Enhanced reconstruction of pseudo-individual patient data using quadratic programming

Andrew Titman

School of Mathematical Sciences, Lancaster University

16th June 2026

Outline

Lancaster University

- Why individual patient data (IPD) reconstruction?
- Constrained optimization formulation
- Applications and examples
- *CIFresolve* package

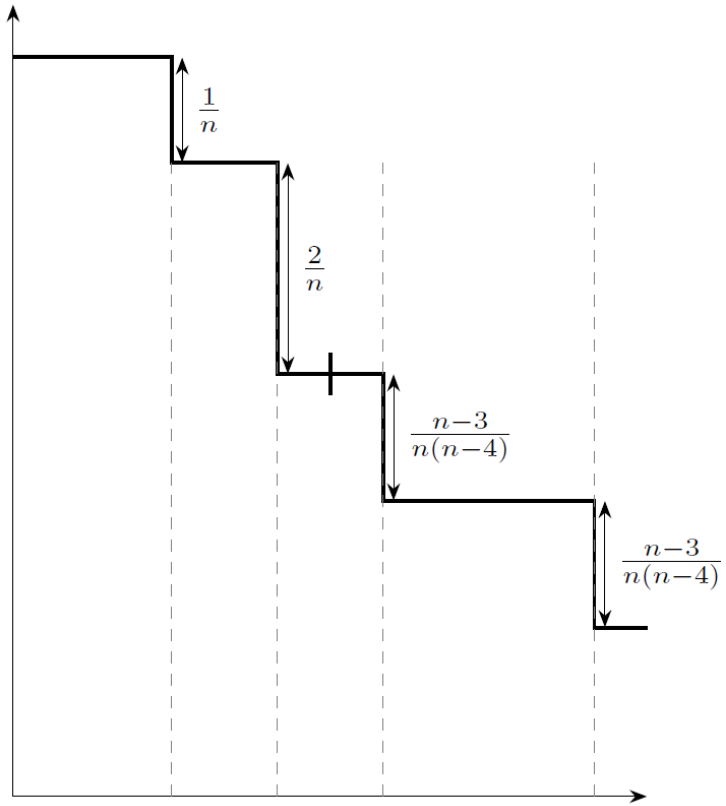
Reconstructing individual patient level data

Lancaster University

- Medical researchers frequently need to perform secondary data analysis of published studies
  - Meta-analysis/network meta-analysis
  - Indirect comparisons
  - Survival extrapolation for health technology assessment
- Published summary measures (e.g. point estimates and standard errors) may be insufficient or sub-optimal for the desired analysis
- Obtaining access to the original data may either be not possible or too time consuming

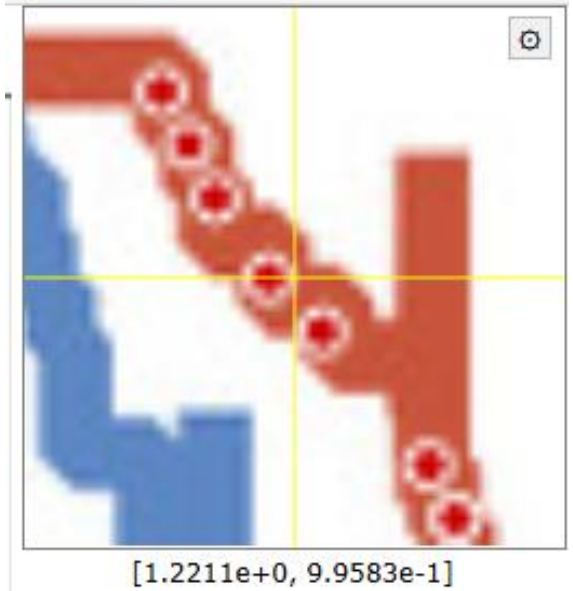
Reconstructing individual patient level data

- For survival data, the Kaplan-Meier estimate of the survivor function is almost universally reported for all treatment arms or exposure groups in a study.
- It has long been recognised that the Kaplan-Meier curve contains substantial information about the raw data in the study
  - Decrements of the curve tell us the unique uncensored event times
  - Size of decrements tell us something about the number of events and/or number at risk
  - The plot or paper may also give further information about the number at risk
- Many proposals for data reconstruction dating back to Parmar et al (1998)



