and Quality of Life Endpoints in Cancer Clinical Trials

# IMI-SISAQOL and STRATOS

- develops guidelines for
  evaluating the effect of novel treatments on PROMs in oncology populations.
- by seeking consensus internationally and across stakeholders
  - industry
  - academics
  - patients
  - trial organizations
  - regulators (e.g. FDA and EMA)

STRATOS helps develop statistical guidelines as a project partner

# Broad consensus guidelines...

- guidelines are growing in reaching consensus...
- as more people agree, they become more general/conceptual

What are we specifically recommending/accepting in practice?

Thanks to the collaboration with Pfizer: two case studies for statistical learning. [1]

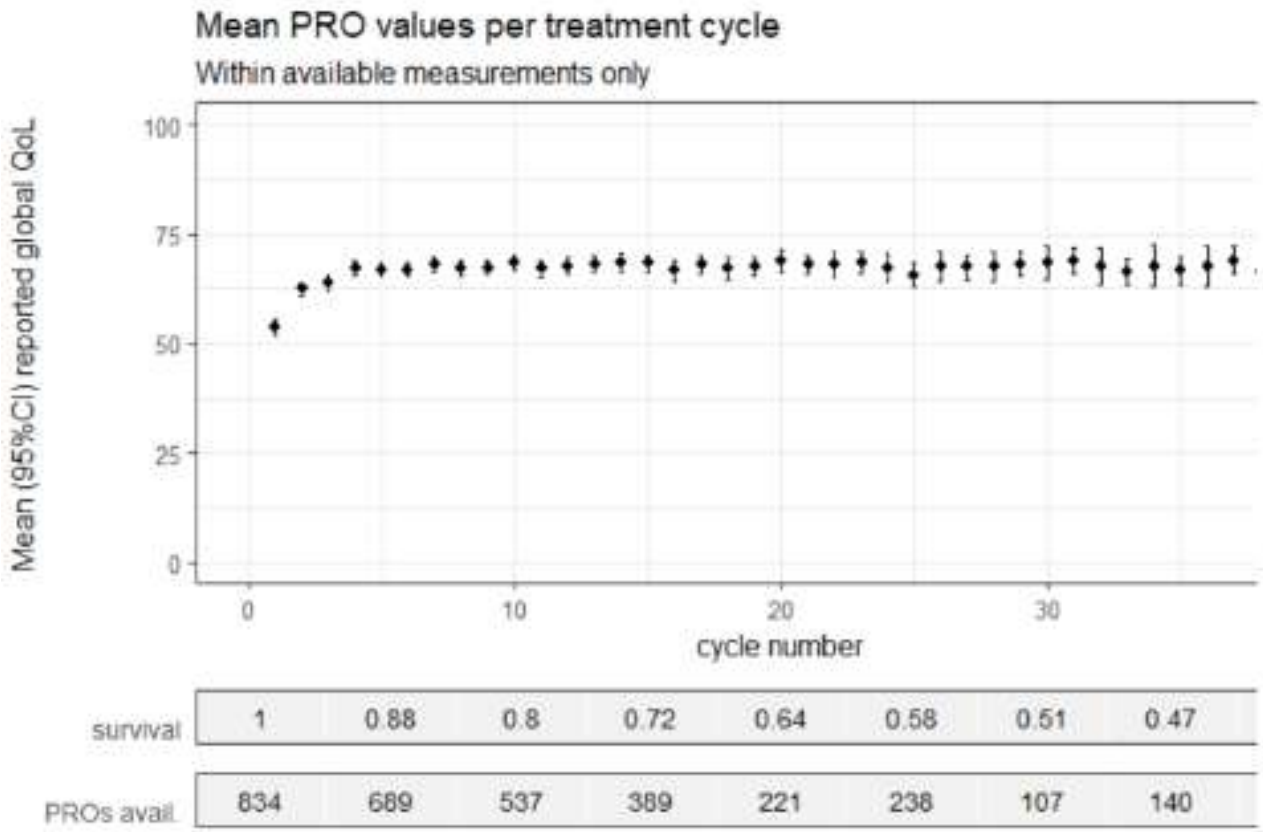
- one single arm: PROFILE 1005
- RCT on the same drug: PROFILE 1007

We start with the single arm and refer to the control of the RCT
we also learn lessons for the RCT. [2]

[1] Blackhall F, et al. ESMO Open 2017;2:e000219. doi:10.1136/esmoopen-2017- 000219; 'Crizotinib versus Chemotherapy in Advanced ALK-Positive Lung Cancer', n engl j med 368;25 nejm.org june 20, 2013

[2] Our data resemble but are not equal to the data previously published on.

# Outcome descriptives in our single arm study




30

Average QLQ-C30 [0-100] score on 'available' data...
for highly variable sample size ...

# ICH E9 [3] (R1) Addendum: Estimands

The definition of an estimand contains 5 components:

- The **population**.
- The **treatments** or **intervention** under **comparison** .
- The **outcome** variable.
- How to account for **intercurrent events**.
- The **variable summary** that is used to compare the treatment arms.

[3] The **International Council for Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH)** brings together regulatory authorities and the pharmaceutical industry. It makes recommendations towards achieving greater harmonisation in the interpretation and application of technical guidelines and requirements for pharmaceutical product registration.

