# Follow-up time in clinical trials with a time-to-event endpoint: Redefining the question(s)

*Kaspar Rufibach*
*Methods, Collaboration, and Outreach Group, Roche Basel*
*ISCB43, Newcastle, 22nd August 2022*

Paper written within Oncology estimand WG

**Lynda Grinsted**, AstraZeneca

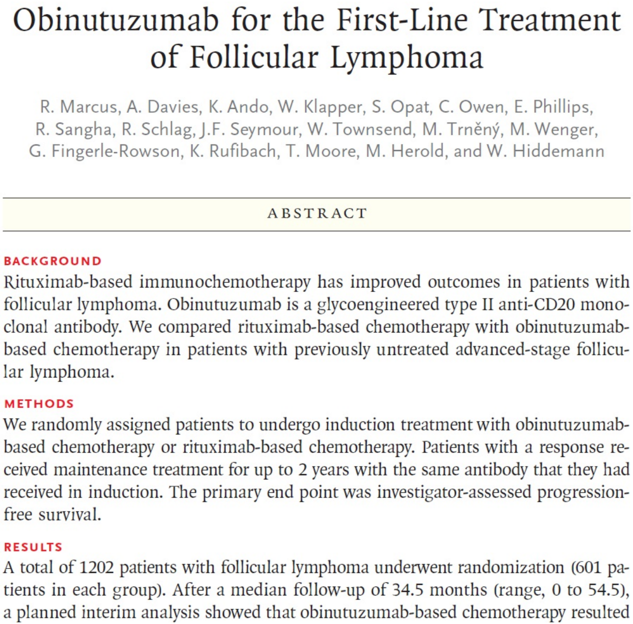
**Jiang Li**, BeiGene

**Yue Shentu**, Daiichi Sankyo

**Hans Jochen Weber**, Novartis

**Cheng Zheng**, Zentalis

**Jiangxiu Zhou**, J&J



# Obinutuzumab for the First-Line Treatment of Follicular Lymphoma

R. Marcus, A. Davies, K. Ando, W. Klapper, S. Opat, C. Owen, E. Phillips,
R. Sangha, R. Schlag, J.F. Seymour, W. Townsend, M. Trněný, M. Wenger,
G. Fingerle-Rowson, K. Rufibach, T. Moore, M. Herold, and W. Hiddemann

## ABSTRACT

### BACKGROUND

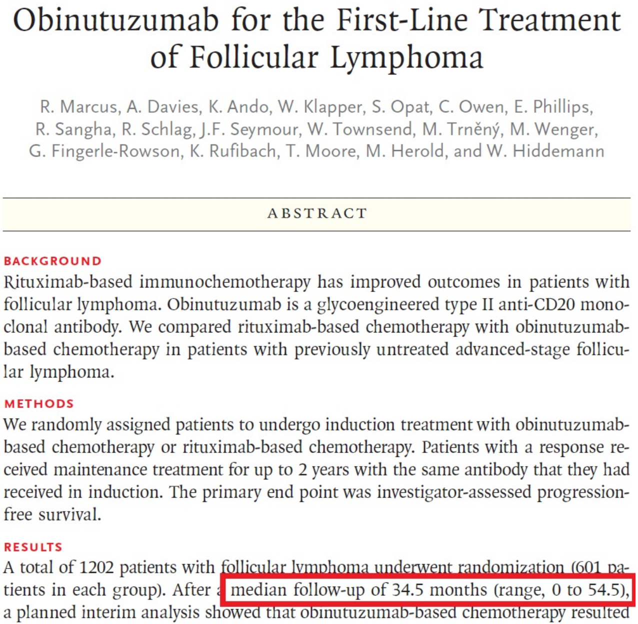
Rituximab-based immunochemotherapy has improved outcomes in patients with follicular lymphoma. Obinutuzumab is a glycoengineered type II anti-CD20 monoclonal antibody. We compared rituximab-based chemotherapy with obinutuzumab-based chemotherapy in patients with previously untreated advanced-stage follicular lymphoma.

### METHODS

We randomly assigned patients to undergo induction treatment with obinutuzumab-based chemotherapy or rituximab-based chemotherapy. Patients with a response received maintenance treatment for up to 2 years with the same antibody that they had received in induction. The primary end point was investigator-assessed progression-free survival.

### RESULTS

A total of 1202 patients with follicular lymphoma underwent randomization (601 patients in each group). After a median follow-up of 34.5 months (range, 0 to 54.5), a planned interim analysis showed that obinutuzumab-based chemotherapy resulted



# Obinutuzumab for the First-Line Treatment of Follicular Lymphoma

R. Marcus, A. Davies, K. Ando, W. Klapper, S. Opat, C. Owen, E. Phillips,
R. Sangha, R. Schlag, J.F. Seymour, W. Townsend, M. Trněný, M. Wenger,
G. Fingerle-Rowson, K. Rufibach, T. Moore, M. Herold, and W. Hiddemann

