

The Special Structure of Time-to-Event Data

Many of the most consequential questions in Nigerian health research concern not just whether an event occurs but when: time from HIV diagnosis to antiretroviral therapy initiation, time from tuberculosis treatment start to cure or default, time from birth to a child's death, or time from schistosomiasis treatment to re-infection. These outcomes share a defining feature that ordinary regression cannot handle correctly: censoring. A participant who has not yet experienced the event by the end of a study (loss to follow-up, migration, or the study simply ending) contributes partial information — we know they survived at least that long — and discarding them (treating them as missing) or treating them as if the event never occurred both introduce bias. Survival analysis is the family of methods purpose-built to use this partial information correctly.

The Kaplan-Meier Estimator and the Log-Rank Test

The Kaplan-Meier estimator computes the non-parametric survival curve — the probability of remaining event-free beyond each observed event time — correctly incorporating censored observations, and it is almost always the first thing plotted in a survival analysis, before any regression model is fitted. Comparing

survival curves between two or more groups (for example, treatment adherence support versus standard care, in terms of time to viral suppression) uses the log-rank test, a non-parametric test of whether the underlying survival distributions differ, which is the time-to-event analogue of the two-sample tests introduced in Chapter 7.

The Cox Proportional Hazards Model

To adjust for multiple covariates simultaneously (age, sex, baseline CD4 count, facility) while modelling time-to-event data, the Cox proportional hazards model estimates a hazard ratio for each covariate — the multiplicative effect on the instantaneous risk of the event at any given time — without requiring the analyst to specify the baseline hazard’s functional form, which is what makes it the dominant model in applied survival analysis. Its key assumption, proportional hazards (that the hazard ratio for a covariate is constant over time), should be checked using Schoenfeld residuals; a covariate whose effect changes meaningfully over follow-up time (for example, a treatment effect that is strong initially but fades) requires a time-varying coefficient extension rather than a standard Cox model.

R example

```
# R: Cox model for time from ART initiation to viral suppression
library(survival)

cox_model <- coxph(Surv(time_to_suppression, suppressed) ~ age + sex +
  baseline_cd4 + adherence_support_arm,
  data = art_cohort)
summary(cox_model)      # hazard ratios
cox.zph(cox_model)      # test of proportional hazards assumption
```

Parametric Survival Models and Applications to NTD and Reinfection Data

Where the shape of the hazard over time is of direct scientific interest — for example, characterising the pattern of

schistosomiasis re-infection risk after praziquantel (PZQ) treatment, which is expected to be low immediately post-treatment and rise progressively with time since treatment as re-exposure to contaminated water accumulates — parametric survival models (exponential, for a constant hazard; Weibull, allowing the hazard to increase or decrease monotonically over time; log-normal or log-logistic, allowing a non-monotonic hazard) are preferred over the semi-parametric Cox model, because they yield an explicit, interpretable functional form for the hazard and allow smooth extrapolation and simulation of future risk, which is valuable in stepped-wedge and re-treatment-interval study designs of the kind used in Nigerian schistosomiasis control research.

Worked Example

Kaplan-Meier by Hand for a Small Cohort

Consider a small illustrative cohort of 8 patients followed for time (in months) to viral suppression, with ‘+’ denoting censored (suppression not yet observed): 2, 3+, 4, 5, 5, 7+, 8, 10. The Kaplan-Meier estimator computes the survival probability (here, probability of NOT yet suppressing) stepwise at each observed event time, multiplying the conditional probability of surviving past each event time by all prior conditional survival probabilities. At month 2 (1 event, 8 at risk): $S(2) = 7/8 = 0.875$. At month 4 (1 event, 6 at risk, since one was censored at month 3): $S(4) = 0.875 \times 5/6 = 0.729$. At month 5 (2 events, 5 at risk): $S(5) = 0.729 \times 3/5 = 0.438$. At month 8 (1 event, 2 at risk, since one was censored at month 7): $S(8) = 0.438 \times 1/2 = 0.219$. At month 10 (1 event, 1 at risk): $S(10) = 0.219 \times 0/1 = 0$. This step-by-step calculation is exactly what R’s ‘survfit()’ or Python’s ‘lifelines’ package automate, and working through it by hand once is the best way to understand why censored observations (months 3 and 7 here) still contribute information — they remain ‘at risk’ up to the point they are censored, they simply do not experience the event.