# Concerning death as intercurrent event

ICH-E9 foresees

1. Evaluate QOL- while alive
2. Integrate death in composite outcome
3. Treat death as missing outcome (mixed models) implicitly imputed

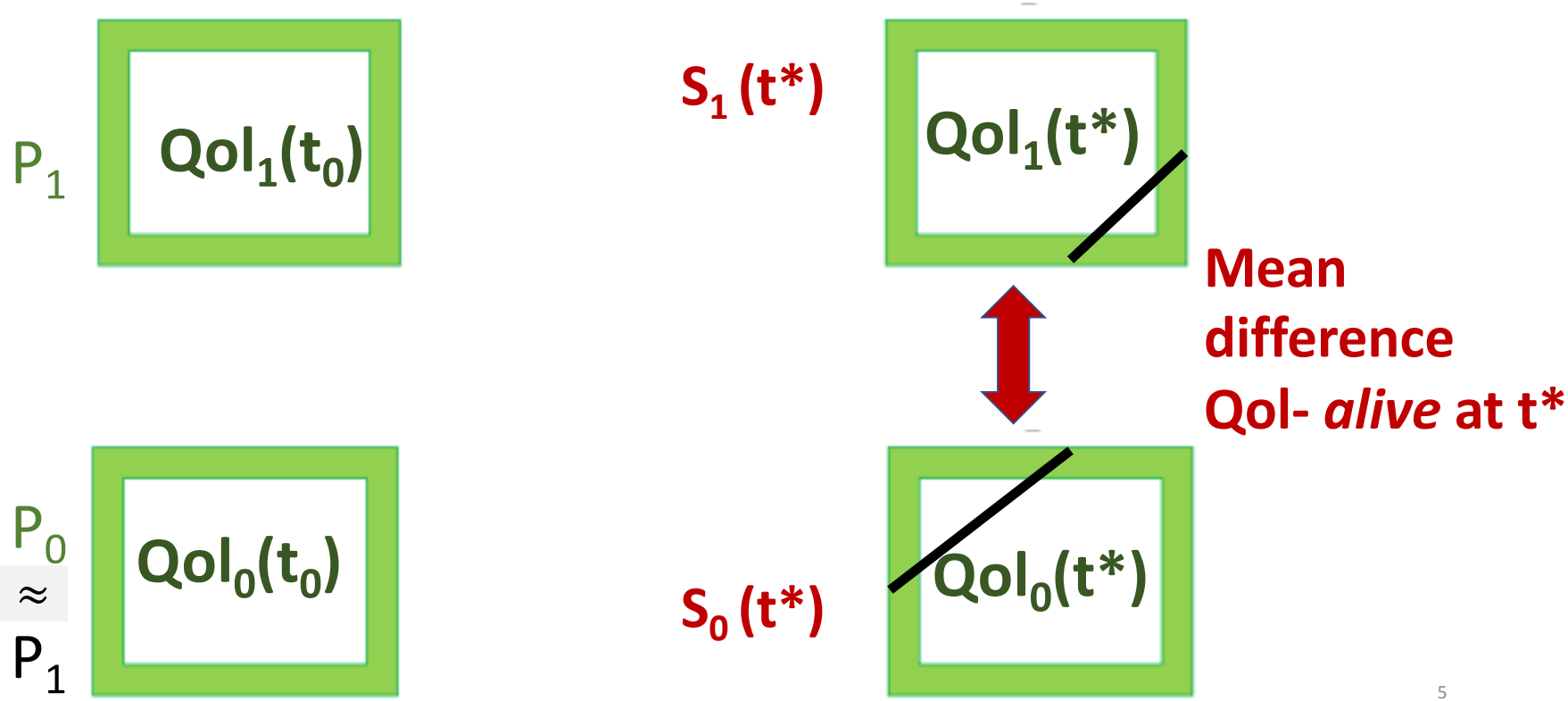
We consider:

- evaluate QOL- while alive = real PROM that matters in the real world
- plus the survival curve