## ABSTRACT

**BACKGROUND**

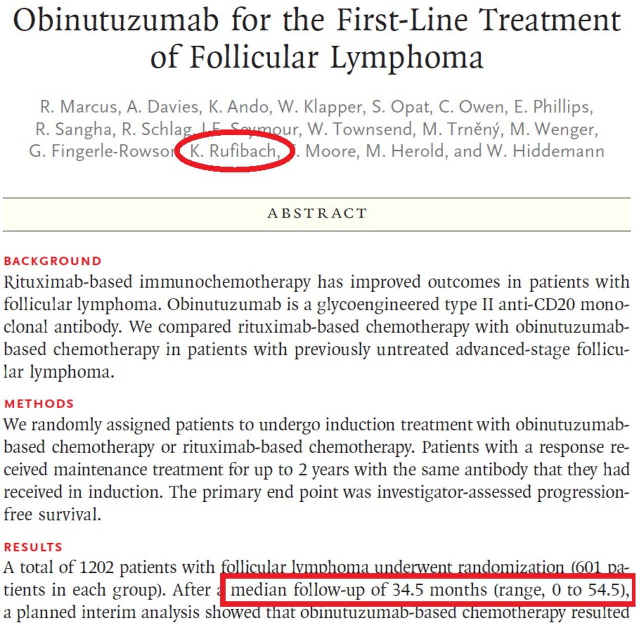
Rituximab-based immunochemotherapy has improved outcomes in patients with follicular lymphoma. Obinutuzumab is a glycoengineered type II anti-CD20 monoclonal antibody. We compared rituximab-based chemotherapy with obinutuzumab-based chemotherapy in patients with previously untreated advanced-stage follicular lymphoma.

**METHODS**

We randomly assigned patients to undergo induction treatment with obinutuzumab-based chemotherapy or rituximab-based chemotherapy. Patients with a response received maintenance treatment for up to 2 years with the same antibody that they had received in induction. The primary end point was investigator-assessed progression-free survival.

**RESULTS**

A total of 1202 patients with follicular lymphoma underwent randomization (601 patients in each group). After a median follow-up of 34.5 months (range, 0 to 54.5), a planned interim analysis showed that obinutuzumab-based chemotherapy resulted



# Obinutuzumab for the First-Line Treatment of Follicular Lymphoma

R. Marcus, A. Davies, K. Ando, W. Klapper, S. Opat, C. Owen, E. Phillips, R. Sangha, R. Schlag, J.F. Seymour, W. Townsend, M. Trněný, M. Wenger, G. Fingerle-Rowson, K. Rufibach, T. Moore, M. Herold, and W. Hiddemann

## ABSTRACT

### BACKGROUND

Rituximab-based immunochemotherapy has improved outcomes in patients with follicular lymphoma. Obinutuzumab is a glycoengineered type II anti-CD20 monoclonal antibody. We compared rituximab-based chemotherapy with obinutuzumab-based chemotherapy in patients with previously untreated advanced-stage follicular lymphoma.

### METHODS

We randomly assigned patients to undergo induction treatment with obinutuzumab-based chemotherapy or rituximab-based chemotherapy. Patients with a response received maintenance treatment for up to 2 years with the same antibody that they had received in induction. The primary end point was investigator-assessed progression-free survival.

### RESULTS

A total of 1202 patients with follicular lymphoma underwent randomization (601 patients in each group). After a median follow-up of 34.5 months (range, 0 to 54.5), a planned interim analysis showed that obinutuzumab-based chemotherapy resulted

# What do these 34.5 months mean?

**What do these 34.5 months mean?**

**What can we conclude from it?**

**Nothing!**

**Nothing!**

**Do not report such numbers at all!**

# What do trialists believe 34.5 months means?

**What do trialists believe 34.5 months means?**

**Shuster (1991): interviews with oncologists.**

# Median of...

# Median of...

**”Follow-up among those who did not have the event yet.”**

**Median of...**

**"Follow-up among those who did not have the event yet."**

**"Follow-up of all patients."**

**Median of...**

**”Follow-up among those who did not have the event yet.”**

**”Follow-up of all patients.”**

**”Time from trial entry to clinical cut-off date.”**

**Median of...**

**”Follow-up among those who did not have the event yet.”**

**”Follow-up of all patients.”**

**”Time from trial entry to clinical cut-off date.”**

**”Censoring distribution, estimated through inverse KM.”**

# What do trialists want to know?

# ”Maturity” of the estimated survival function.

**”Maturity” of the estimated survival function.**

**”Stability” of the estimated survival function.**

**”Maturity” of the estimated survival function.**

**”Stability” of the estimated survival function.**

**Time interval where Kaplan-Meier estimate is ”valid”.**

**”Maturity” of the estimated survival function.**

**”Stability” of the estimated survival function.**

**Time interval where Kaplan-Meier estimate is ”valid”.**

**”Quality” of follow-up.**

# ”Maturity”?

## ”Maturity”?

## ”Stability”?

## ”Maturity”?

## ”Stability”?

## ”Validity”?

**”Maturity”?**

**”Stability”?**

**”Validity”?**

**”Quality”?**

**”Maturity”?**

**”Stability”?**

**”Validity”?**

**”Quality”?**

**Trials compared based on vague concept of ”follow-up”.**

# Problem statement

”Follow-up quantification”:

# Problem statement

”Follow-up quantification”:

- Unclearly defined concept. **Confusion**!

# Problem statement

”Follow-up quantification”:

- Unclearly defined concept. **Confusion**!
- Different quantities used to ”answer” question. **Heterogeneity**!

# Problem statement

"Follow-up quantification":

- Unclearly defined concept. **Confusion**!
- Different quantities used to "answer" question. **Heterogeneity**!
- Precise computation rarely mentioned in publication. **Comparability**!