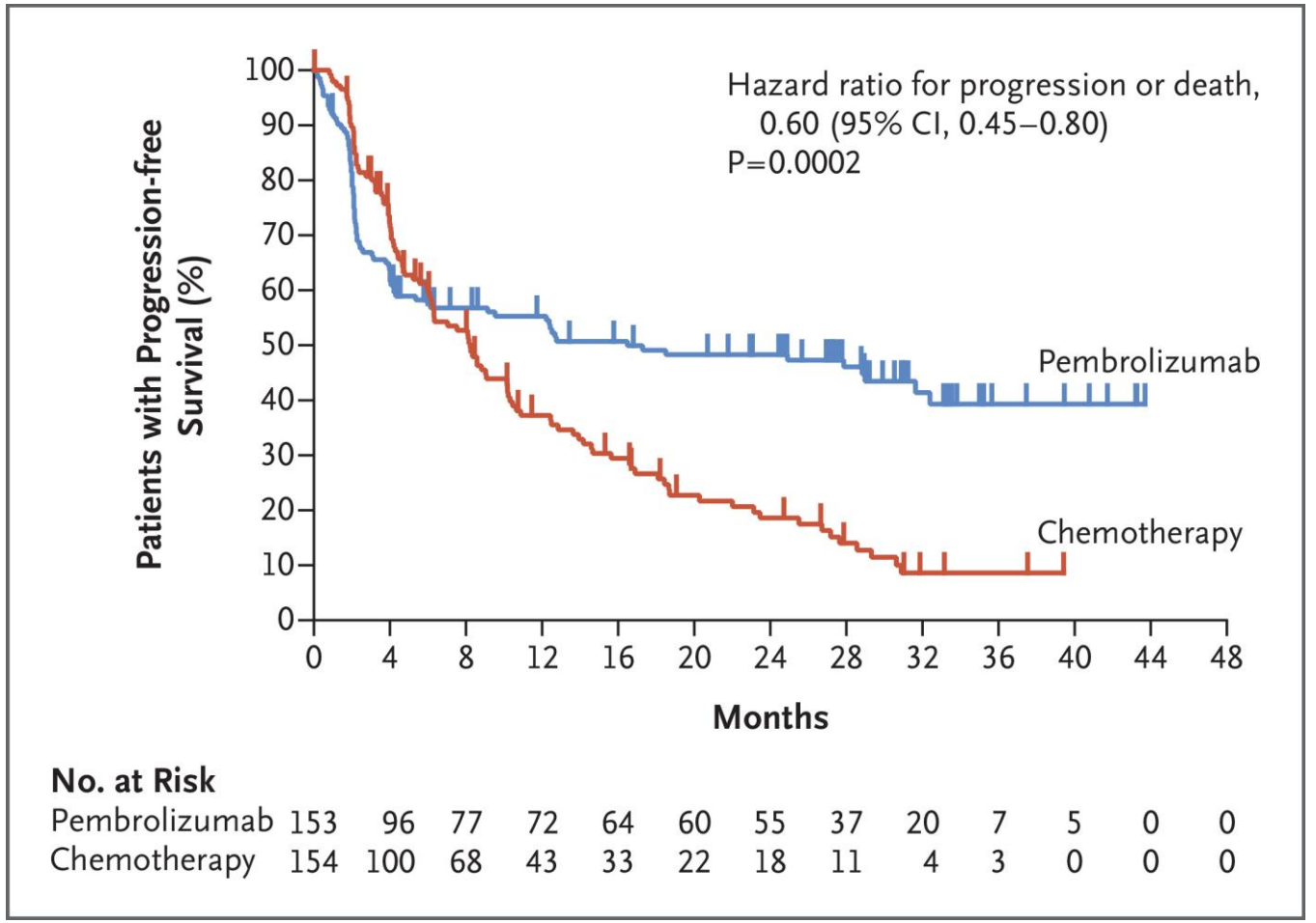
Digitizing for pseudo-IPD

- The location of the decrements can be determined by digitizing the curve
- Will necessarily introduce some degree of error
  - Misalignment of the time of the decrement
  - Misalignment of the survivor estimate at the decrement
  - Omission of decrements
- Most common existing algorithm is that of Guyot et al (2012)