# Our focus and bivariate estimand - in RCT

$$E^1[QOL_1(t* \mid \text{alive}) - QOL_0(t* \mid \text{alive})]$$

RCT equivalent control group at $t_0$, *then death (>60%)*

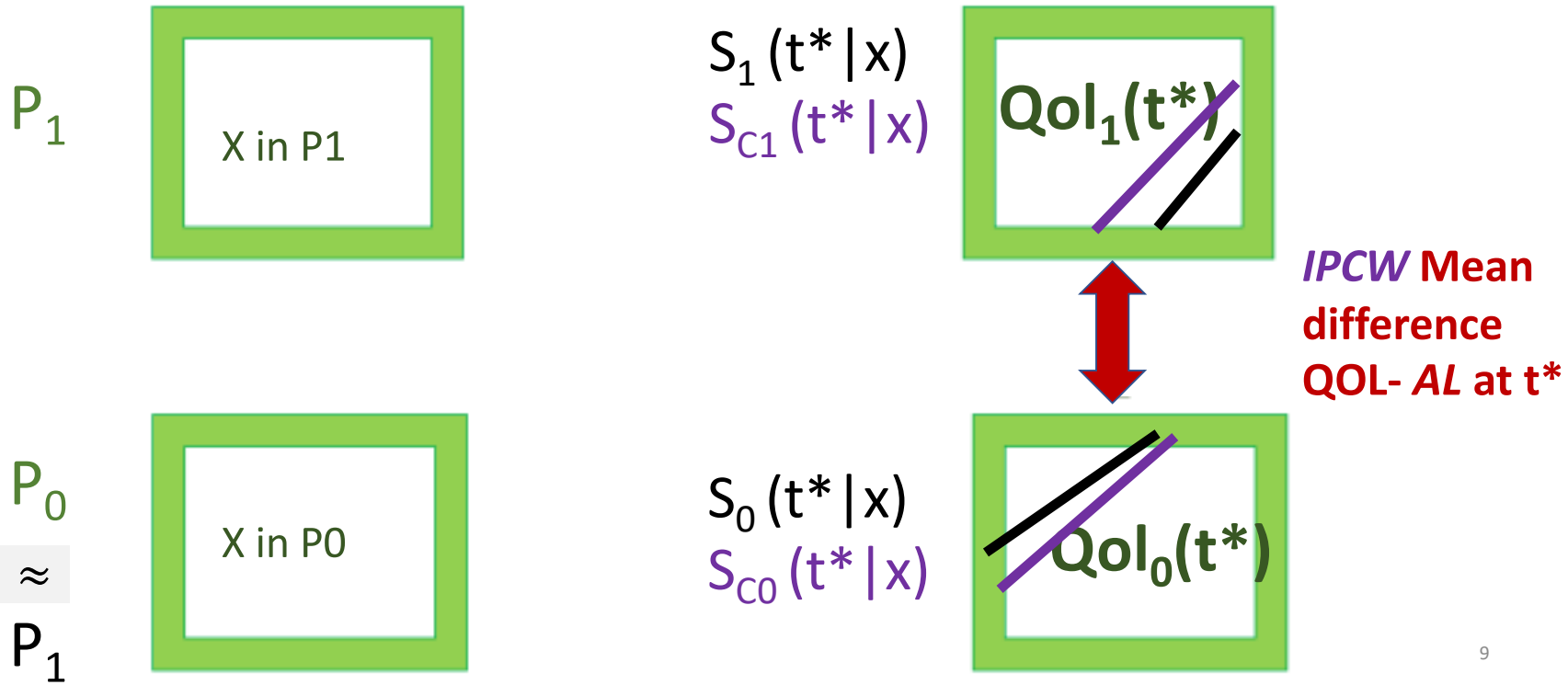



5

# Data challenges

$$E^1[QOL_1(t*|X, \text{alive}) - QOL_0(t*|X, \text{alive})]$$

## Realistic: Equivalent control group *and death + (admin) censoring*

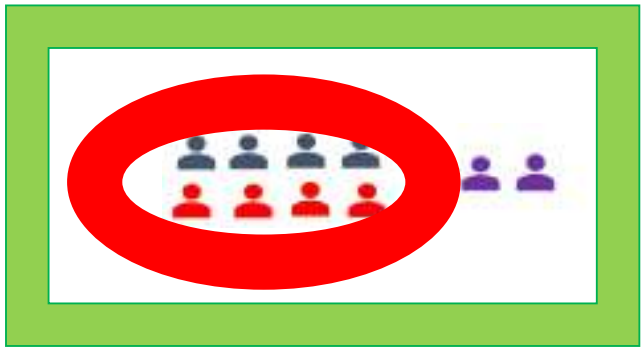



NUCC: $X = \{X_T, X_C\}$

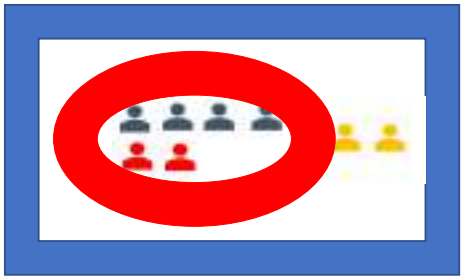
# Our focus and estimand - in Single arm

Involving **external control** under NUC and beyond, currently from RCT could come from disease register or meta-analysis.

Single Arm **P1 at t0**



RCT Control Arm **$P_0$ at t0**



- **Positive chance of membership of P1 *and* P0**
- Still different x-distributions in P1 and P0 -> $\prod_1 (\mathbf{x})$
- **Focus on the P1** ➡ direct standardization

  $P_0$ -> $P_1$

  ➡ **ATT**

11

# Handeling positivity: some lack of overlap

N (%)
Median (IQR)

| Characteristic | RCT control N = 158 | Treated single arm, N = 879 |
|---|---|---|
| ECOG cat ... 3: | 0 (0%) | 28 (3.2%) |
| Unknown | 0 | 3 |
| Number of Prior Drug Therapy Regimens | | |
| 1 | 128 (81%) | 181 (21%) |
| 2 | 29 (18%) | 323 (37%) |
| 3 | 1 (0.6%) | 189 (22%) |
| 4 | 0 (0%) | 91 (10%) |
| 5 | 0 (0%) | 49 (5.6%) |
| 6 | 0 (0%) | 22 (2.5%) |
| 7 | 0 (0%) | 9 (1.0%) |
| 8 | 0 (0%) | 6 (0.7%) |
| 9 | 0 (0%) | 4 (0.5%) |
| 10 | 0 (0%) | 1 (0.1%) |
| 12 | 0 (0%) | 1 (0.1%) |
| Unknown | 0 | 3 |
| baseqol | 58 (42, 75) | 50 (33, 75) |
| Unknown | 10 | 41 |

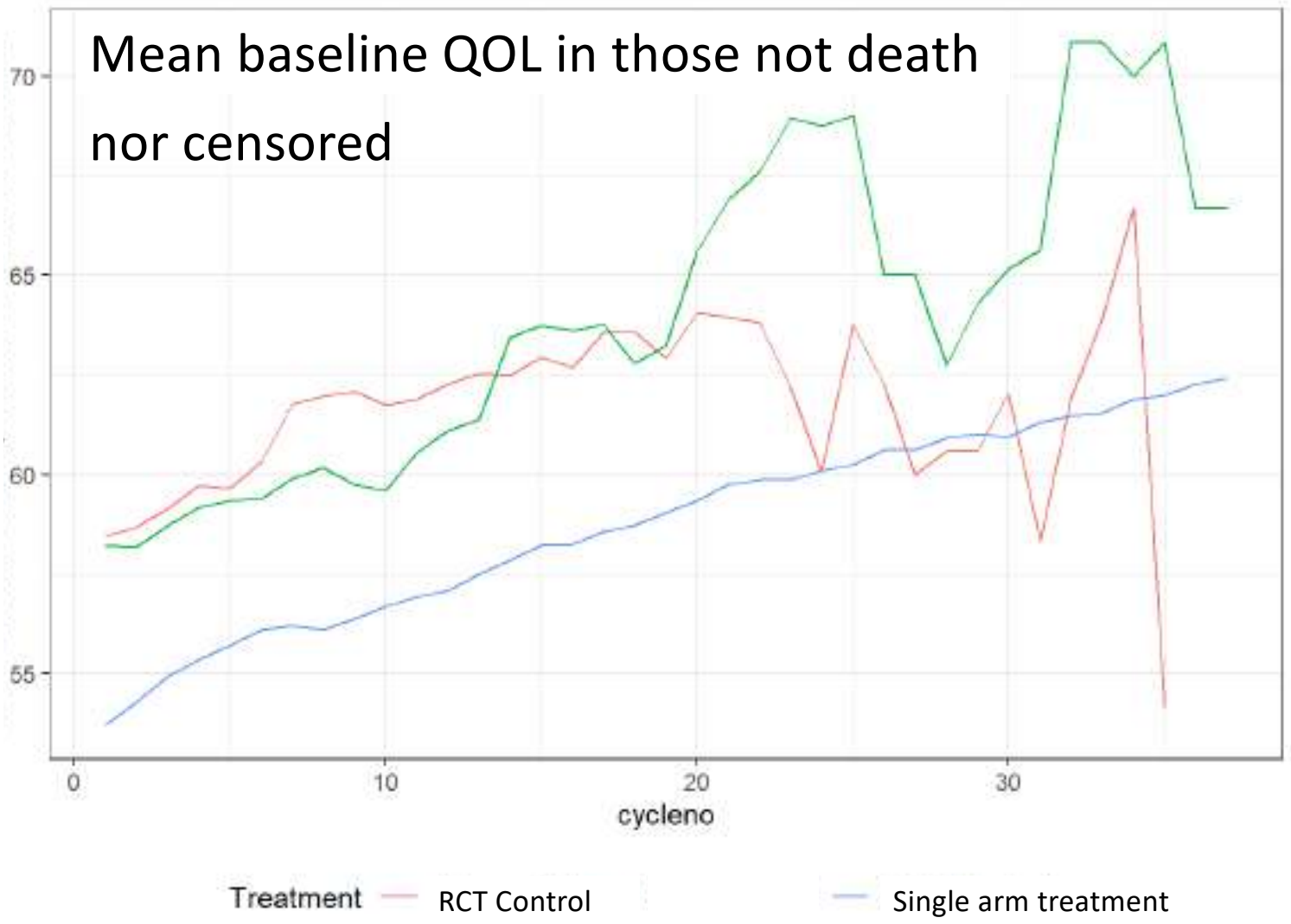
# Handeling positivity - reducing the scope