# Problem statement

”Follow-up quantification”:

- Unclearly defined concept. **Confusion**!
- Different quantities used to ”answer” question. **Heterogeneity**!
- Precise computation rarely mentioned in publication. **Comparability**!

Proposal:

# Problem statement

”Follow-up quantification”:

- Unclearly defined concept. **Confusion**!
- Different quantities used to ”answer” question. **Heterogeneity**!
- Precise computation rarely mentioned in publication. **Comparability**!

Proposal:

- Inspiration from estimand framework to **start with scientific question(s)** trialists want answers to.

# Problem statement

"Follow-up quantification":

- Unclearly defined concept. **Confusion**!
- Different quantities used to "answer" question. **Heterogeneity**!
- Precise computation rarely mentioned in publication. **Comparability**!

Proposal:

- Inspiration from estimand framework to **start with scientific question(s)** trialists want answers to.
- Analyze existing quantities.

# Problem statement

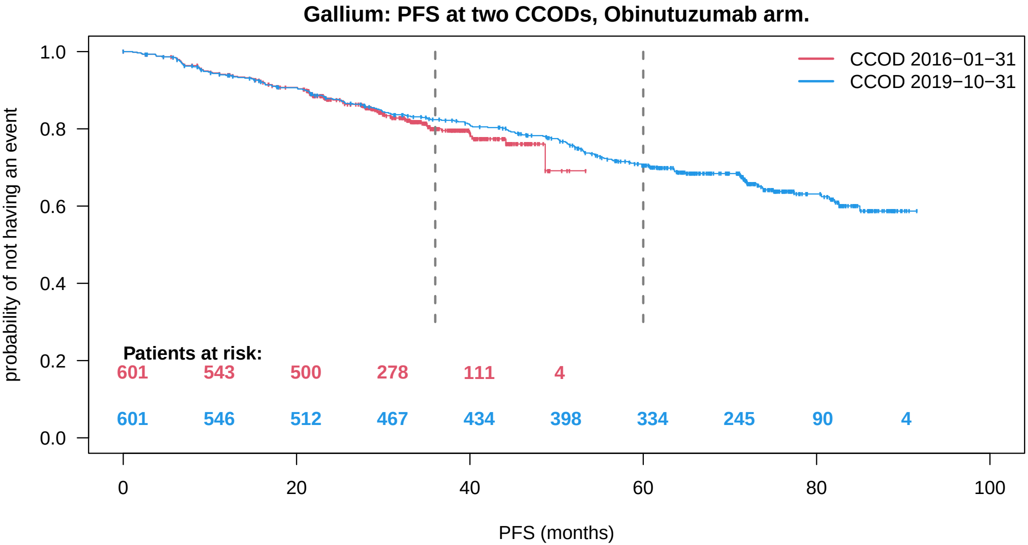
"Follow-up quantification":

- Unclearly defined concept. **Confusion**!
- Different quantities used to "answer" question. **Heterogeneity**!
- Precise computation rarely mentioned in publication. **Comparability**!

Proposal:

- Inspiration from estimand framework to **start with scientific question(s)** trialists want answers to.
- Analyze existing quantities.
- Extend considerations to 2-sample case. Separately for PH and non-PH.