Figure 9.1 — Kaplan-Meier Curve for the 8-Patient Cohort (Worked Example, Sec. 9.5)

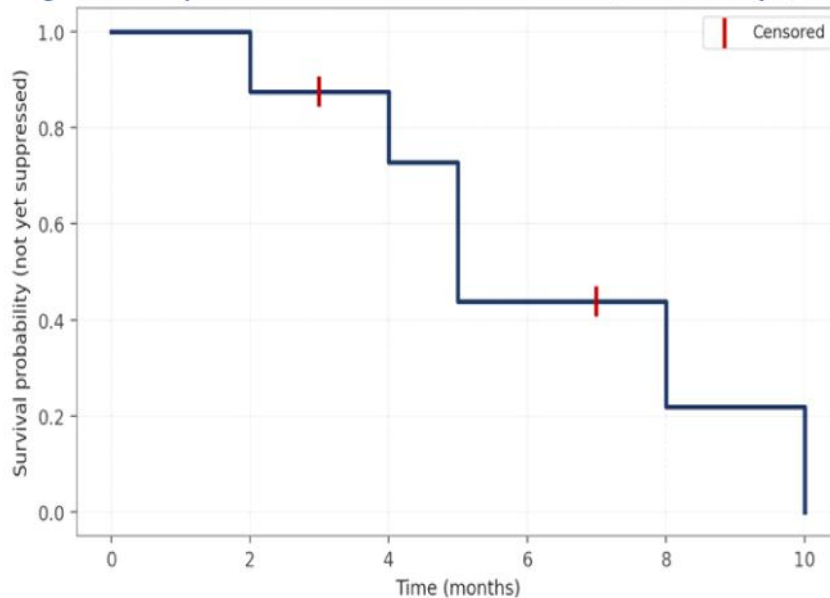


Fig. 9.1 — Kaplan-Meier survival curve for the 8-patient cohort (Section 9.5).

PYTHON example

```
# Python: Kaplan-Meier estimation and plotting with the lifelines package
from lifelines import KaplanMeierFitter
import pandas as pd

durations = [2, 3, 4, 5, 5, 7, 8, 10]
event_observed = [1, 0, 1, 1, 1, 0, 1, 1] # 1 = event, 0 = censored

kmf = KaplanMeierFitter()
kmf.fit(durations, event_observed, label="Time to viral suppression")
print(kmf.survival_function_)
kmf.plot_survival_function()
```

Worked Example**Interpreting a Cox Model and Checking Its Assumption**

A Cox model fitted to a tuberculosis treatment cohort yields a hazard ratio of 1.9 (95% CI 1.3-2.8) for baseline HIV co-infection, predicting time to treatment default. This means that, at any given point during treatment, HIV-co-infected patients default at 1.9 times the instantaneous rate of HIV-negative patients, adjusting

for other covariates in the model. A Schoenfeld residual test for this covariate returns $p = 0.41$ (non-significant), supporting the proportional hazards assumption; had this test returned $p = 0.01$, it would suggest the HIV co-infection effect on default risk changes meaningfully over the course of treatment (perhaps stronger early in treatment and weaker later), and a time-varying coefficient or a stratified Cox model would be a more appropriate specification than a standard Cox model reporting a single, constant hazard ratio.

R example

```
# R: full Cox model workflow with hazard ratio and proportional-hazards check
library(survival)

tb_model <- coxph(Surv(time_to_default, defaulted) ~ hiv_status + age + sex,
                  data = tb_cohort)
summary(tb_model)

ph_test <- cox.zph(tb_model)
print(ph_test)      # p-value near 1 supports the PH assumption for each covariate
plot(ph_test[1])    # visual check for the hiv_status covariate specifically
```

Competing Risks

When More than One Event Type is Possible

Standard survival analysis, as presented in this chapter, assumes a single event type of interest, with all other reasons for follow-up ending treated as censoring. This assumption can be problematic when a competing risk exists — for example, in a study of time to malaria treatment failure, a patient who dies from an unrelated cause can no longer experience treatment failure, and simply treating that death as an ordinary censoring event (as a standard Kaplan-Meier or Cox analysis would) can overstate the cumulative incidence of treatment failure, because it implicitly assumes the patient would eventually have failed treatment had they lived. The competing risks framework, using the cumulative incidence function and Fine-Gray subdistribution hazard models, correctly

accounts for this by explicitly modelling both event types together, and should be considered whenever a genuinely competing event (a different cause of the same kind of outcome, or death from an unrelated cause) is common in the study population.