N (%)
Median (IQR)

| Characteristic | RCT control N=157 | Single arm treatment N=490 |
|---|---|---|
| Sex | | |
| FEMALE | 88 (56%) | 288 (59%) |
| MALE | 69 (44%) | 202 (41%) |
| Age | 49 (39, 59) | 53 (42, 63) |
| n (%); Median (IQR) | | |
| Baseline ECOG Score | | |
| 0 | 62 (39%) | 160 (33%) |
| 1 | 84 (54%) | 260 (53%) |
| 2 | 11 (7.0%) | 70 (14%) |
| 3 | 0 (0%) | 0 (0%) |
| Number of Prior Drug Therapy Regimens | | |
| 1 | 128 (82%) | 176 (36%) |
| 2 | 29 (18%) | 314 (64%) |
| 3 | 0 (0%) | 0 (0%) |
| baseqol | 58 (42, 75) | 58 (33, 75) |
| Unknown | 10 | 34 |

$\Rightarrow$: exclude ECOG at baseline 3 and # prior drug treatments $> 2$ ($>3$ is trouble)
Age-distributions are overlapping.

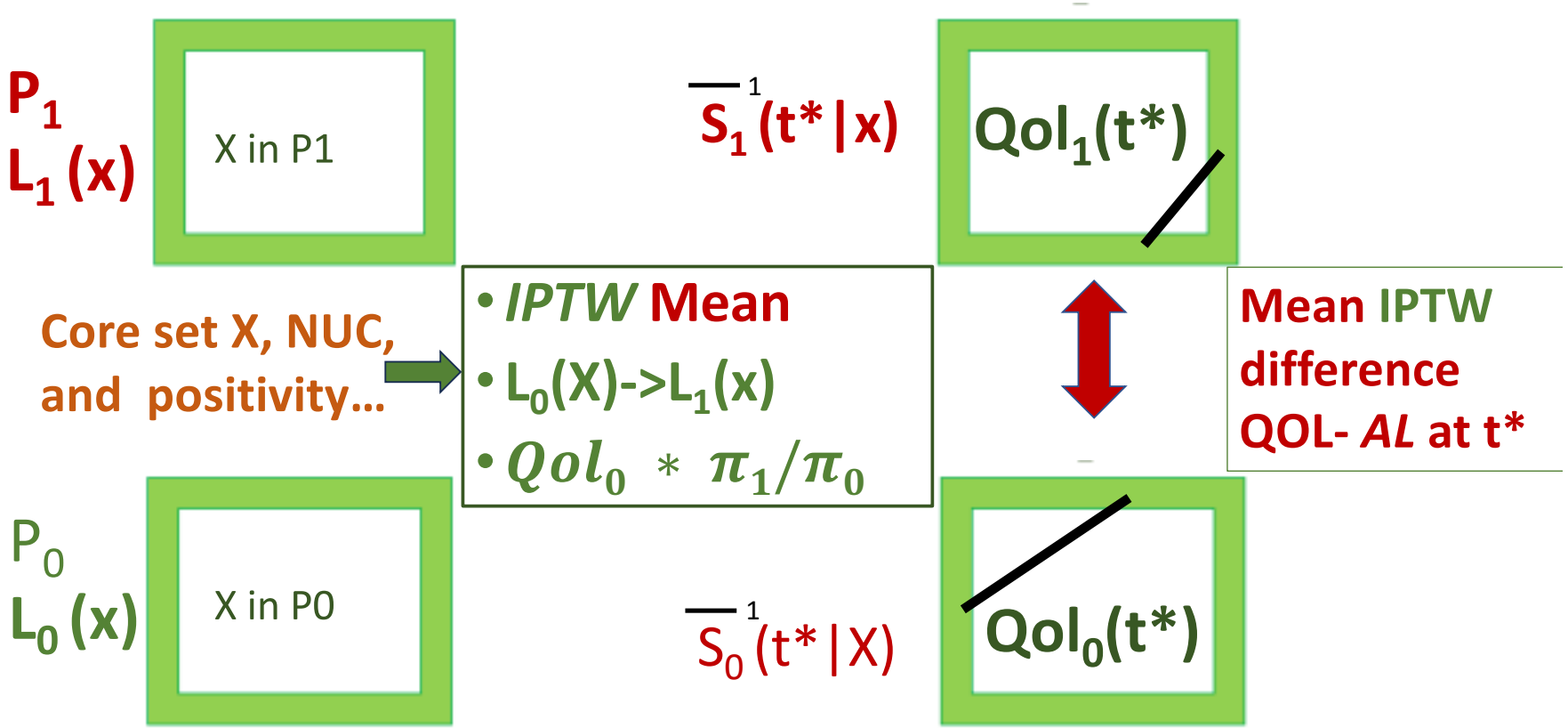
# Changing baseline distribution in QOL-*data*



