# Matching

## Nonparametric Survival Comparison with KM

Jih-Chang Yu

October 6, 2025

# Context: Modeling vs. Nonparametric Approach

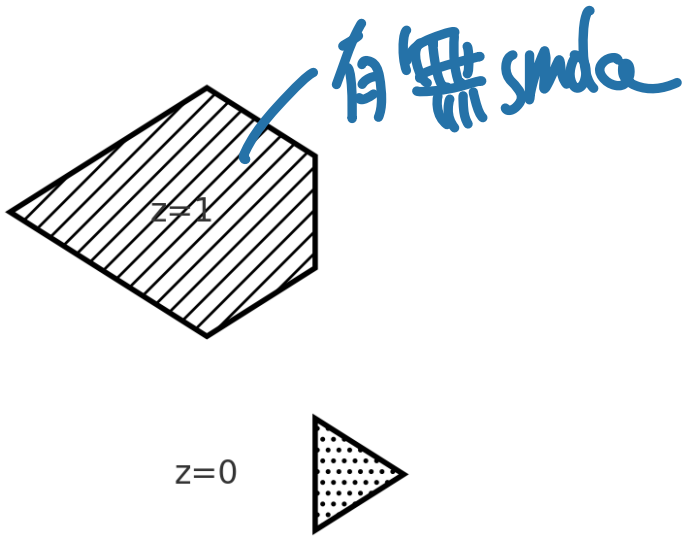
- Target: the **survival curve under a specific treatment**.
- Observational data: treatment assignment depends on covariates $\Rightarrow$ naive KM shows **associations**.
- Cox PH offers efficient modeling but needs assumptions; here we emphasize a **design-first** route via KM + matching.

# Kaplan–Meier and the Need for Matching

- KM is **nonparametric**; differences across $Z$ may reflect both treatment and baseline imbalance.
- Goal: approximate a **counterfactual** comparison by balancing observed covariates before KM.
- **Matching** helps create comparable groups on $X$ so KM reflects treatment rather than baseline differences.

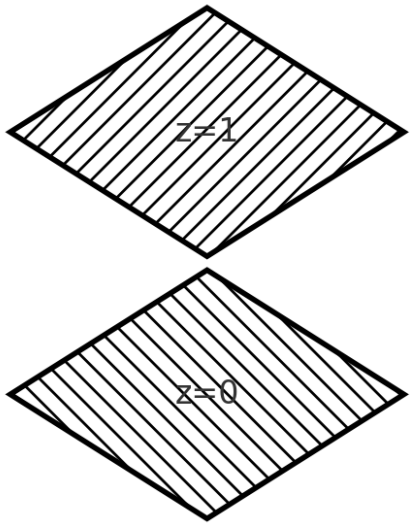
# Association vs. Counterfactual (Design Idea)

**Association (Observational study)**




Observed: different shapes across z;
triangle is the piece cut from the diamond.

**Causal (Counterfactual outcomes)**

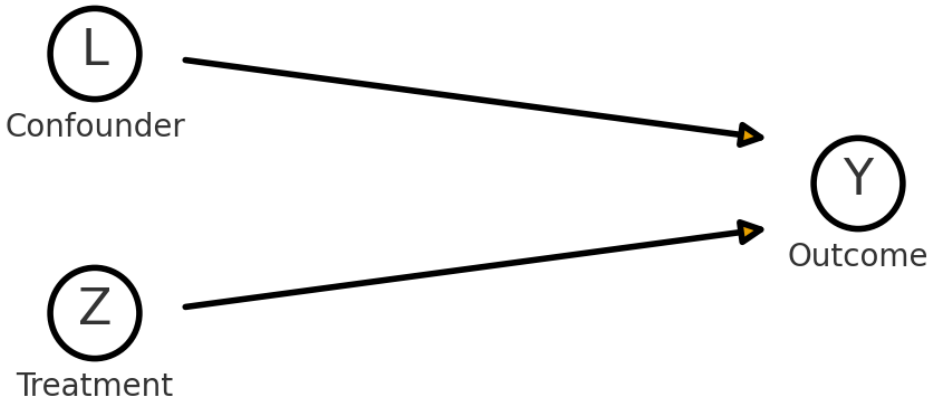



Counterfactual: identical shapes for z=1 and z=0.