Extended Case Study

Comparing Survival Across Three Treatment Eras

A common and methodologically rich application of the survival methods in this chapter, well beyond a simple two-arm treatment comparison, is comparing outcomes across successive calendar-time treatment ‘eras’ as a health programme’s protocols evolve — for example, comparing time-to-viral-suppression among ART patients enrolled during three successive eras of Nigeria’s national HIV treatment guidelines, each associated with different first-line regimens and different service delivery models (facility-based only, versus later differentiated/community-based models). A naive comparison of raw suppression rates across the three eras risks conflating genuine treatment or service-delivery improvements with two confounding calendar-time trends: patients enrolled in the most recent era have, by construction, shorter maximum possible follow-up time than patients from the earliest era (since the study’s end date is fixed), and background factors unrelated to treatment (improving overall health system capacity, changing patient risk profiles as earlier, sicker at-risk populations are progressively already enrolled in prior eras) also shift over calendar time.

A defensible analysis instead fits a single Cox model across the pooled multi-era cohort, with treatment era included as a covariate (and, where clinically plausible, tested for interaction with baseline CD4 count, since a new regimen’s benefit might differ for patients presenting with more advanced immunosuppression), correctly using each patient’s actual follow-up time and censoring status rather than a simple end-of-study

snapshot proportion, and checks the proportional hazards assumption for the era variable specifically, since it is entirely plausible that an early treatment-era advantage from a genuinely superior regimen might narrow over time as later-era service delivery improvements (which benefit all patients, including those measured for the era comparison) diffuse across the health system. Reporting a hazard ratio for each later era relative to the earliest, together with a check of whether these hazard ratios remain stable across the follow-up period, gives a far more defensible basis for attributing a genuine improvement (or lack thereof) to the guideline changes than a simple before-and-after comparison of raw suppression percentages ever could.

Choosing the Right Time Origin

An easily overlooked design decision in any survival analysis is the choice of time origin — the point defined as time zero, from which follow-up time is measured. For an ART cohort, time zero might reasonably be defined as date of diagnosis, date of treatment initiation, or date of enrolment into a specific study cohort, and these choices are not interchangeable: measuring time to viral suppression from date of diagnosis, when there is often a substantial and variable delay before treatment actually begins, conflates two very different processes (time to treatment initiation, and time from treatment initiation to suppression) into a single duration, and a covariate that predicts one of these sub-processes well may show a diluted or even reversed association with the combined duration. Selecting the time origin that most precisely matches the specific biological or programmatic process under study — most often treatment initiation, when the question of interest is specifically about how quickly treatment achieves suppression — is a design decision that should be made deliberately and stated explicitly, not defaulted to whichever date happens to be most conveniently available in the dataset.

Frailty Models

Extending Cox Regression to Grouped Survival Data

Just as Section 8.12 introduces mixed-effects models to handle grouped data in ordinary regression, the survival analysis methods of this chapter have an analogous extension for grouped time-to-event data: a frailty model adds a random effect (the ‘frailty’ term) to a Cox proportional hazards model, allowing the model to represent unobserved, shared risk factors within a group — for example, patients treated at the same facility potentially sharing an unmeasured facility-level quality-of-care factor that affects all of that facility’s patients’ time to treatment default similarly. Ignoring this shared, facility-level frailty when it is genuinely present, by fitting a standard Cox model directly to individual-level, facility-clustered survival data, risks the same kind of understated standard errors and potentially misleading hazard ratio estimates that ignoring clustering produces in the ordinary regression and survey-inference contexts discussed elsewhere in this book, making a frailty model (or, alternatively, a Cox model stratified by facility) the more defensible choice whenever a meaningful number of patients are nested within a meaningful number of distinct facilities or treatment sites.