# Our focus and estimand - in Single arm

$$E^1[QOL_1(t*|X, \text{alive}) - QOL_0(t*|X, \text{alive})]$$

*External* control group - *and death*

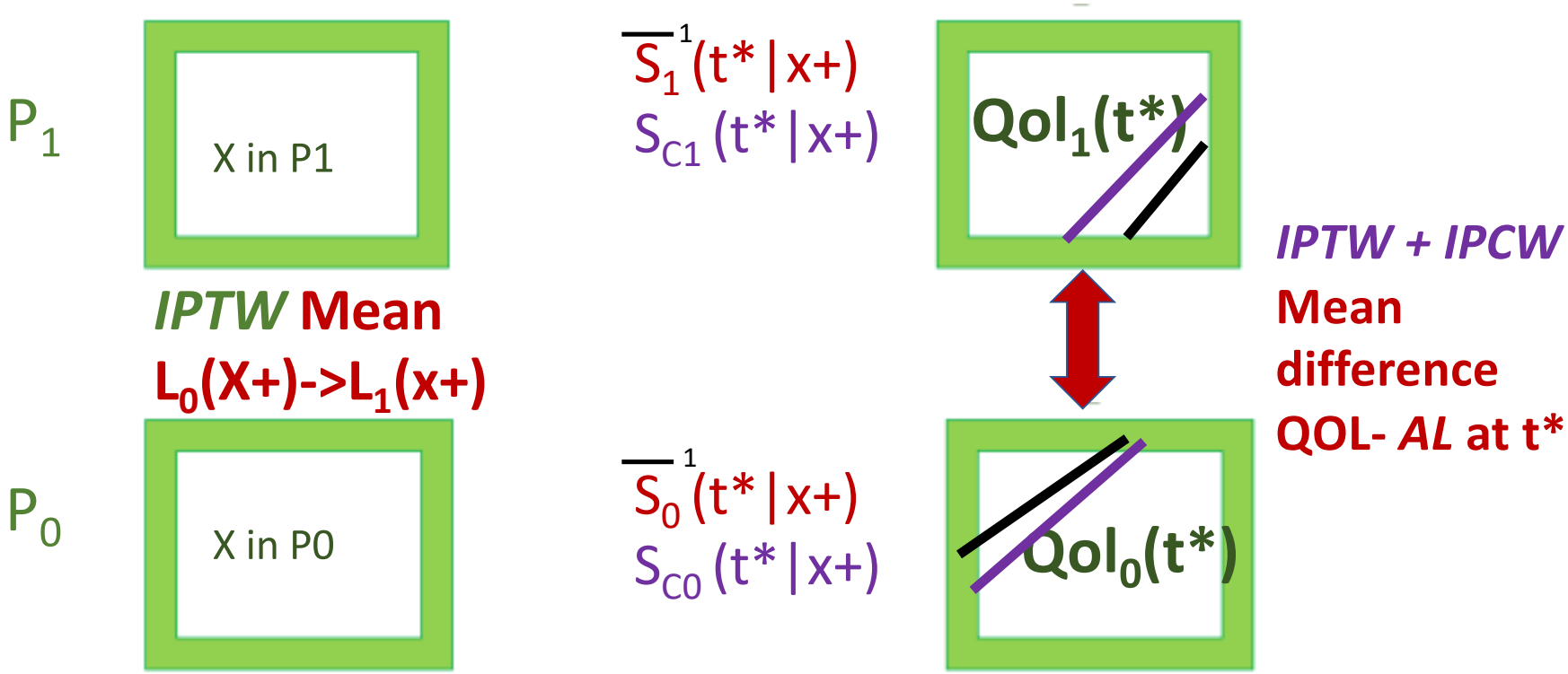


# Data challenges - in Single arm

$$\{T, QOL\} \perp C|X+$$

(RCT trt to be taken out)