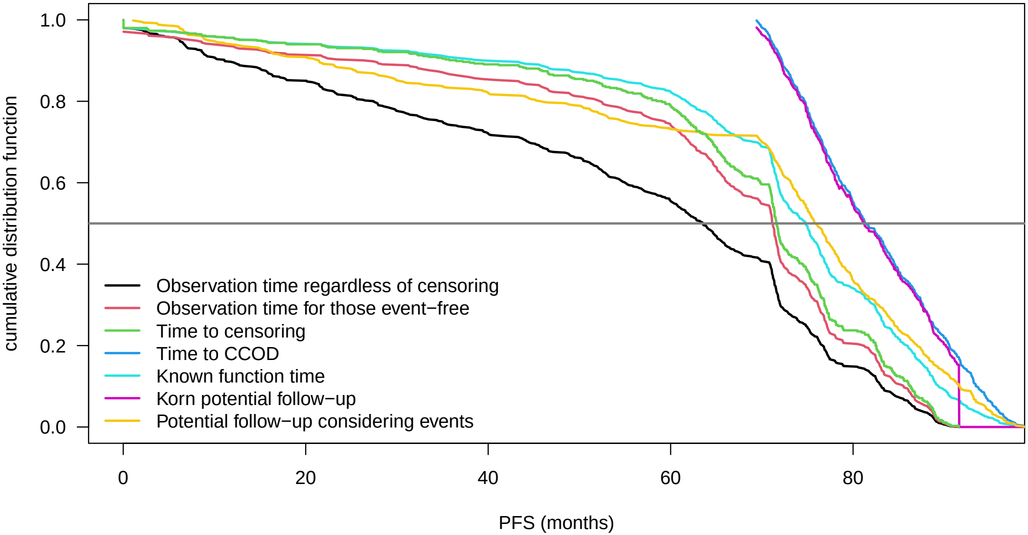
# One-sample case



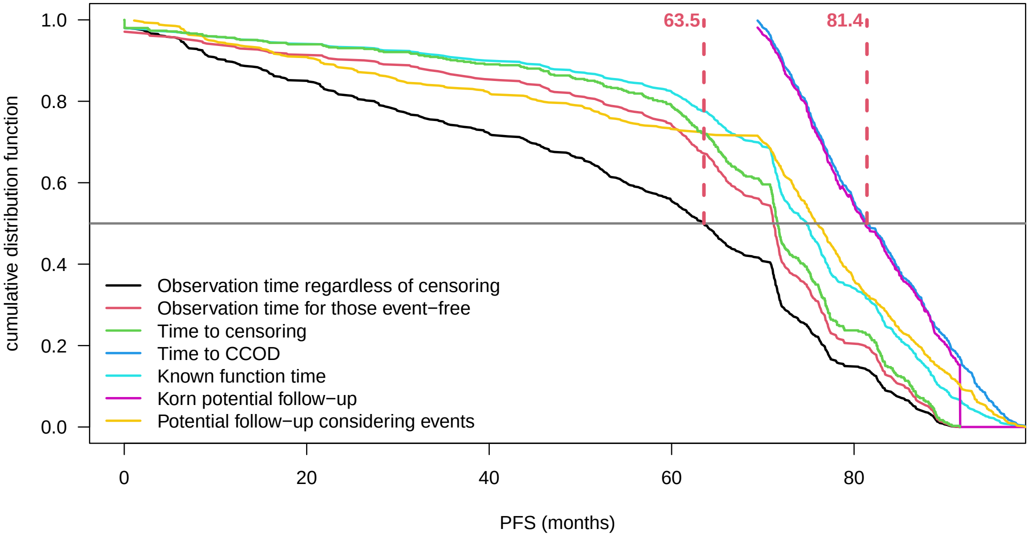
# Commonly used quantities

| Follow-up quantifier | Patient subset | Primary event | Censoring: administrative | Censoring: LTFU | CCOD |
|---|---|---|---|---|---|
| **Observation time regardless of censoring** | | event | event | event | ignored |
| **Observation time for those censored** | censored | ▨ | event | event | ignored |
| | event | excluded | | | |
| **Time to censoring** | | censored | event | event | ignored |
| **Time to CCOD, Potential follow-up** | | ignored | ignored | ignored | event |
| **Known function time** | censored | ▨ | event | event | ignored |
| | event | ignored | ▨ | ▨ | event |
| **Korn's potential follow-up time** | | Generalization of time to censoring, estimates P(under follow-up at $t$), distinguishes lost-to-follow-up and administrative censoring. | | | |
| **Potential follow-up considering events** | censored | ▨ | ignored | ignored | event |
| | event | event | ▨ | ▨ | ignored |

# One-sample case, second CCOD



# One-sample case, second CCOD



## What do trialists **really** want to know?

| Term | Question | How to best answer |
|---|---|---|

| Term | Question | How to best answer |
|---|---|---|
| **Precision** | How precise is estimate of underlying true survival function $S_X$? | Pointwise CIs or confidence bands. |

| Term | Question | How to best answer |
|---|---|---|
| **Precision** | How precise is estimate of underlying true survival function $S_X$? | Pointwise CIs or confidence bands. |
| **Reliability** | How far out to extend KM estimate? | Pointwise CIs or confidence bands. Disregard KM estimate if less than $m$ patients remain at risk. |

| Term | Question | How to best answer |
|---|---|---|
| **Precision** | How precise is estimate of underlying true survival function $S_X$? | Pointwise CIs or confidence bands. |
| **Reliability** | How far out to extend KM estimate? | Pointwise CIs or confidence bands. Disregard KM estimate if less than $m$ patients remain at risk. |
| **Stability** | How much can KM estimate possibly change in future data snapshot? | Consider all currently censored patients to either (1) have event day after censoring or (2) being censored at latest observed event time. Betensky (2015). |

| Term | Question | How to best answer |
|---|---|---|
| **Precision** | How precise is estimate of underlying true survival function $S_X$? | Pointwise CIs or confidence bands. |
| **Reliability** | How far out to extend KM estimate? | Pointwise CIs or confidence bands. Disregard KM estimate if less than $m$ patients remain at risk. |
| **Stability** | How much can KM estimate possibly change in future data snapshot? | Consider all currently censored patients to either (1) have event day after censoring or (2) being censored at latest observed event time. Betensky (2015). |
| **Information** | How much of information necessary to achieve targeted power for hypothesis test, either for milestone timepoint or median, has been collected? | Power depends on inverse of variance of parameter of interest. |

**No ”measure of follow-up” needed whatsoever!**

# Two-sample case

Why different from one-sample case?

# Two-sample case

Why different from one-sample case?

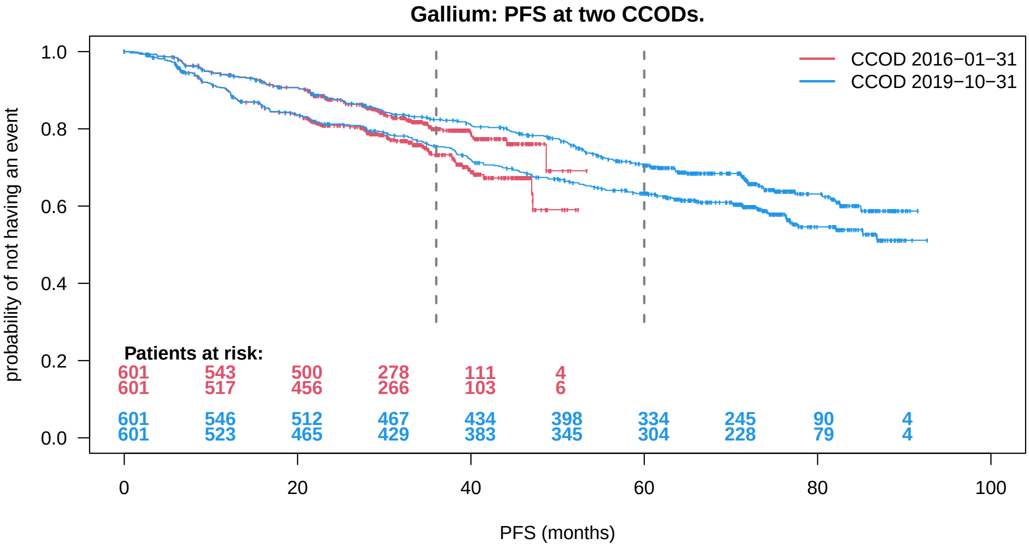
- Interest in **relative** effect.