# Association vs. Counterfactual Outcomes

- **Left (observational)**: Top is a diamond with a wedge removed ($z$=1); bottom is the removed triangle ($z$=0). Different shapes reflect covariate imbalance—an *association*, not causal.
- **Right (counterfactual/balanced)**: Two *identical* diamonds ($z$=1 on top, $z$=0 below). After balancing, the contrast targets a *causal effect*.
- **Takeaway**: Balance first, then compare KM on the balanced (matched) sample.

# DAG: Prognostic vs. Confounder

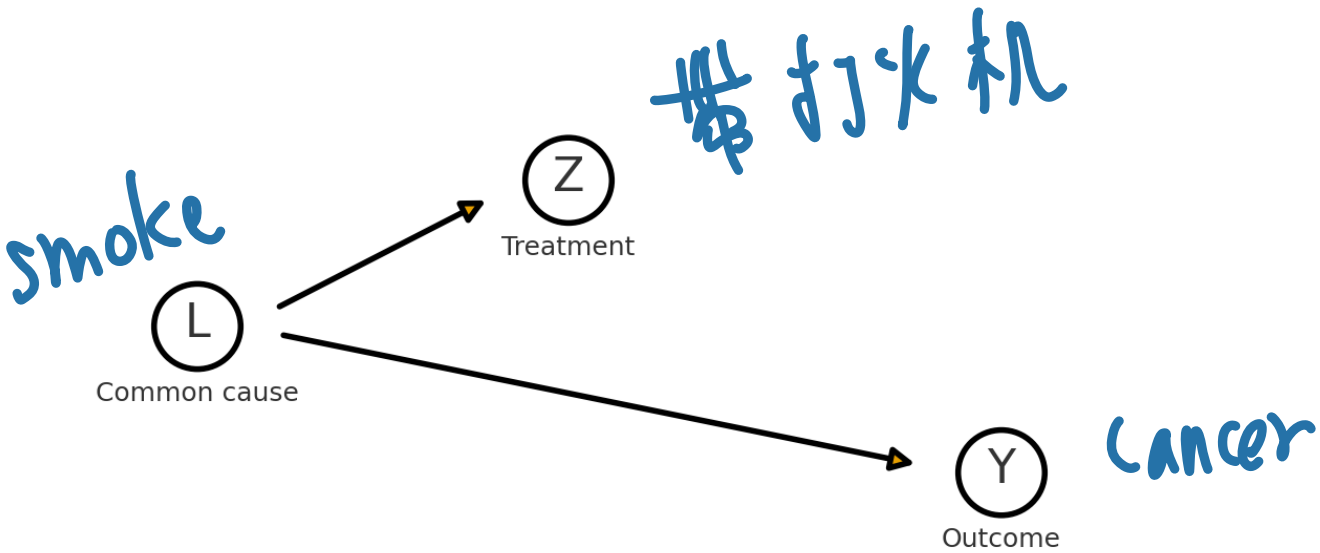



- $L \to Y$, $Z \to Y$, no $L \to Z$: $L$ is **prognostic**, not a confounder.
- Omitting $L$ does **not** bias $Z \to Y$; adjusting $L$ can improve precision.

# DAG: $L \to Z \to Y$ (No $L \to Y$)


- $L \to Z$, $Z \to Y$, no $L \to Y$: $L$ is **not** a confounder of $Z \to Y$.
- Omitting $L$ does not bias $Z \to Y$; variables predicting $Z$ but not $Y$ need not be adjusted.

# DAG: $L \to Z$, $L \to Y$ (Spurious Association if $L$ Unadjusted)




- $L$ is a **common cause** of $Z$ and $Y$: $L \to Z$, $L \to Y$; no $Z \to Y$.
- **Without adjusting** $L$, $Z$–$Y$ association is **spurious**; block $Z \leftarrow L \to Y$ by matching/weighting/stratification.

# From $L$ to $X$: We Do Not Pre-Label Confounders

- In practice we **do not know** which covariates are true confounders.
- Use $X$ to denote all observed covariates; target **balance in** $X$ between $Z{=}1$ and $Z{=}0$.