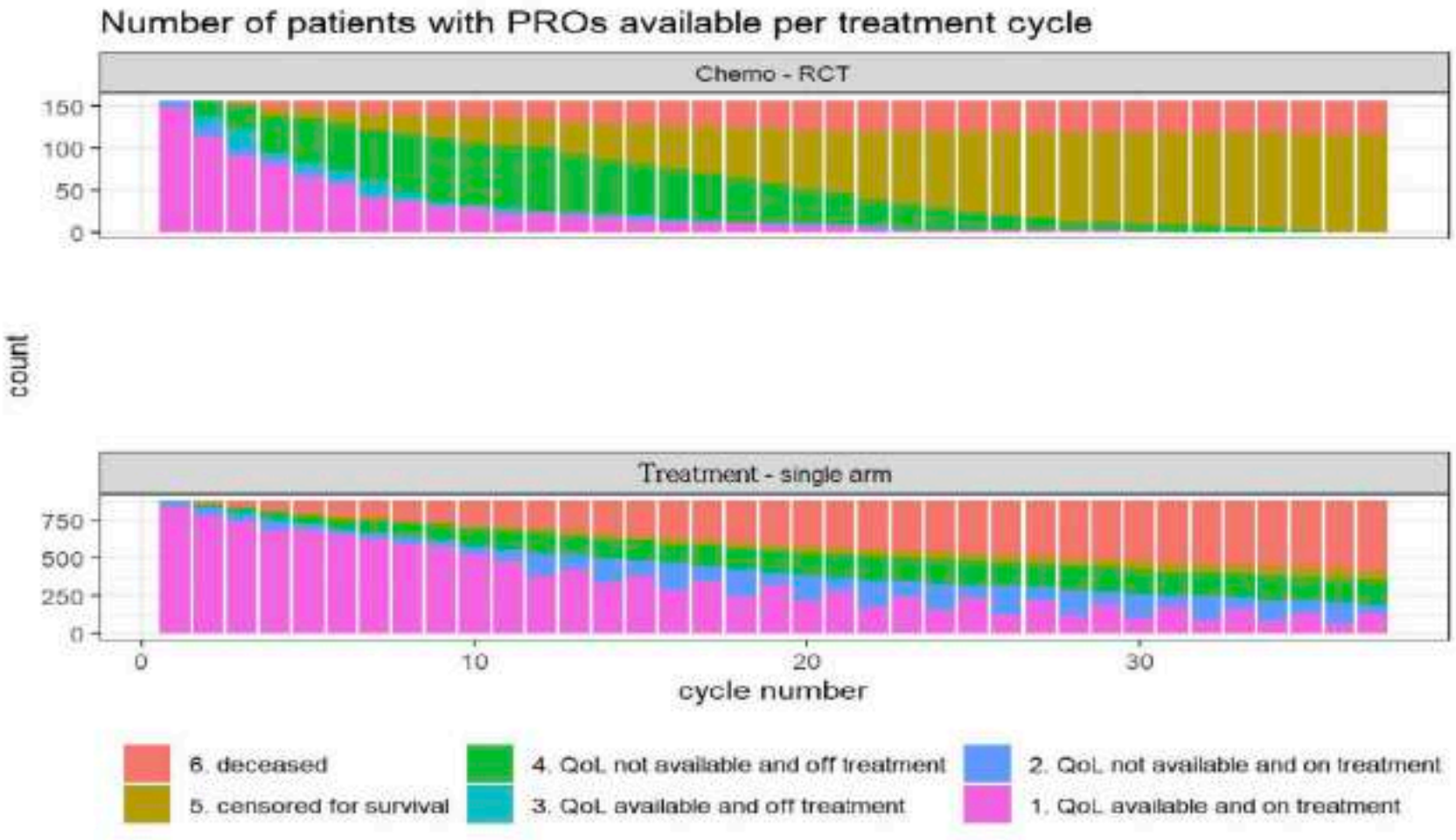
## Equivalent control group *and death + censoring*



# Our focus and estimand

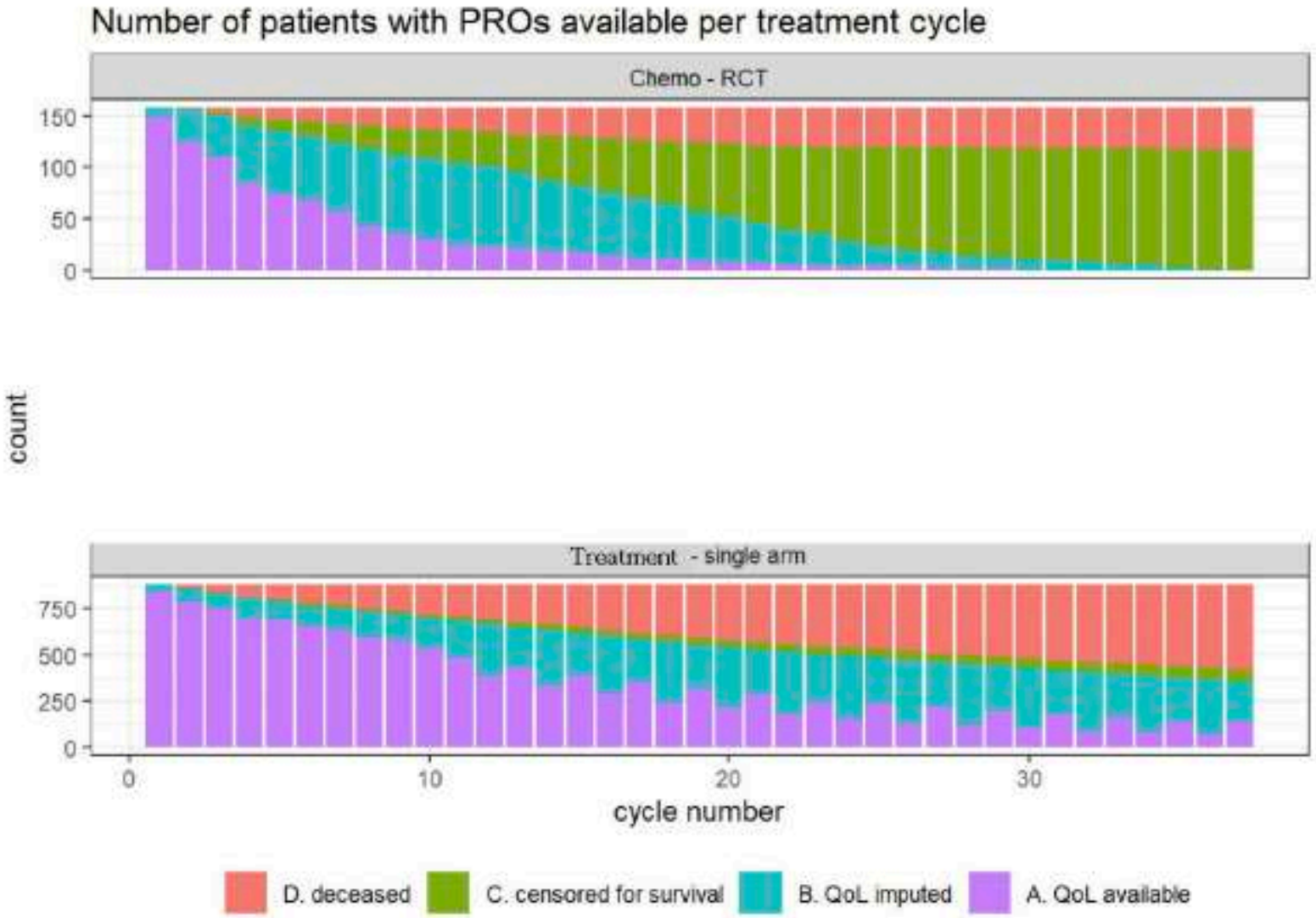
- The Single arm **population** P1: real world inclusion/exclusion (designed)
- The **treatments:** similar cycles in the single arm and external control
- The **outcome** variable: QoI (t*) [actual QLQ-C30 value at $t*$]
- **intercurrent/terminal death**
- The **variable summary** Mean QoI over the P1- alive at t*