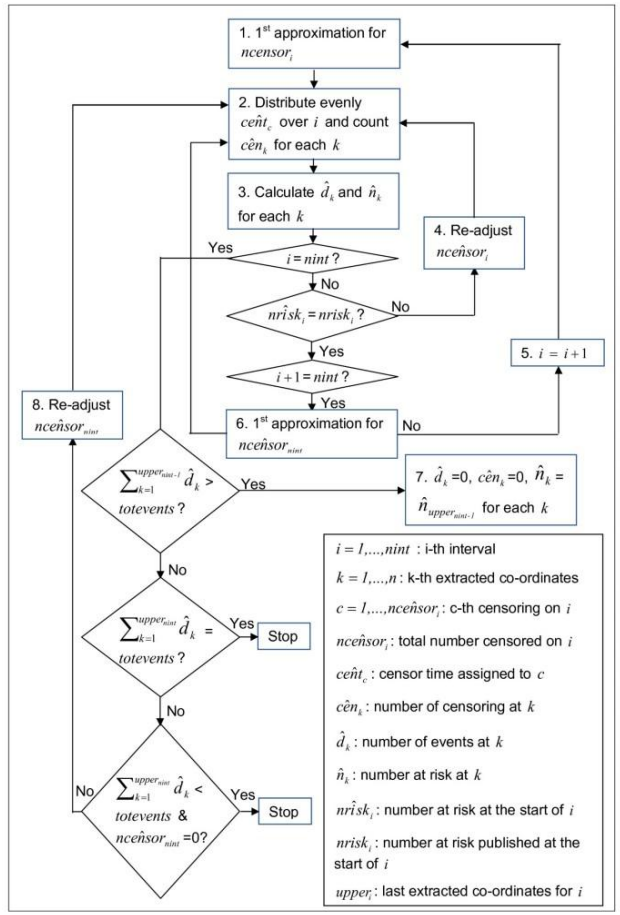
Example: PFS in KEYNOTE-177 trial

- Trial of patients with microsatellite-instability-high or mismatch-repair – deficient colorectal cancer
- Clear non-proportionality between pembrolizumab and chemotherapy
- Total number of events in each arm also reported



Issues with existing approaches

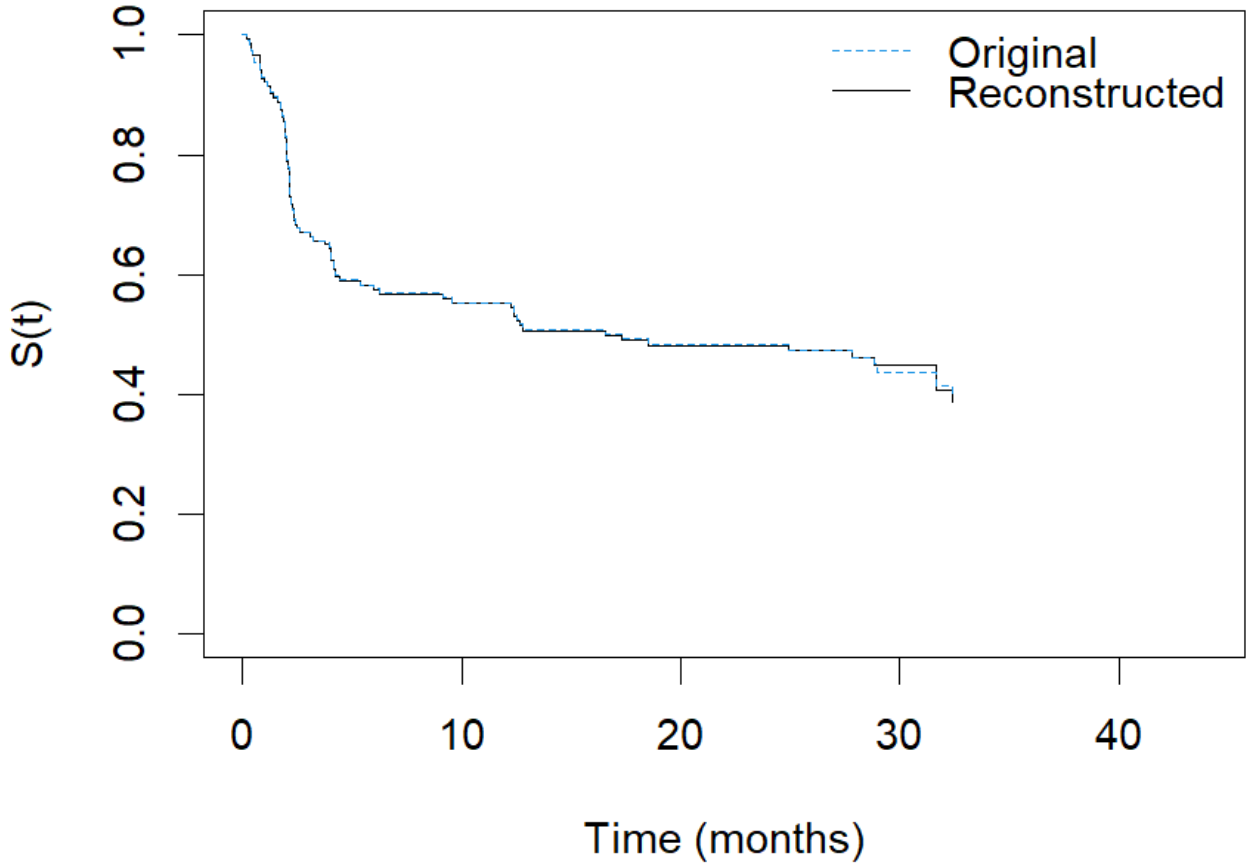
- Many Kaplan-Meier curves give information on the timing of censoring via marked censoring points or tick points
  - Is not incorporated into the Guyot et al method
  - Rogula et al introduced a method which uses the tick points but not the information on numbers at risk.
- Guyot et al method uses information on numbers at risk and number of events, but doesn't guarantee agreement
  - Algorithm complicated and seemingly not easy to extend to other cases