# Matching: Design-Stage Alignment (no hazard model)

- Build groups **comparable in** $X$; then KM contrasts are closer to causal (under assumptions).
- Not every unit must be used: lack of overlap $\Rightarrow$ trimming; estimand typically becomes **ATT**.

# Propensity Score: Design-Stage Adjustment

- $e(X) = P(Z = 1 \mid X)$, the probability of treatment given covariates $X$.
- Estimate $\hat{e}(X)$ (e.g., logistic regression or ML with nonlinearity/interactions).
- Use $\hat{e}(X)$ to **match** treated and control units so that distributions of $X$ are similar.

# Logistic Regression (Binary Treatment)

- Goal: model $e(X) = P(Z{=}1 \mid X)$.
- Model the **log-odds** (logit):

$$\log \frac{e(X)}{1 - e(X)} = \beta_0 + \beta^\top X$$

- Recover probability (sigmoid):

$$e(X) = \frac{1}{1 + \exp\{-(\beta_0 + \beta^\top X)\}}$$

- Use when $Z \in \{0, 1\}$.

# Reading & Building the Model

- $\beta_j$: change in **log-odds** per unit of $X_j$.
- $\exp(\beta_j)$: **odds ratio**.
- Include key covariates; allow **nonlinearity/interaction** if needed.
- Quick checks: separation problems? reasonable predicted $e(X)$ in $(0, 1)$?

# Caliper in Propensity Score Matching

- The **caliper** limits max distance when matching treated and control.
- Rule of thumb (Rosenbaum & Rubin, 1985):

$$\text{caliper} = 0.2 \times SD(\text{logit}(\hat{e}))$$

- Smaller calipers reduce bias but may drop more treated units (bias–variance trade-off).

# Why Logit and How to Compute SD

- Propensity scores $\hat{e}_i \in (0, 1)$; near 0/1 distances are compressed.
- $\text{logit}(\hat{e}_i) = \log\left(\frac{\hat{e}_i}{1-\hat{e}_i}\right)$ expands to $(-\infty, +\infty)$.
- Sample SD of logit scores:

$$SD(\text{logit}(\hat{e})) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\text{logit}(\hat{e}_i) - \overline{\text{logit}(\hat{e})}\right)^2}$$

- Keep matches with $|\text{logit}(\hat{e}_i) - \text{logit}(\hat{e}_j)| < \text{caliper}$.

# Matching Design Choices

- **With replacement**: a control can serve multiple treateds (lower bias, higher variance).
- **Without replacement**: each control used once; simpler variance.
- **Greedy NN** vs **Optimal** (minimize total distance).
- **Ratios** 1:$k$ / **Full matching** to use data efficiently.

# Caliper & Common Support

- Enforce a caliper on logit(PS); drop pairs beyond threshold.
- Trim units outside the **overlap region**; accept that some pairs are not comparable.
- Result: $X$ distributions become similar; target estimand is typically **ATT**.

# Standardized Mean Difference (SMD)

- **SMD** measures covariate balance on a common scale.
- Used **after matching/weighting** to check balance.

# SMD for Continuous Covariates

Average Standard Mean Deviation

$$\text{SMD}_k = \frac{\bar{X}_{1k} - \bar{X}_{0k}}{\sqrt{\frac{1}{2}\left(s_{1k}^2 + s_{0k}^2\right)}}$$

- $\bar{X}_{1k}, \bar{X}_{0k}$: means in $Z{=}1$ and $Z{=}0$.
- $s_{1k}^2, s_{0k}^2$: variances in each group.

# SMD for Binary Covariates

$$\text{SMD} = \frac{p_1 - p_0}{\sqrt{\frac{1}{2}\left(p_1(1-p_1) + p_0(1-p_0)\right)}}$$

- $p_1, p_0$: proportions in $Z{=}1$ and $Z{=}0$.
- Same idea: difference scaled by variability.

# How to Read SMD

- $|\mathrm{SMD}| < 0.10 \quad \Rightarrow$ **good balance**.
- $0.10 \sim 0.20 \quad \Rightarrow$ caution; may need tweaks.
- $> 0.20 \quad \Rightarrow$ **imbalance**; revisit design.