# Data challenges



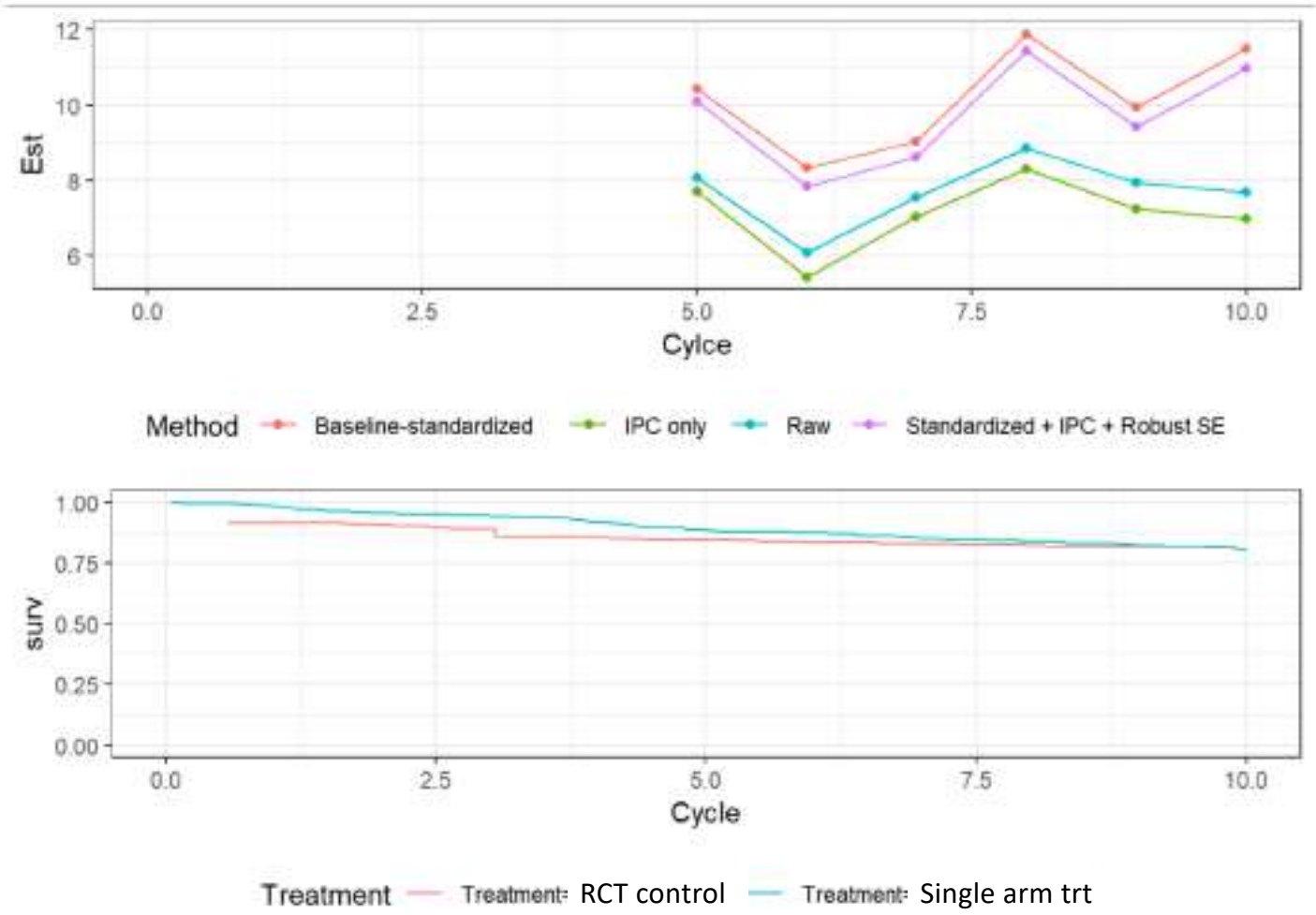
# Data challenges - upon imputing missingness

Building on work from Thomassen et al.