Guyot et al algorithm flowchart

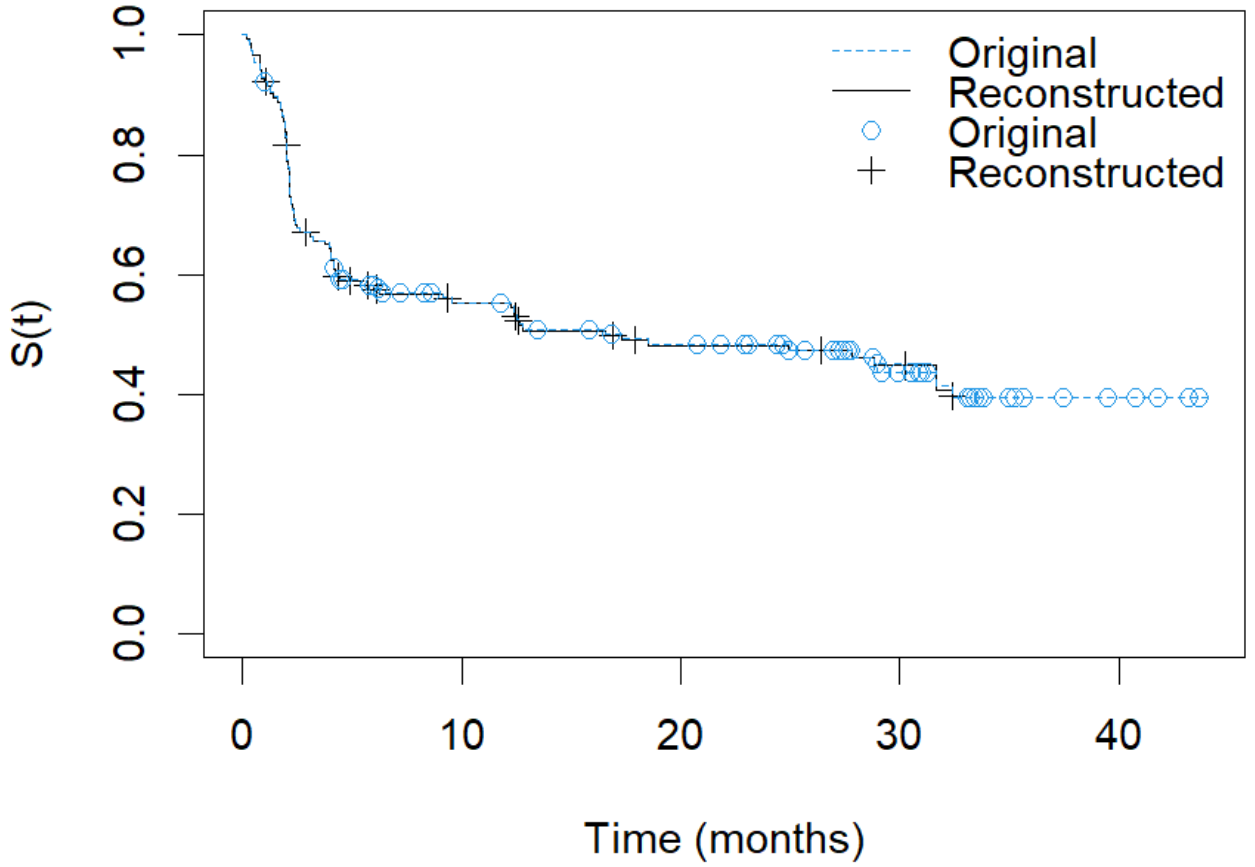
Example: PFS in KEYNOTE-177 trial

- Using the Guyot method via the IPDfromKM package in R
- Good agreement with KM curve