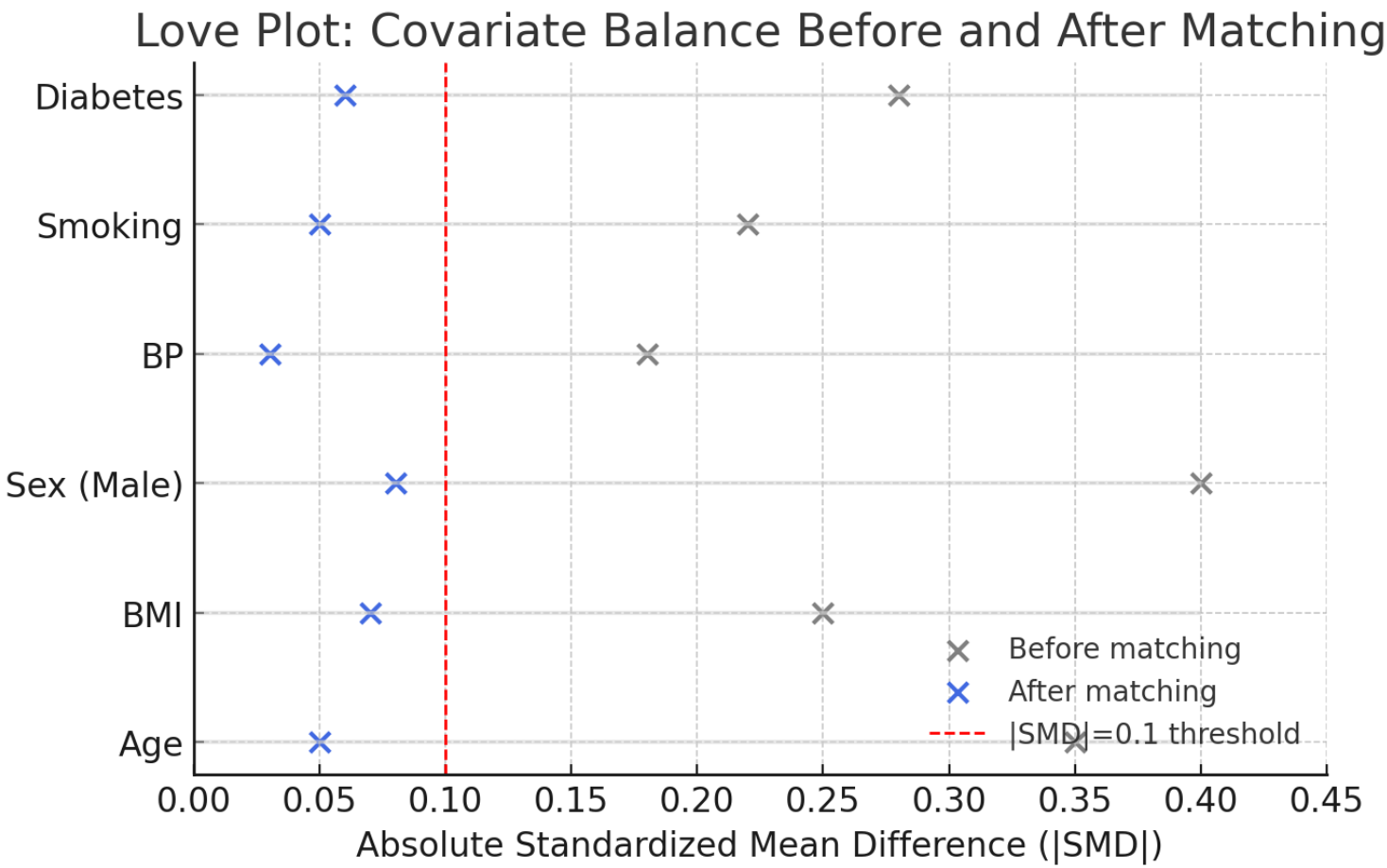
## Goal

Keep $|\mathrm{SMD}|$ small *for all covariates*.