# Two-sample case

Why different from one-sample case?

- Interest in **relative** effect.
- Proportional (PH) vs. non-proportional hazards (NPH).

# Two-sample case - PH



## What does power of logrank (any) test depend on **in general**?

**What does power of logrank (any) test depend on in general?**

**Inverse of variance of parameter of interest.**

**What does power of logrank (any) test depend on in general?**

**Inverse of variance of parameter of interest.**

**Only under PH and for unweighted logrank test proportional to #events!**

| Term | Question | How to best answer |
|---|---|---|
| **Precision** | How precise is HR estimate? | CI. |
| **Stability** | How much can HR estimate change in future data snapshot? | If PH assumption applies then estimate of HR will (on average) simply become more precise over time. |
| **Information** | How much of information necessary to achieve targeted power for hypothesis test for HR within group-sequential design has already been collected? | Information fraction $d_{\text{int}}/d_{\text{fin}}$. |
| **PH (reliability)** | Do hazard functions remain proportional? | Standard tools to assess PH, e.g. plot nonparametric estimates of (cumulative) hazard functions over time, and ratio thereof, or hypothesis tests. |
| **Censoring pattern** | Is censoring distribution same in both arms? Are distributions of censoring reasons same in both arms? | Plot nonparametric estimates of censoring distribution per arm, potentially split by censoring reason. |

**”We need enough FU for safety.”**

**”We need enough FU for safety.”**

**HTA assessments: use FU to assess ”Evidence base sufficient for evidence synthesis?”**

# Vague questions!

**Vague questions!**

**No FU-quantifier whatsoever can answer these questions!**

**Vague questions!**

**No FU-quantifier whatsoever can answer these questions!**

**Formulate questions precisely!**

**More follow-up is better in any case!**

**More follow-up is better in any case!**

**Depends on quantity of interest.**

# Conclusions

# Conclusions

- Follow-up quantifiers used in literature **highly heterogeneous**.

# Conclusions

- Follow-up quantifiers used in literature **highly heterogeneous**.
- Focus on **scientific question**, answer that using suitable quantities: precision, stability, information, assumptions for any quantity of interest.

# Conclusions

- Follow-up quantifiers used in literature **highly heterogeneous**.
- Focus on **scientific question**, answer that using suitable quantities: precision, stability, information, assumptions for any quantity of interest.
- No hope that any of these questions can be answered with one single number, however defined.

# Conclusions

- Follow-up quantifiers used in literature **highly heterogeneous**.
- Focus on **scientific question**, answer that using suitable quantities: precision, stability, information, assumptions for any quantity of interest.
- No hope that any of these questions can be answered with one single number, however defined.

PH vs. NPH:

# Conclusions

- Follow-up quantifiers used in literature **highly heterogeneous**.
- Focus on **scientific question**, answer that using suitable quantities: precision, stability, information, assumptions for any quantity of interest.
- No hope that any of these questions can be answered with one single number, however defined.

PH vs. NPH:

- Assumption matters for stability!

# Conclusions

- Follow-up quantifiers used in literature **highly heterogeneous**.
- Focus on **scientific question**, answer that using suitable quantities: precision, stability, information, assumptions for any quantity of interest.
- No hope that any of these questions can be answered with one single number, however defined.

PH vs. NPH:

- Assumption matters for stability!
- NPH: need to chose effect measure.

# Conclusions

- Follow-up quantifiers used in literature **highly heterogeneous**.
- Focus on **scientific question**, answer that using suitable quantities: precision, stability, information, assumptions for any quantity of interest.
- No hope that any of these questions can be answered with one single number, however defined.

PH vs. NPH:

- Assumption matters for stability!
- NPH: need to chose effect measure.
- Information depends on #events (PH) or many more quantities (NPH).

**Do not provide a quantification of follow-up.**

**Do not provide a quantification of follow-up.**

**Not only useless, but confusing.**

# Resources

Paper: https://arxiv.org/abs/2206.05216.

Markdown with all code: https://oncoestimand.github.io/quantFU/quantFU.html.

Oncology estimand WG: http://www.oncoestimand.org.

# Thank you for your attention.

**kaspar.rufibach@roche.com**
**http://go.roche.com/dss-mco**

**http://www.kasparrufibach.ch**
**numbersman77**
**numbersman77**

# References I

Betensky, R. A. (2015). Measures of follow-up in time-to-event studies: Why provide them and what should they be? *Clin Trials*, **12**(4), 403–408.

Shuster, J. J. (1991). Median follow-up in clinical trials. *J. Clin. Oncol.*, **9**(1), 191–192.