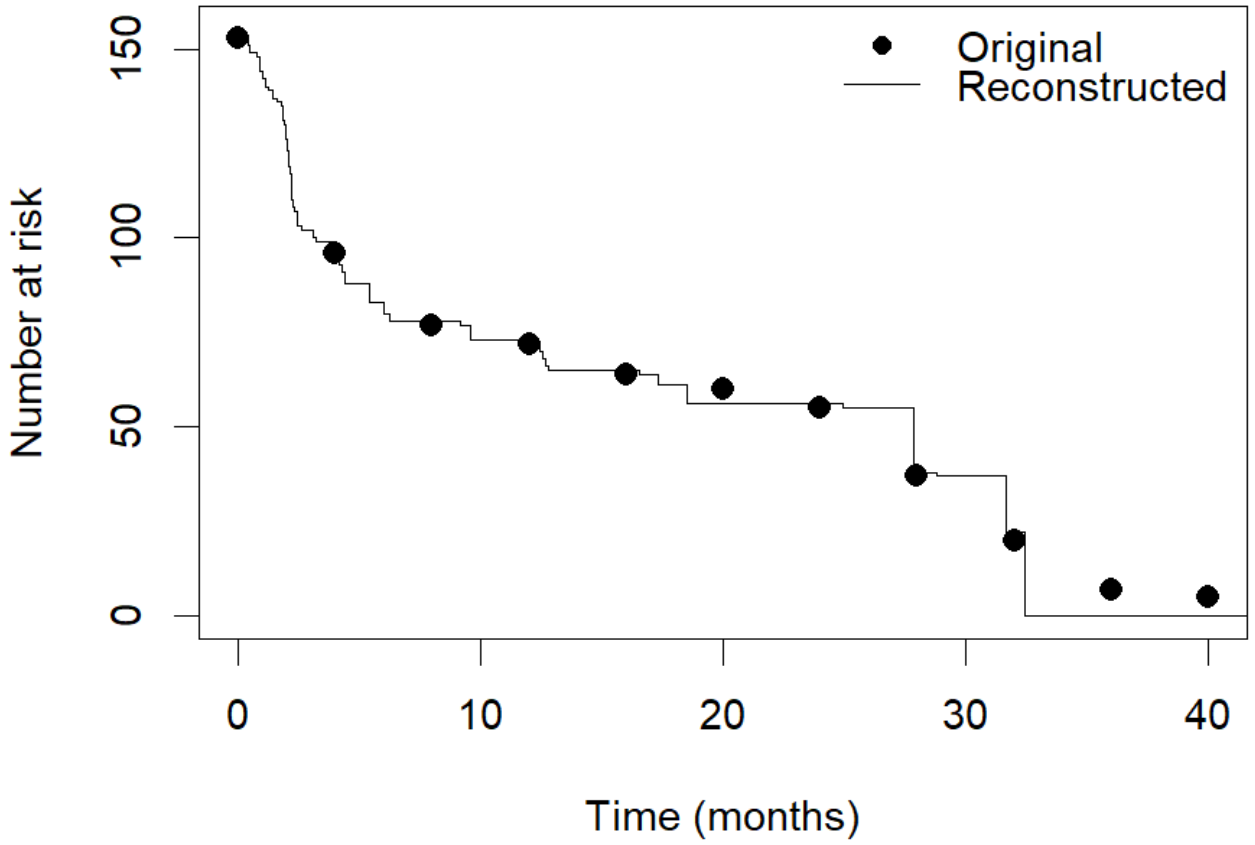
Example: PFS in KEYNOTE-177 trial

- Using the Guyot method via the IPDfromKM package in R
- Good agreement with KM curve
- Censoring times not correct



Example: PFS in KEYNOTE-177 trial

- Using the Guyot method via the IPDfromKM package in R
- Good agreement with KM curve
- Censoring times not correct
- Numbers at risk incorrect



Constrained optimization framework

Lancaster University

- Overall aim is to give a set of times $\{(t_i, \delta_i), i = 1, \ldots, n\}$ which replicates the Kaplan-Meier estimate as close as possible
- Digitization effectively gives two sources of information
  - Locations of decrements (subject to error)
  - Reported numbers at risk, number of events (should be exactly stated)
- Motivates framing the problem as a formal constrained optimization problem
- At each observed decrement point, $t_{(k)}$, observe $s_k = \hat{S}\left(t_{(k)}\right)$ and look to estimate
  - Number of events at that time (d_k)
  - Number of patients censored at that time (c_k)

Constrained optimization framework

Lancaster University

- Need to choose an appropriate measure of fit of the reconstruction
- Could try to use the discrepancy between the reconstructed and digitized Kaplan-Meier directly.
  - But the cumulative product of the Kaplan-Meier makes this unwieldly.
- However, if we use define $o_k := 1 - \frac{s_k}{s_{k-1}}$ can show that $o_k = \frac{d_k}{r_k}$
- Motivates trying to minimize

$$\sum_{k=1}^{K} \left(o_k \, \tilde{r}_k - \tilde{d}_k\right)^2$$

  subject to equality constraints e.g. $\tilde{r}_m = R_m$ for some m and $\sum_{k=1}^{K} d_k = n_D$
- This formulation ensures it is a *quadratic program* with linear constraints for which well-established algorithms and software exist