Landmark Analysis

A Practical Alternative to Time-Varying Covariates

When a covariate of interest is itself measured partway through follow-up (for example, viral load suppression status at 6 months, as a predictor of subsequent mortality), including it naively as a baseline covariate in a standard Cox model introduces immortal time bias, since patients must, by definition, have survived to 6 months to have this covariate measured at all. Landmark analysis addresses this by selecting a fixed landmark time point (here, 6 months), restricting the analysis to patients who are still under observation and event-free at that point, and then analysing survival

from the landmark time forward using the covariate's value at that landmark — a simpler and more transparent alternative to a full time-varying covariate Cox model, and one well suited to the kind of applied clinical cohort analysis common in Nigerian HIV and tuberculosis treatment research.

Multi-State Models

Beyond a Single Terminal Event

The survival models presented so far in this chapter track a single transition from an initial state to one terminal event. Many Nigerian clinical pathways instead involve multiple, sequential states — a tuberculosis patient may move from treatment initiation, to sputum conversion, to treatment completion, to either cure or relapse, each transition potentially depending on time already spent in the current state and on covariates that may themselves change over the pathway. Multi-state models extend the Cox framework to model each possible transition between states simultaneously (typically via separate, transition-specific hazard models sharing the same underlying data structure), providing a considerably richer picture of a patient's full clinical trajectory than a single time-to-terminal-event analysis, at the cost of requiring more detailed, accurately-timestamped data on every intermediate state transition.

A Reusable Survival Analysis Function

Combining Sections 9.3 and 9.6 into a single function ensures the proportional hazards check is never accidentally omitted when a Cox model is fitted — a real risk when the model-fitting and assumption-checking code live in two separate, easily-separated code cells rather than a single function call that returns both results together, flagging clearly whenever the assumption check fails so the analyst cannot overlook it.

R example

```
# R: fitting a Cox model and its proportional-hazards check in a single call
library(survival)

fit_cox_with_ph_check <- function(formula, data, alpha = 0.05) {
  model <- coxph(formula, data = data)
  ph_test <- cox.zph(model)
  failing_covariates <- rownames(ph_test$table)[ph_test$table[, "p"] < alpha]
  failing_covariates <- setdiff(failing_covariates, "GLOBAL")
  if (length(failing_covariates) > 0) {
    warning("Proportional hazards assumption violated for: ",
            paste(failing_covariates, collapse = ", "))
  }
  list(model = model, ph_test = ph_test, ph_violations = failing_covariates)
}

result <- fit_cox_with_ph_check(Surv(time_to_default, defaulted) ~ hiv_status + age,
tb_cohort)
```

Chapter Summary

- Censoring is the defining feature of time-to-event data; survival analysis methods use partial information from censored observations correctly, unlike ordinary regression.
- The Kaplan-Meier estimator provides a non-parametric survival curve; the log-rank test compares survival between groups.
- The Cox proportional hazards model adjusts for multiple covariates via hazard ratios without specifying the baseline hazard shape; the proportional hazards assumption must be checked with Schoenfeld residuals.
- Parametric survival models (Weibull, log-normal, etc.) are preferred when the shape of the hazard itself is scientifically important, such as post-treatment reinfection risk.
- Working through a Kaplan-Meier calculation by hand clarifies why censored observations still contribute information up to their censoring time.

Mini-Project**Mini-Project 9: A Complete Survival Analysis Workflow**

Analyse a time-to-event dataset (real or simulated) from raw data through Kaplan-Meier curves to a fully checked Cox model.

Step 1. Simulate or obtain a dataset of at least 100 subjects with a time-to-event outcome, a censoring indicator, and at least two covariates (one categorical, one continuous).

Step 2. Produce and interpret a Kaplan-Meier survival curve stratified by the categorical covariate, including a log-rank test.

Step 3. Fit a Cox proportional hazards model adjusting for both covariates, and report hazard ratios with 95% confidence intervals.

Step 4. Check the proportional hazards assumption using Schoenfeld residuals for each covariate, and report the result.

Step 5. Write a short paragraph interpreting the hazard ratios in plain language, appropriate for a technical report to a disease programme.

Deliverable: *A Kaplan-Meier plot, log-rank test result, Cox model summary table, proportional-hazards check, and a half-page interpretation.*

End-of-Chapter Exercises

1. Explain why treating a participant lost to follow-up at 6 months as if they were event-free for the whole study period would bias a survival analysis.
2. A Cox model shows a hazard ratio of 0.6 for a treatment arm. Interpret this in words, including the direction of the effect on risk.
3. Why might a Weibull model be preferred over a Cox model when characterising the time course of schistosomiasis reinfection risk after treatment?
4. Using the by-hand Kaplan-Meier calculation in Section 9.5, extend the cohort by one additional patient who is censored at month 6, and recompute $S(8)$ and $S(10)$.
5. In the worked example of Section 9.6, explain what a significant Schoenfeld residual test ($p = 0.01$) would imply about the appropriateness of reporting a single hazard ratio for HIV co-infection across the whole follow-up period.