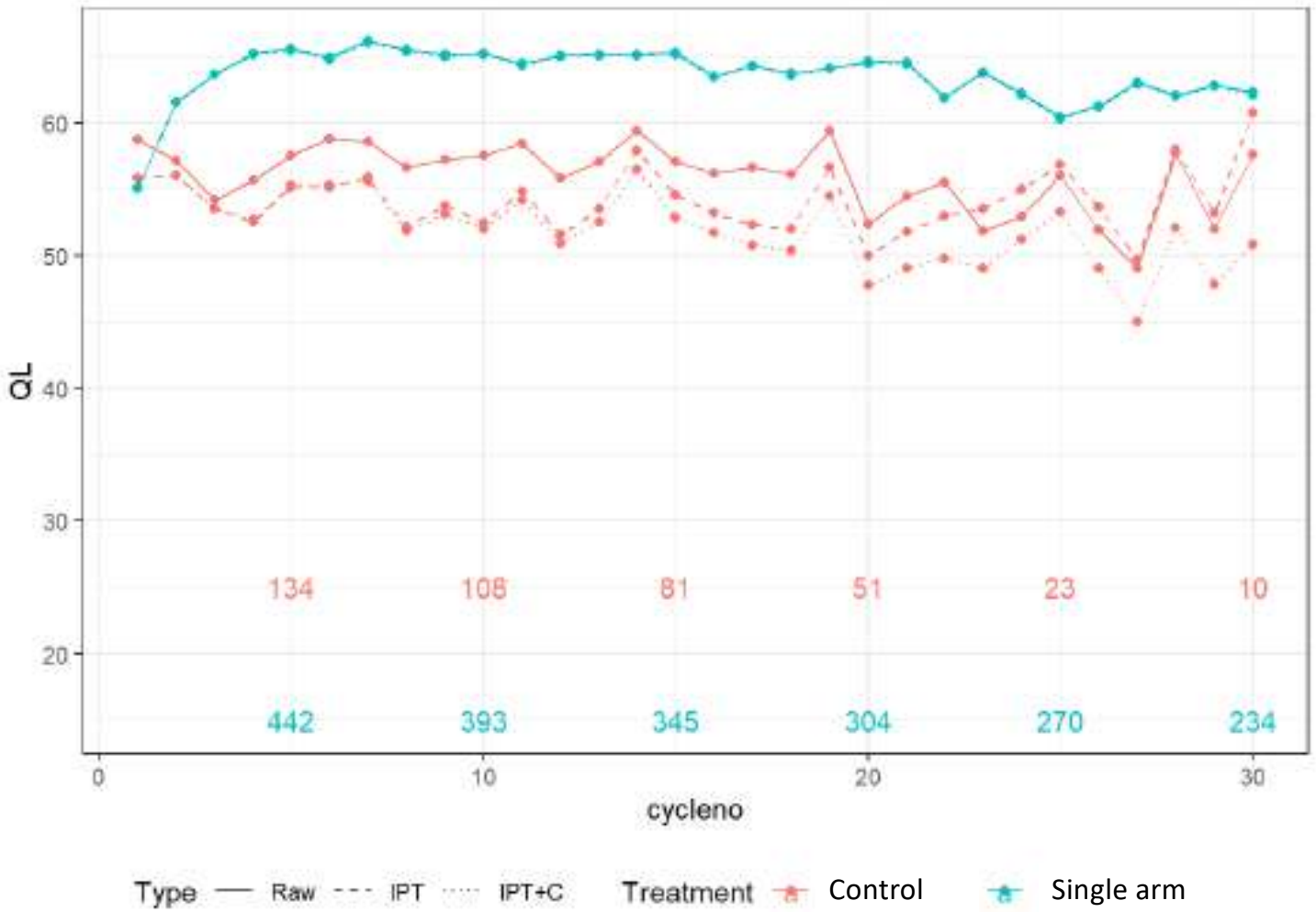


# ATT: Baseline-standardized QoL while alive analysis

IPTW involving $P(\text{Single arm}|X, T)$



# Impact of IPT and IPC weights



# On the variance estimation

1. Conditional on observed sample(s):

Standardization + Censoring weighted difference in observed QoL
Robust standard errors using the marginaleffects-package in R.
Results over timepoints are calculated separately (no GEE . . . )

| Method | Est | SE | Cylce |
|---|---|---|---|
| Raw | 8.06 | 2.02 | 5 |
| Baseline-standardized | 10.4 | 1.73 | 5 |
| IPC only | 7.72 | 1.98 | 5 |
| Standardized + IPC + Robust SE | 10.1 | 3.17 | 5 |
| Raw | 6.06 | 2.15 | 6 |
| Baseline-standardized | 8.33 | 1.86 | 6 |
| IPC only | 5.43 | 2.1 | 6 |
| Standardized + IPC + Robust SE | 7.85 | 3.43 | 6 |
| Raw | 7.55 | 2.01 | 7 |
| Baseline-standardized | 9.02 | 1.73 | 7 |
| IPC only | 7.03 | 1.94 | 7 |
| Standardized + IPC + Robust SE | 8.61 | 3.28 | 7 |
| Raw | 8.84 | 2.08 | 8 |
| Baseline-standardized | 11.9 | 1.76 | 8 |
| IPC only | 8.31 | 2 | 8 |
| Standardized + IPC + Robust SE | 11.4 | 3.23 | 8 |
| Raw | 7.95 | 2.13 | 9 |
| Baseline-standardized | 9.93 | 1.82 | 9 |
| IPC only | 7.23 | 2.04 | 9 |
| Standardized + IPC + Robust SE | 9.42 | 3.37 | 9 |

26

# 2. Acknowledging between study variation

We have seen one treatment and one control study so far

- we did a small review of older studies in similar population
- evaluated between study variance in (standardised) means
- added variation component to sampling variation seen within our two studies

# Alternatives to inverse weighting

1. Adjust $QOL(t*)$ & survival/censoring for baseline covariates can be done thru
   - outcome regression: $E(QOL(t*)|X+)$
   - then standardize involving weights $S(t|X+)$
2. When parts of the $X$ distribution are gone by $t^*$ due to censoring
   better to use GEE for the longitudinal QOL-while-alive
   thus borrowing on x-dependence at earlier times

   Pitfalls?
3. Can turn into a double robust analysis.

# Further issues

Compare mean $QOL_1(t*)$ while alive in treated population with 'comparable' $QOL_0(t*)$ while alive in control population

- Much to say about missing data past treatment change and disease progression
  should we restrict inference to QOL-while on treatment?
  should we avoid control groups of RCTs in this late stage oncology population?
- MAR dependent on time of death
  - What with censored for death? (with STRATOS TG8?)
  - What about MNAR (with new STRATOS guidelines TG1?)
- Measurement timing and frequency matters -> deserves its own paper

Harmonise with WP2?

# Core set of variables

Potential solution:

- All studies (single-arm, RCT, disease registries, ...) in a disease domain measure the same core of baseline variables
- Why? To facilitate comparisons of PRO results (also meta-analysis)

# In conclusion/for discussion

- We can learn from single arm studies, but it is never simple
- Much left to do, causal inference helps us greatly
  instead of inverse weighting, we can use outcome regression, imputation, double robust modelling
- Suggestions?