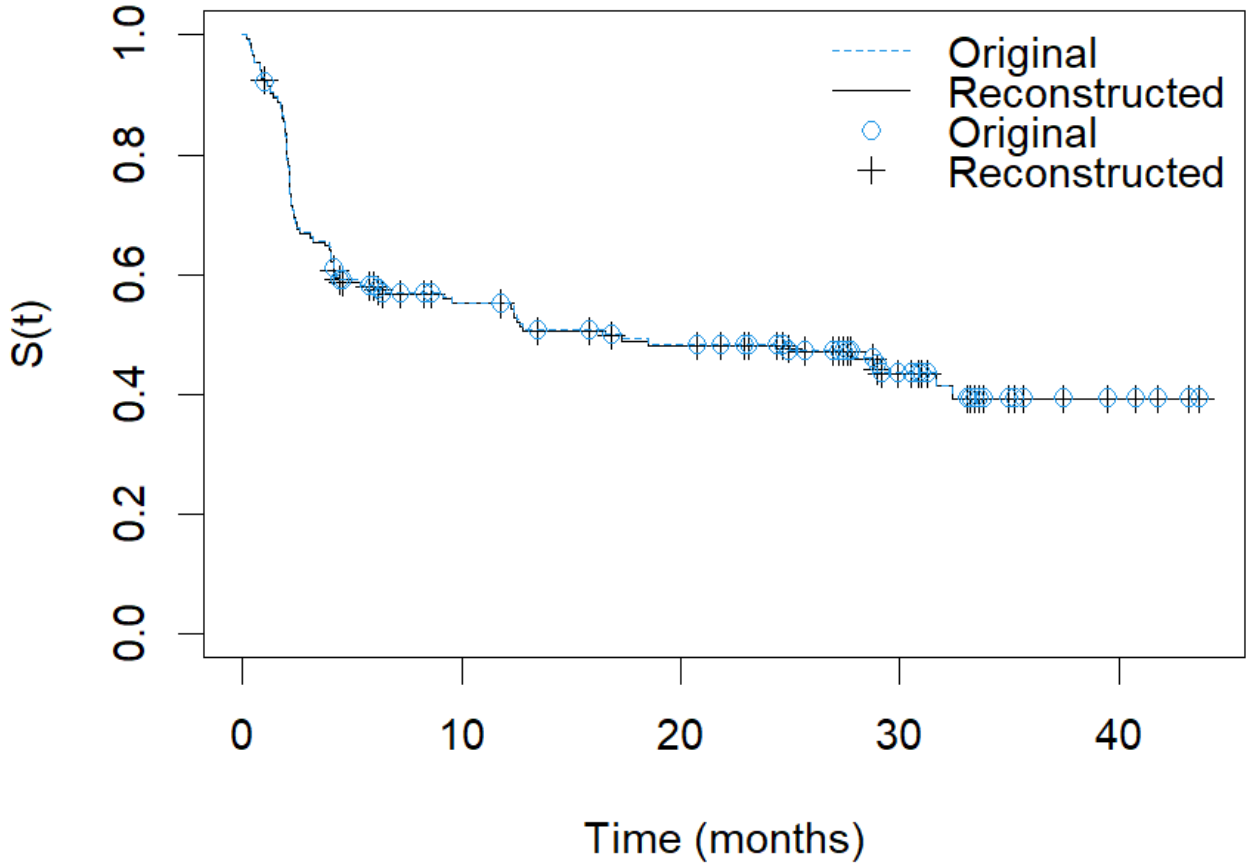
Constrained optimization framework

Lancaster University

- If marked censoring times are available, they can be added to the set of possible times $t_1, \ldots, t_k$ with the constraint that censoring can only occur at those times.
- Exact formulation is a *mixed-integer quadratic program* since number of events and number censored must take integer values
- Either use an approximate solution based on integer rounding of the continuous QP solution
    - Can use e.g. R package quadprog to do the quadratic programming.
- Or use a restricted amount of computation time to find a good solution
    - Requires commercial software like IBM CPLEX