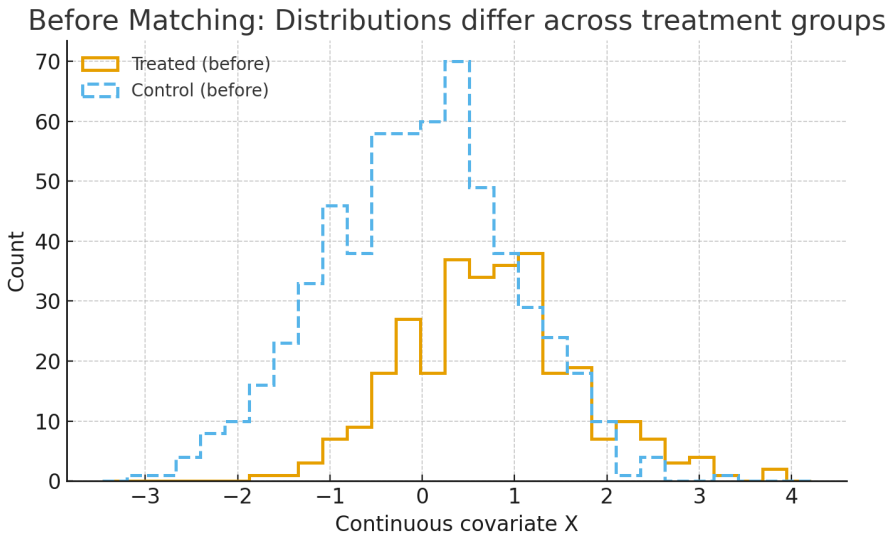
# Love Plot: Visualizing Covariate Balance

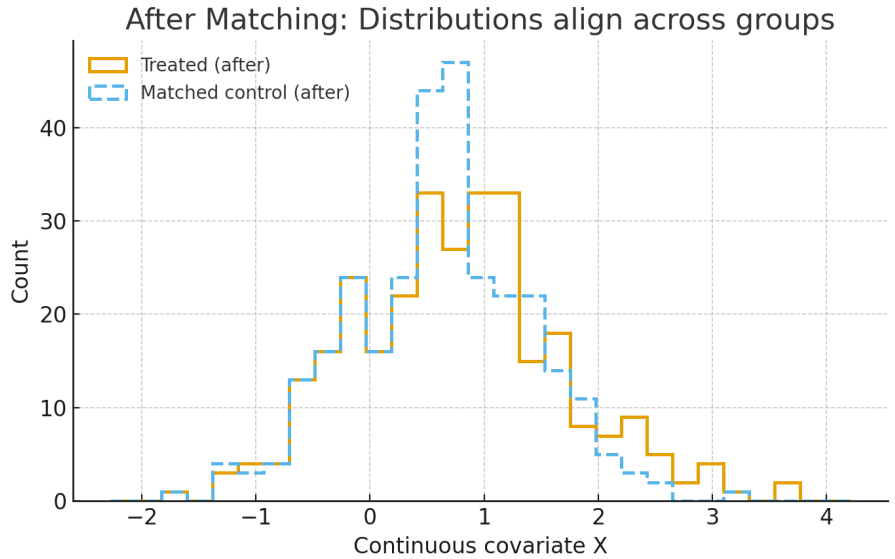
- A **Love plot** shows SMDs for each covariate: before vs after matching/weighting.
- X-axis: $|\mathrm{SMD}|$; goal: points move toward 0 (within $\pm 0.1$).

# Love Plot: Example



# Continuous $X$: Before vs After Matching



# Matching ⇒ KM: Workflow

1. **Choose covariates** $X$: affect both treatment and outcome.
2. **Estimate PS** $\hat{e}(X)$: logistic/ML.
3. **Match**: 1:1 nearest neighbor; common caliper $0.2 \times \mathrm{SD}(\mathrm{logit}(\hat{e}))$; decide with/without replacement.
4. **Check balance**: SMD as primary; examine PS overlap.
5. **KM on matched sample**: compare survival (and optionally log-rank) with pairing/cluster-aware variance.