6. Using the Python code in Section 9.5, modify the ‘durations’ and ‘event_observed’ lists to add two more censored patients at months 9 and 11, and describe (without necessarily running the code) how you would expect the tail of the survival curve to change.
7. A colleague proposes simply excluding all censored patients from a survival analysis to ‘simplify the data.’ Explain, in two to three sentences, what would be lost or biased by doing this.
8. Explain, using the malaria treatment failure example in Section 9.7, why treating an unrelated death as ordinary censoring can overstate the true cumulative incidence of treatment failure.
9. In what circumstance would a standard Kaplan-Meier analysis (ignoring competing risks) still give an approximately correct answer despite the presence of a competing risk?
10. Name one other Nigerian health research context (besides malaria treatment failure) where competing risks would need to be considered.
11. In the Section 9.8 case study, explain why comparing raw end-of-study suppression percentages across the three treatment eras would be misleading, even before considering any treatment differences.
12. Why is testing an interaction between treatment era and baseline CD4 count a reasonable thing to explore in this analysis, given what is known clinically about ART regimens?
13. Explain, in your own words, what it would mean if the proportional hazards check showed the era effect was NOT constant over follow-up time in this study.

14. Propose one additional covariate (beyond era and CD4 count) you would want included in the Cox model described in Section 9.8, and explain why.
15. Why does using each patient's actual follow-up time and censoring status give a more accurate era comparison than a simple 'percentage suppressed by end of study' snapshot?
16. Using Section 9.9, explain why measuring time from date of diagnosis (rather than date of treatment initiation) could dilute a covariate's true association with time to viral suppression.
17. Propose a research question for which date of diagnosis WOULD be the more appropriate time origin, rather than date of treatment initiation.
18. Why should the choice of time origin be made deliberately and stated explicitly, rather than defaulting to whichever date is most conveniently available?
19. Explain, in your own words, how combining two different sub-processes (time to treatment initiation, and time to suppression after initiation) into one duration could obscure a genuine finding.
20. If a dataset only records date of diagnosis and final outcome, with treatment initiation date missing for many patients, what would you consider doing before proceeding with a survival analysis?
21. Using Section 9.10, explain what a frailty term represents in a Cox model fitted to facility-clustered survival data.
22. Why would ignoring shared, facility-level unobserved risk factors in a survival analysis risk the same kind of problem as ignoring clustering in ordinary regression (Section 8.12)?

23. What is one alternative to a frailty model, mentioned in Section 9.10, for handling facility-clustered survival data?
24. Propose a scenario (beyond ART treatment default) where a frailty model would be an appropriate extension of a standard Cox model.
25. Explain, in your own words, why ‘a meaningful number of patients nested within a meaningful number of distinct facilities’ is the condition that makes a frailty model worth considering.
26. Summarise, in your own words, why censoring is described as the defining structural feature that separates survival analysis from ordinary regression.
27. Explain the common thread linking the competing risks discussion (Section 9.7) and the frailty models discussion (Section 9.10).
28. Why does this chapter emphasise choosing the correct time origin (Section 9.9) as a design decision, not a minor technicality?
29. Propose a time-to-event research question from your own field, and identify which extension (competing risks, frailty, or careful time-origin choice) would be most relevant to it.
30. If you had to explain to a colleague why ‘just excluding the censored patients’ is a bad idea, in one sentence, what would you say?
31. Using Section 9.12, describe a Nigerian clinical pathway (other than tuberculosis treatment) that could be modelled as a multi-state process.
32. Why does a multi-state model require more detailed data than a standard single-event survival analysis?

33. Explain how a multi-state model could reveal information that a single time-to-terminal-event analysis would miss.
34. What practical data collection challenge would you anticipate in accurately timestamping every intermediate state transition in a real clinical cohort?
35. How does the multi-state framework relate conceptually to the competing risks framework introduced earlier in this chapter (Section 9.7)?