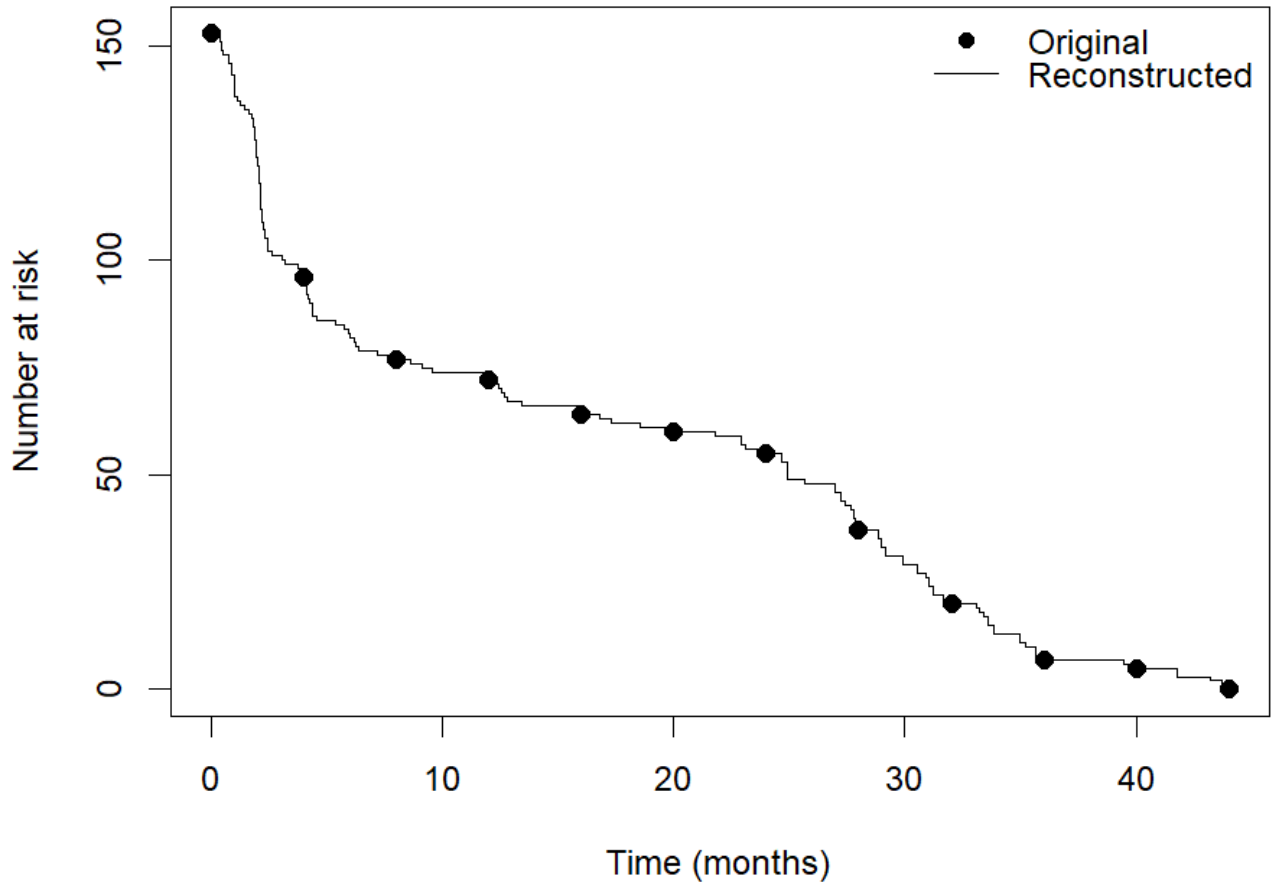
Example: PFS in KEYNOTE-177 trial

- Using the QP method, gives a similar visual fit to the Kaplan-Meier
- Censoring times directly align to those on the plot