# Technical Details: Setup & Estimand

- Design: **1:1 nearest-neighbor matching, without replacement** (caliper applied as needed).
- Analysis set: **only** successfully matched pairs; unmatched units are excluded.
- Target estimand: typically
- All subsequent estimation and inference are performed on the matched sample.

# Technical Details: KM Construction on Matched Sample

- Form two groups within the matched sample:
  - Treated group: all matched treated units.
  - Control group: their matched controls (one per treated).
- Compute standard Kaplan–Meier curves:

$$\widehat{S}_1(t) \text{ for } Z{=}1, \qquad \widehat{S}_0(t) \text{ for } Z{=}0.$$

- Weights: **unit weight = 1** for all individuals (no reweighting in 1:1, no-replacement matching).

# Technical Details: 1:$k$ Matching (No Replacement)

- Design: **1:$k$ nearest-neighbor**, each control used at most once (caliper on logit(PS) as needed).
- Analysis set: **only** matched units; unmatched are excluded.
- Estimand: typically **ATT** (effect for the matched treated population).
- Keep a **set ID** for each treated and its $k$ controls (for stratified inference).

# KM Under Constant Weights: They Cancel

- KM update at time $t$: $\widehat{S}(t^-) \to \widehat{S}(t^-) \times \left(1 - \frac{d_w(t)}{Y_w(t)}\right)$, where $d_w(t)$ and $Y_w(t)$ are weighted events and risk set sizes.
- If **all individuals in the same arm** are multiplied by the same constant $c$, then $d_w(t)$ and $Y_w(t)$ both scale by $c \Rightarrow$ the ratio is unchanged.
- Hence, in the **ideal** 1:$k$ case (every control has the same weight, e.g., all $1/k$): the control arm's KM curve is **identical** to the unweighted KM.

# Unequal Set Sizes: Set-Normalized Weights

- In practice, some sets are 1:$m_i$ with $m_i \in \{k, k-1, k-2, \dots\}$ due to caliper/overlap.
- Two coherent choices:
  - **Unweighted KM**: each individual has weight 1 (simple; larger sets contribute more).
  - **Set-normalized weights (recommended)**: for set $i$, give treated weight 1 and each of its $m_i$ controls weight $1/m_i$, so the *total control weight per set equals 1*.
- With set-normalized weights, weights **differ across sets**, so KM can **change** (compared to unweighted) in a way that equalizes *set influence*.

# KM Construction on the Matched Sample

- Build KM curves on the **matched sample only**:

$$\widehat{S}_1(t) \text{ for } Z{=}1, \qquad \widehat{S}_0(t) \text{ for } Z{=}0.$$

- 1:$k$ (no replacement):
  - Treated arm: all matched treated units.
  - Control arm: matched controls (up to $m_i$ per set).
- Weights:
  - **Unweighted**: all units weight $= 1$.
  - **Set-normalized**: treated $= 1$, each control in set $i = 1/m_i$.

# Technical Details: Comparing Curves

不見得要交料上の

- Visual contrast: overlay $\widehat{S}_1(t)$ and $\widehat{S}_0(t)$.
- Pointwise CIs for each KM curve: standard **Greenwood** variance.
- For differences (e.g., $\widehat{S}_1(t) - \widehat{S}_0(t)$, RMST difference):
  - **Pair-stratified** variance / log-rank test, or
  - **Paired bootstrap** (resample by *pairs* as the resampling unit).

加分題 (5分)
5-6頁模擬

# Inverse Probability Weighting (IPW)

- Goal: use **all subjects** and re-weight individuals by their propensity score (PS) to make the comparison across $Z{=}1$ vs $Z{=}0$ more credible in practice.
- Idea: subjects who are *under-represented* in a treatment arm receive **larger weights**.
- Output: two **weighted KM** curves, one per arm, built on the full dataset (no matching, no trimming by design).

# Propensity Score and Unit Weights

- Propensity score: $e(X) = P(Z{=}1 \mid X)$ (estimated via logistic regression or flexible ML).
- Common (practical) weights for two arms:

$$w_i = \frac{Z_i}{e(X_i)} + \frac{1 - Z_i}{1 - e(X_i)}$$

- Stabilized variant to temper extreme weights:

$$w_i^{\text{stab}} = \frac{Z_i \ \Pr(Z{=}1)}{e(X_i)} + \frac{(1 - Z_i) \ \Pr(Z{=}0)}{1 - e(X_i)}$$

- Use whichever keeps weights well-behaved; diagnostics on the next slides.