# Backup

| Term | Question | How to best answer |
|---|---|---|

| Term | Question | How to best answer |
|---|---|---|
| **Stability** | How much can HR estimate change in future data snapshot? | If PH assumption applies then estimate of HR will (on average) simply become more precise over time. |

| Term | Question | How to best answer |
|---|---|---|
| **Stability** | How much can HR estimate change in future data snapshot? | If PH assumption applies then estimate of HR will (on average) simply become more precise over time. |

- Exclusively relying on PH to ”predict” future course of the trial: **strong and statistically motivated assumption**.

| Term | Question | How to best answer |
|---|---|---|
| **Stability** | How much can HR estimate change in future data snapshot? | If PH assumption applies then estimate of HR will (on average) simply become more precise over time. |

- Exclusively relying on PH to ”predict” future course of the trial: **strong and statistically motivated assumption**.
- Drug development perspective: **stability** assessed also based on accumulating data, not ”only” on PH assumption.

| Term | Question | How to best answer |
|---|---|---|
| **Stability** | How much can HR estimate change in future data snapshot? | If PH assumption applies then estimate of HR will (on average) simply become more precise over time. |

- Exclusively relying on PH to ”predict” future course of the trial: **strong and statistically motivated assumption**.
- Drug development perspective: **stability** assessed also based on accumulating data, not ”only” on PH assumption.

Stability cannot be assessed using whatever measure of follow-up!

# Results for Gallium (PH)

# Precision, stability, information

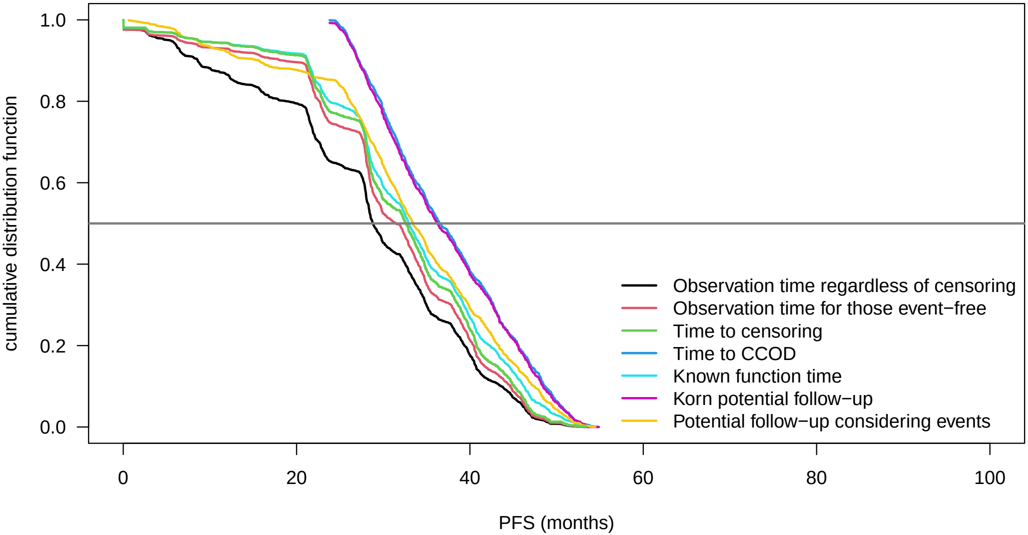
| | CCOD 2016-01-31 | CCOD 2019-10-31 |
|---|---|---|
| HR | 0.66 | 0.76 |
| 95% CI | [0.51, 0.85] | [0.62, 0.92] |
| Number of events $d$ | 245 | 419 |
| Proportion of patients with event | 20.4% | 34.9% |

Table: Key efficacy results for Gallium.

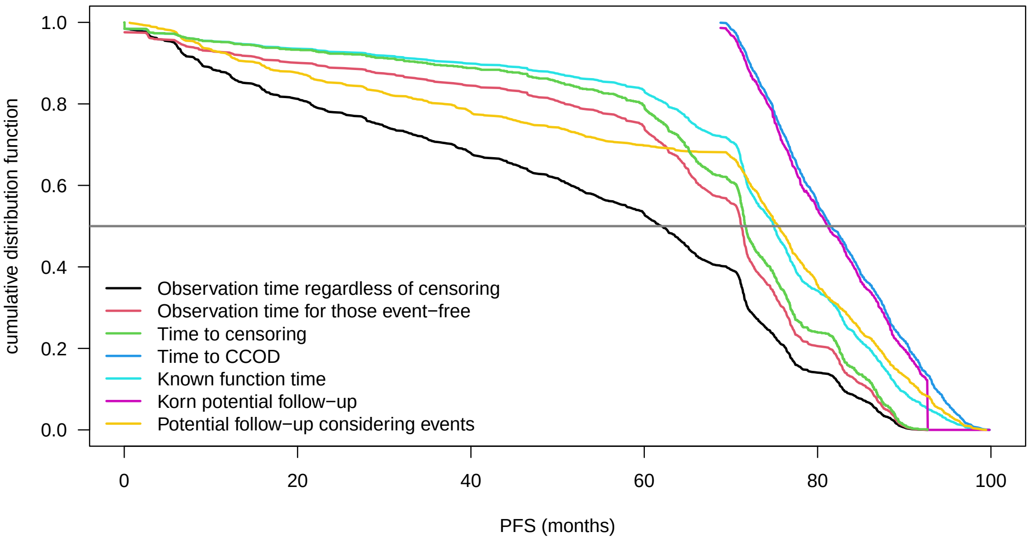
| milestone | treatment arm | CCOD 2016-01-31 | CCOD 2019-10-31 |
|---|---|---|---|
| 36 | Rituximab | 0.73 [0.66, 0.80] | 0.76 [0.71, 0.79] |
| 36 | Gazvya | 0.80 [0.73, 0.85] | 0.82 [0.79, 0.86] |
| 36 | Difference Obinutuzumab - Rituximab | 0.07 [0.01, 0.12] | 0.07 [0.02, 0.12] |
| 60 | Rituximab | - | 0.63 [0.58, 0.68] |
| 60 | Gazvya | - | 0.70 [0.65, 0.75] |
| 60 | Difference Obinutuzumab - Rituximab | - | 0.07 [0.02, 0.13] |