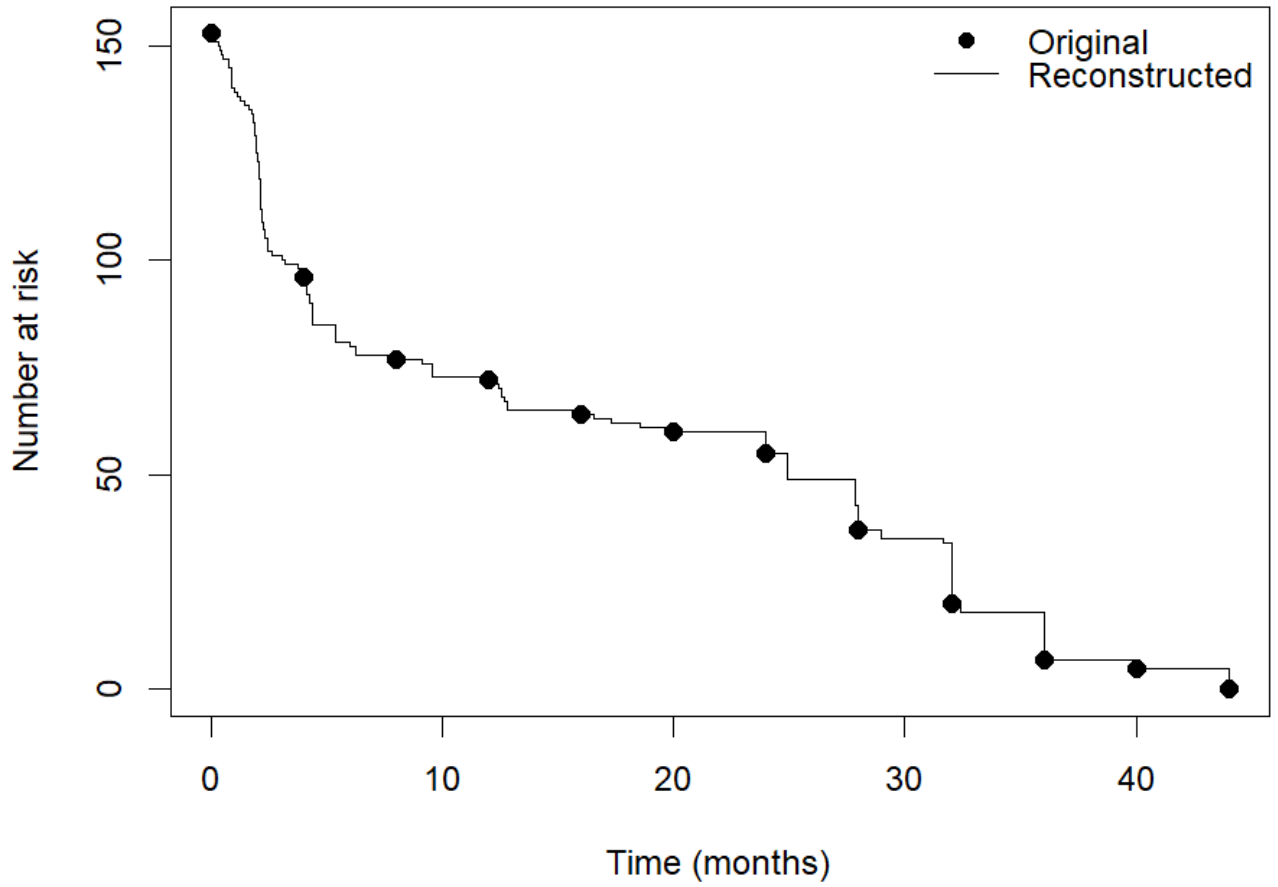
Example: PFS in KEYNOTE-177 trial

- Numbers at risk exactly agree with those reported on the Kaplan-Meier curve



Example: PFS in KEYNOTE-177 trial

- Numbers at risk exactly agree with those reported on the Kaplan-Meier curve
- Similar pattern if do not use the marked censoring times in the reconstruction.



Example: PFS in KEYNOTE-177 trial

Lancaster University

Method	Total patient time at risk
Original Data	2238.8
QP including marked times	2240.5
QP w/o marked times	2242.9
Guyot et al method	2139.2
Rogula et al method	2540.3

Simulation results

Lancaster University

- On simulated data find QP approaches are more accurate than Guyot et al method with respect to
  - The absolute fit to the Kaplan-Meier curve
  - The absolute fit to the true number at risk curve
  - Mean squared deviation from IPD fit of Weibull parametric model
- Difference between approximate QP and MIQP solutions quite modest
- Also outperforms Rogula et al method when only the marked censoring times are given
  - Can achieve high accuracy if the total number of events is known

Extensions

Lancaster University

- Total time at risk can be added as an additional constraint
- Method can be extended to allow reconstruction of pseudo-IPD from competing risks data using cumulative incidence curves
- Can be used to reconstruct the sufficient statistics for left-truncated survival data
  - Aggregate number of events at each time point
  - Aggregate number at risk at each time point
- Can be used to jointly reconstruct PFS and OS from same study in a way that is mutually consistent

CIFresolve package

Lancaster University

- R package available on Github
  https://github.com/andrewtitman/CIFresolve
- Sets up the necessary QP and then calls either solve.QP in the quadprog package or IBM CPLEX via the Rcplex package
- Accommodates standard survival, competing risks data and left-truncated survival data
- Produces both the aggregate summary table and pseudo-IPD dataset
- Provides goodness-of-fit plots for the reconstructed curve and the numbers at risk

References

- Guyot P, Ades AE, Ouwens MJ, Welton NJ. (2012) Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Medical Research Methodology. 12:1
- Parmar MKB, Torri V, Stewart L. (1998) Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Statistics in Medicine. 17: 2815—2834.
- Rogula B, Lozano-Oretga G, Johnston KM. (2022) A method for reconstructing individual patient data from Kaplan-Meier survival curves that incorporate marked censoring times. MDM Policy and Practice. 7(1).
- Titman AC (2026) Using quadratic programming to reconstruct data from published survival and competing risk analyses. Statistics in Medicine. DOI:10.1002/sim.70474