# Weighted Kaplan–Meier with $Y_w(t)$

- For each arm ($Z$=1 and $Z$=0), at each event time $t$:

$$Y_w(t) = \sum_{i \in \mathcal{R}(t)} w_i \qquad \text{(weighted risk set)}$$

$$d_w(t) = \sum_{i \in \mathcal{D}(t)} w_i \qquad \text{(weighted number of events)}$$

$$\widehat{S}(t) = \widehat{S}(t^-)\left(1 - \frac{d_w(t)}{Y_w(t)}\right)$$

- Plot one weighted curve per arm: $\widehat{S}^{(1)}(t)$ and $\widehat{S}^{(0)}(t)$.
- Censoring is handled as in standard KM (assumed non-informative for this stage).

# Weight Hygiene (Keep It Stable)

- **Check overlap**: visualize PS by arm; avoid regions where $e(X) \approx 0$ or 1.
- **Stabilize**: prefer stabilized weights if raw weights are highly variable.
- **Trim/Cap (practical)**: clip $e(X)$ to $[\varepsilon, 1-\varepsilon]$ (e.g., $\varepsilon \in [0.01, 0.05]$) or cap weights at a high percentile.
- Re-check that large weights are rare and do not dominate a few individuals.

# Reading & Reporting (Practice-First)

- **Reading**: compare the two weighted KM curves; optionally report differences at key times or RMST up to a horizon $\tau$.
- **Report** succinctly:
  - how $e(X)$ was estimated (model/features),
  - whether weights were stabilized and/or trimmed,
  - basic diagnostics (PS overlap, weight distribution).
- Reminder: this IPW approach **does not use matching**; it **keeps all observations** and re-weights them.

# IPW for KM: Quick Checklist

1. Estimate $e(X)$ with a flexible but transparent model.
2. Compute individual weights $w_i$ (stabilized if needed).
3. Build **two** weighted KMs using $Y_w(t)$ and $d_w(t)$ in each arm.
4. Inspect PS overlap and weight tails; trim/cap if necessary and re-run.
5. Present curves + a brief note on PS model and weight handling.

# Matching vs. IPW: Two Practical Paths

- **Matching**: design first, build a **comparable subsample** and then compare.
- **IPW**: keep **all observations**, re-weight to improve comparability across arms.
- Goal: make Kaplan–Meier comparisons **more credible and interpretable**.

# Data Usage & What Each Curve Represents

- **Matching**: analyze **matched units only**; units without good neighbors (e.g., outside caliper) are excluded.
- **IPW**: **use the full dataset**; underrepresented subjects are **up-weighted** to reduce baseline differences.
- **Interpretation**:
  - Matching: KM curves on the **matched subsample** (clean, comparable subset).
  - IPW: **weighted KM** curves representing the reweighted full sample.

# How the KM Is Constructed (No Inference Yet)

- **Matching**: on the matched sample, draw two KM curves (treated vs. control); no weights.
- **IPW (weighted KM)**: for each arm, at each event time $t$,

$$Y_w(t) = \sum_{i \in \mathcal{R}(t)} w_i, \quad d_w(t) = \sum_{i \in \mathcal{D}(t)} w_i, \quad \widehat{S}(t) = \widehat{S}(t^-)\Big(1 - \frac{d_w(t)}{Y_w(t)}\Big).$$

- Practical tip: estimate the propensity score $e(X)$ reasonably; use **stabilization/trimming** to avoid domination by extreme weights.

# Variability & Common Misconceptions

- "IPW keeps all data, so variance must be smaller" — **not necessarily**. If weights are extreme, the **effective sample size drops** and variance can increase.
- "Matching discards units, so variance must be larger" — **not necessarily**. On the matched subsample (no extreme weights), curves often look **more stable**; the tradeoff is that they represent the subset.
- Key message: focus on **PS overlap** and the **weight distribution**, not only nominal sample size.