Table: Milestone KM estimates for Gallium.

# Quantification of follow-up, CCOD1



# Quantification of follow-up, CCOD2



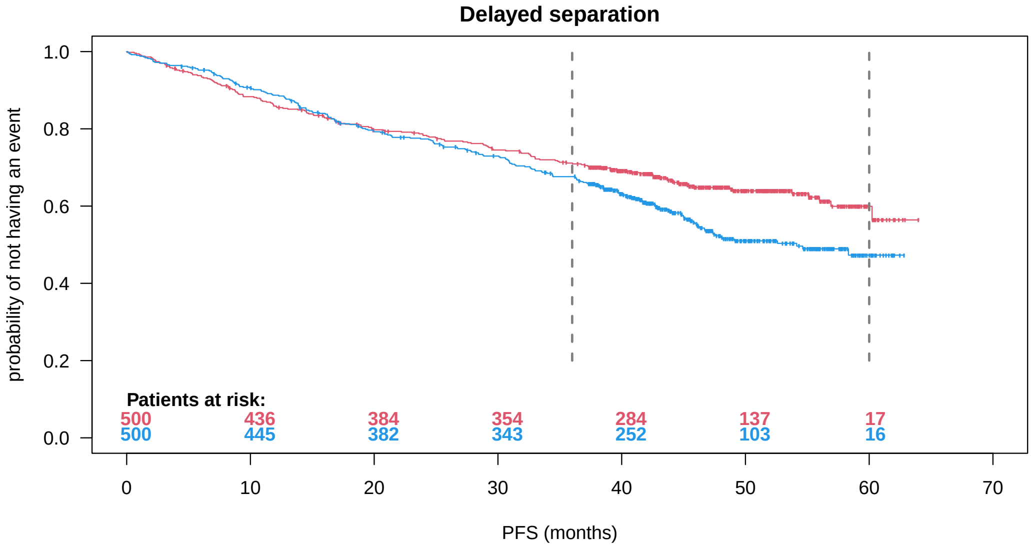
# Quantification of follow-up

| Quantity | CCOD 2016-01-31 | CCOD 2019-10-31 | $\Delta$ | $\Delta$% |
|---|---|---|---|---|
| Observation time regardless of censoring | 28.8 | 62.0 | 33.2 | +115% |
| Observation time for those event-free | 31.5 | 71.2 | 39.8 | +126% |
| Time to censoring | 32.6 | 71.7 | 39.1 | +120% |
| Time to CCOD | 36.5 | 81.5 | 45.0 | +123% |
| Known function time | 32.9 | 75.0 | 42.0 | +128% |
| Korn potential follow-up | 36.2 | 81.2 | 44.9 | +124% |
| Potential follow-up considering events | 33.4 | 75.5 | 42.1 | +126% |

Table: Different quantifications of follow-up for Gallium, in months.

# Delayed separation example

# Two-sample case - NPH: delayed separation



# NPH: need effect quantifier $\neq$ HR.

**NPH: need effect quantifier $\neq$ HR.**

**Milestone or median difference, RMST, variants of logrank test.**

**NPH: need effect quantifier $\neq$ HR.**

**Milestone or median difference, RMST, variants of logrank test.**

**Precision depends on inverse of variance of parameter of interest.**

**NPH: need effect quantifier $\neq$ HR.**

**Milestone or median difference, RMST, variants of logrank test.**

**Precision depends on inverse of variance of parameter of interest.**

**No hope #events tells us everything!**

# Illustration for RMST

Variance of RMST depends on:

- total number of patients,
- randomization ratio,
- KM estimate of the pooled sample,
- estimated censoring distribution in each arm (which can be taken as pooled if random censoring is assumed),
- observed number of events at $t_0$,
- observed number of patients still at risk at $t_0$.

| Term | Question | How to best answer |
|---|---|---|

| **Term** | **Question** | **How to best answer** |
|---|---|---|
| **Precision** | How precise is estimate to quantify effect of interest? | CI. |

| Term | Question | How to best answer |
|---|---|---|
| **Precision** | How precise is estimate to quantify effect of interest? | CI. |
| **Stability** | How much can effect estimate change in future data snapshot? | Look at extreme scenarios as proposed by Betensky (2015). |

| Term | Question | How to best answer |
|---|---|---|
| **Precision** | How precise is estimate to quantify effect of interest? | CI. |
| **Stability** | How much can effect estimate change in future data snapshot? | Look at extreme scenarios as proposed by Betensky (2015). |
| **Information** | How much of information necessary to achieve targeted power for hypothesis test for effect of interest has already been collected, if group-sequential design is used? | Not only related to #events / information fraction.<br>Effect measure specific. |