# When to Prefer Which? (Practical Cheat Sheet)

- **Poor overlap / extreme PS**: start with **Matching** (caliper / trimming / full matching) to avoid exploding weights.
- **Good overlap & desire to retain all data**: **IPW** is convenient; stabilize and trim tails if needed.
- **Teaching / communication**: show **Matching** first (intuitive), then **IPW** as a reweighted full-sample view; agreement between the two is reassuring.

# How to Compare Fairly & What to Report

1. **Design both** on the same dataset: Matching and IPW.
2. **Diagnostics**:
   - Matching: matching rate, post-match SMDs, PS overlap (on matched sample).
   - IPW: **weighted** SMDs, PS overlap, **weight distribution/tails** (stabilized? trimmed?).
3. **Plot**:
   - Matching: two KM curves on the matched subsample.
   - IPW: two **weighted KM** curves on the full sample (via $Y_w(t)$ and $d_w(t)$).
4. **Read** differences at key times or via RMST (formal inference in the next chapter).
5. **Explain** design-driven differences: subset (Matching) vs. reweighted full sample (IPW). If results diverge, revisit overlap and weight tails.

# Relevant Packages for KM in R

- `survival`: provides `Surv()`, `survfit()`, and the example dataset `aml`.
- Outcome coding: `status` is typically 1 = event, 0 = censored.
- Optional: `survminer` (`ggsurvplot`) for publication-ready plots.

**Example (AML Maintenance Study)**

```
library(survival)
leukemia.surv <- survfit(Surv(time, status) ~ x, data = aml)
plot(leukemia.surv, lty = 2:3)
legend(100,0.9,c("Maintenance","No Maintenance"),lty = 2:3)
title("Kaplan{Meier Curves\nfor AML Maintenance Study")
```

# Setup & 1:1 Nearest-Neighbor Matching

- Data: `lalonde` (from `MatchIt`); treatment = `treat`, covariates include age.
- Design: 1:1 nearest-neighbor PS matching, no replacement (defaults).

```
library(MatchIt)
library(cobalt)   # for SMDs / Love plot (optional)

data("lalonde", package = "MatchIt")

# 1:1 NN matching without replacement (default)
m.out1 <- matchit(
  treat ~ age + educ + race + nodegree + married + re74 + re75
  data = lalonde
)

m.out1            # quick look
summary(m.out1)   # balance before/after (SMDs, etc.)
```

# Extract Matched Sample

- Use only the **matched sample** to draw KM or diagnostics.
- If replacement/full matching were used, `weights` will reflect reuse/set-weights.

```
# Matched (post-design) sample
m.dat <- match.data(m.out1)   # contains treat, age, and weigh

# (Optional quick splits if you need them)
treated_dat <- subset(m.dat, treat == 1)
control_dat <- subset(m.dat, treat == 0)
```

# Age: Prepare Data (Before vs After)

- Build a combined dataset with a `stage` flag.
- `Before`: original `lalonde`; `After`: matched `m.dat`.

```
library(ggplot2)
# Build a combined dataset: Before vs After
before_dat <- lalonde[, c("age","treat")]
before_dat$weights <- 1
before_dat$stage   <- "Before"
after_dat  <- m.dat[, c("age","treat","weights")]
after_dat$stage    <- "After"
plot_dat <- rbind(before_dat, after_dat)
plot_dat$group <-
ifelse(plot_dat$treat == 1, "Treated", "Control")
plot_dat$group <-
factor(plot_dat$group, levels = c("Control","Treated"))
```

# Age Histograms: Before vs After Matching

- Overlaid histograms only (no smoothing), high transparency.
- Left panel: Before; Right panel: After.

```
ggplot(plot_dat, aes(x = age, fill = group)) +
  geom_histogram(aes(y = after_stat(density),
  weight = weights),
                 position = "identity",
                 bins = 30, color = NA, alpha = 0.25) +
  facet_wrap(~ stage, ncol = 2) +
  labs(x = "Age", y = "Density", fill = "Group",
       title = "Age Histograms: Before vs After Matching") +
  theme_minimal() +
  theme(legend.position = "top")
```