| Term | Question | How to best answer |
|---|---|---|
| **Precision** | How precise is estimate to quantify effect of interest? | CI. |
| **Stability** | How much can effect estimate change in future data snapshot? | Look at extreme scenarios as proposed by Betensky (2015). |
| **Information** | How much of information necessary to achieve targeted power for hypothesis test for effect of interest has already been collected, if group-sequential design is used? | Not only related to #events / information fraction.<br>Effect measure specific. |
| **PH** | Not applicable. | |

| Term | Question | How to best answer |
|---|---|---|
| **Precision** | How precise is estimate to quantify effect of interest? | CI. |
| **Stability** | How much can effect estimate change in future data snapshot? | Look at extreme scenarios as proposed by Betensky (2015). |
| **Information** | How much of information necessary to achieve targeted power for hypothesis test for effect of interest has already been collected, if group-sequential design is used? | Not only related to #events / information fraction.<br>Effect measure specific. |
| **PH** | Not applicable. | |
| **Censoring pattern** | Same as in PH scenario. | |

# Trial design

Assumptions:

- Base event rate: 0.012, corresponding to median time-to-event of 60 months.
- Piecewise exponential survival with no effect between 0 and 12 months, HR $=$ 0.65 thereafter.
- In both arms: probability LTFU follows exponential distribution calibrated such that probability amounts to 0.025 at 12 months. Corresponds to median time-to-LTFU of 329 months.
- After ramp-up of 6 months we recruit 42 patients / month until maximal number of 1000 patients.

# Trial design

Assumptions:

- Base event rate: 0.012, corresponding to median time-to-event of 60 months.
- Piecewise exponential survival with no effect between 0 and 12 months, HR $= 0.65$ thereafter.
- In both arms: probability LTFU follows exponential distribution calibrated such that probability amounts to 0.025 at 12 months. Corresponds to median time-to-LTFU of 329 months.
- After ramp-up of 6 months we recruit 42 patients / month until maximal number of 1000 patients.

CCOD after 389 events: power of

- 80.5% for unweighted logrank test,
- 69.7% using RMST difference based on KM estimates between arms, with data-driven restriction time $t_0$ of lower of two maximal observed times (events and censored) in each arm.
- Based on 10000 simulated trials.

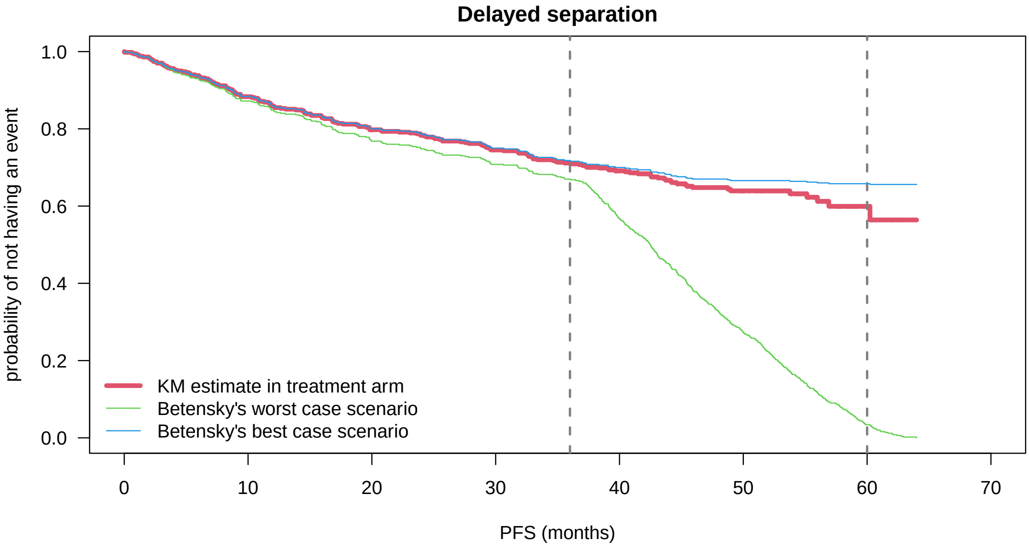
# Precision

| milestone | treatment arm | KM estimates and 95% CIs |
|---|---|---|
| 36 | Control arm | 0.68 [0.62, 0.73] |
| 36 | Treatment arm | 0.71 [0.66, 0.76] |
| 36 | Difference treatment - control | 0.04 [-0.03, 0.09] |
| 60 | Control arm | 0.47 [0.30, 0.65] |
| 60 | Treatment arm | 0.60 [0.46, 0.73] |
| 60 | Difference treatment - control | 0.13 [0.05, 0.21] |

Table: Milestone estimates for delayed separation example.

Difference of RMST between arms: 2.82 months between arms with 95% CI from -0.35 to 5.98.

# Stability


Delayed separation
probability of not having an event
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
70
PFS (months)
KM estimate in treatment arm
Betensky's worst case scenario
Betensky's best case scenario

*Doing now what patients need next*

**R version and packages used to generate these slides:**

R version: R version 4.1.1 (2021-08-10)

Base packages: stats / graphics / grDevices / utils / datasets / methods / base

Other packages: reporttools / xtable / rpact / survminer / ggpubr / survival / forcats / stringr / dplyr / purrr / readr / tidyr / tibble / ggplot2 / tidyverse

This document was generated on 2022-08-03 at 12:05:05.